Computational and Statistical Approaches for Large-Scale Genome-Wide Association Studies

by

Wei Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Cristen J. Willer, Co-Chair
Assistant Professor Seunggeun Lee, Co-Chair
Professor Gonçalo R. Abecasis
Professor Margit Burmeister
Associate Professor Hyun Min Kang
Assistant Professor Stephen C.J. Parker

Wei Zhou

zhowei@umich.edu

ORCID iD: 0000-0001-7719-0859

Acknowledgements

This dissertation would not have been possible without the immense support from many people. First of all, I would like to express my most sincere gratitude to my thesis advisor Dr. Cristen J. Willer, who encouraged me to do a Ph.D. and guided me through the entire process with her endless patience and motivation. Cristen has always given me freedom to pursue different research directions, provided insightful and intellectual suggestions, asked questions to have me think harder, and inspired me to take on and overcome challenges. Not only is she an amazing mentor whose dedication and enthusiasm in science has a profound influence on my career, but also a role model from whom I have learned how to maintain work-life balance as a scientist and mother of two.

I am especially grateful for my co-advisor Dr. Seunggeun (Shawn) Lee for providing me the invaluable opportunities to pursue my research interests in statistical method development. He helped me envision the field of statistical genetics, guided me to overcome all difficulties I have met, and instilled my confidence in working on statistical and computational problems for genetic studies.

I would like to thank Dr. Hyun Min Kang for co-directing me on my second thesis project, from whom I have learned a tremendous amount about computational and statistical methods for genetic data anaysis. I am also very appreciative to Dr. Goncalo R. Abecasis for his guidance on my rotation project, his support to my dissertation projects and his kindness to let me attend the weekly meeting with his group to discuss problems in my research. I would like to particularly thank Dr. Margit Burmeister for always taking the time to answer my questions, encouraging me to apply for awards and fellowships, and giving invaluable suggestions on my work. I would also like to thank Dr. Stephen C.J. Parker and other members of my dissertation committee who I have mentioned above for your relentless support in my dissertation research.

I gratefully acknowledge the funding received towards my PhD from Rackham Predoctoral Fellowship 2017-2018.

It has been a wonderful experience working with and learning from many talented colleagues and lab mates. I would like to thank Jonas B. Nielson, Lars G. Fritsche, He Zhang, Jin Chen, Maoxuan Lin, Ida Suraka, Sarah Graham, Jonathon LeFaive, Peter VandeHaar, Sayantan Das, Sarah Gagliano, Jingjing Yang, Rounak Dey, and all other members from the Willer group, the Lee group, the Abecasis group, and the Center for Statistical Genetics. I am also very grateful for all wonderful collaborators from the HUNT study and the UM Cardiovascular Health Improvement Project (CHIP) Study.

I am so honored to study in the Bioinformatics program and I want to thank the wonderful staff in the program, especially Julia Eussen, for always being so helpful and friendly. I extend gratitude to my fellow graduate students Ellen Schmidt, Brooke Wolford, Xuefang Zhao, Shweta Ramdas, Sai Chen, Fan Zhang, Yaya Zhai and Nguyen Vo for their friendship, support and help, which means so much to me.

Finally, I would like to express my deepest gratitude to my wonderful family. Thank you to my parents and parents-in-law for providing unconditional love and care and sacrificing their lives to help take care of my kids. Thank you to my incredibly supportive husband, whose love, understanding, and inspiration has helped me get through hard times and kept me brave and confident to face any challenge. I want to thank my daughter Emily and my son William for making me smile every day. Without their blessings and encouragement, I would not have been able to finish this work.

Table of Contents

Lists of Tables

Lists of Figures

Abstract

Over the past decade, genome-wide association studies (GWAS) have proven successful at shedding light on the underlying genetic variations that affect the risk of human complex diseases, which can be translated to novel preventative and therapeutic strategies. My research aims at identifying novel disease-associated genetic variants through large-scale GWAS and developing computational and statistical pipelines and methods to improve power and accuracy of GWAS.

Bicuspid aortic valve (BAV) is a congenital heart defect characterized by fusion of two of the normal three leaflets of the aortic valve. As the most common cardiovascular malformation in humans, BAV is moderately heritable and is an important risk factor for valvulopathy and aortopathy, but its genetic origins remain elusive. In Chapter 2, we present the first large-scale GWAS study to identify novel genetic variants associated with BAV. We report association with a non-coding variant 151kb from the gene encoding the cardiac-specific transcription factor, GATA4, and near-significance for p.Ser377Gly in *GATA4*. We used multiple bioinformatics approaches to demonstrate that the *GATA4* gene is a plausible biological candidate. In the subsequent functional follow-up, GATA4 was interrupted by CRISPR-Cas9 in induced pluripotent stem cells from healthy donors. The disruption of *GATA4* significantly impaired the transition from endothelial cells into mesenchymal cells, a critical step in heart valve development.

Genotype imputation is widely used in GWAS to perform in silico genotyping, leading to higher power to identify novel genetic signals. When multiple reference panels are not consented to combine together, it is unclear how to combine the imputation results to optimize the power of genetic association tests. In Chapter 3, we compared the accuracy of 9,265 Norwegian genomes imputed from three reference panels – 1000 Genomes Phase 3 (1000G), Haplotype Reference Consortium (HRC), and a reference panel containing 2,201 Norwegian participants from the HUNT study with low-pass genome sequencing. We observed that the overall imputation accuracy from the population-specific panel was substantially higher than 1000G and was comparable with HRC, despite HRC being 15-fold larger. We also evaluated different strategies to utilize multiple sets of imputed genotypes to increase the power of association studies. We propose that testing association for all variants imputed from any panel results in higher power to detect association than the alternative strategy of testing only the version of each genetic variant with the highest imputation quality metric.

In phenome-wide GWAS by large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, linear mixed model and the recently proposed logistic mixed model, perform poorly -- producing large type I error rates -- in the analysis of phenotypes with unbalanced case-control ratios. In Chapter 4, we propose a scalable and accurate generalized mixed model association test that uses the saddlepoint approximation (SPA) to calibrate the distribution of score test statistics. This method, SAIGE, provides accurate p-values even when case-control ratios are extremely unbalanced. It utilizes state-of-art optimization strategies to reduce computational time and memory cost of generalized mixed model. The computation cost linearly depends on sample size, and hence can be applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK Biobank data of 408,961 white British European-ancestry samples for 1,403 dichotomous phenotypes, we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.

Chapter 1 Introduction

The principles of inheritance that were developed by Gregor Mendel based on his 19th century experiments on pea plant breeding remarkably expanded the world's understanding about genetic inheritance (G., 1866). Since then, large efforts have been made to study how genetic variations contribute to human diseases. As of November, 2017, the genes underlying 75% of all 6,727 known Mendelian disorders have been identified (Amberger, et al., 2015). However, complex human diseases (e.g. heart disease and diabetes) and quantitative traits (e.g. blood lipids and body mass index) are usually caused by genetic variants in multiple genes, each with relatively small effects, and environmental factors.

Complex diseases/traits are more common in the population than Mendelian disorders, which are caused by variants in a single gene, and the majority of them are 30% to 60% heritable(Price, et al., 2015). Identifying genetic risk factors for complex diseases/traits elucidates disease etiology and ultimately translates to novel preventative and therapeutic strategies. In spite of the prevalence, heritability and significance of human complex diseases/traits, the progress of decoding their genetic risk factors was slow until the initial completion of the human genome sequence in 2001(International Human Genome Sequencing, 2001). Within the past decade, genome-wide association study (GWAS) has emerged as the most efficient approach to identify the associations between genetic variants and complex diseases/traits. As of November 2017, 53,069 unique SNP-trait associations have been discovered (MacArthur, et al., 2017).

The parallel advent of DNA array technology, genotype imputation methods and next-generation sequencing technology allows large biobanks to genotype or sequence hundreds of thousands of participants. GWAS based on the human phenome (PheWAS), which consists of thousands of phenotypes that are constructed based on the real-time electronic health records and epidemiological data, is an emerging powerful approach for detangling genetics underlying human disease/traits(Bush, et al., 2016; Denny, et al., 2013). In addition to uncovering novel genotype-phenotype associations, PheWAS systematically examines the cross-phenotype association for each genetic marker, incorporates comprehensive information on the environmental factors(e.g. life style), and allows the investigation of causal relationships among phenotypes(Millard, et al., 2015). Earlier this year, the UK Biobank released both genome and phenome data for ~500,000 participants, which were collected from 2006 to 2010(Bycroft, et al., 2017; Sudlow, et al., 2015). The popularity of the large-scale PheWAS brings novel statistical and computational challenges.

1.1 Genome-wide association studies on human complex diseases/traits

Complex diseases/traits refer to the disorders or phenotypes that do not exhibit gene-phenotype co-segregation relationship as monogenic Mendelian diseases do (Lander and Schork, 2006). The disease susceptibilities are influenced by multiple genetic and environmental factors as well as possible interactions between them. Genetic mapping for human complex diseases/traits has drawn dramatic attention of researchers since the 1990s (Lander and Schork, 1994; Schork, 1997; Weeks and Lathrop, 1995) initially using methods falling into two main categories: linkage analysis and candidate gene association (Hirschhorn and Daly, 2005). As the traditional method widely used to map genes for Mendelian diseases, linkage analysis attempts to detect disease genes that segregate within the affected families through genetic markers flanking the genes(Hirschhorn and Daly, 2005). This method has not gained much success in gene mapping for complex diseases, because the majority of disease-associated genetic variants have relatively small to moderate effects, while linkage analysis lacks power to detect those variants(Hirschhorn and Daly, 2005). The candidate gene association approach, an alternative to linkage analysis, searches for variants in candidate genes that are depleted or enriched in disease groups. Although candidate gene association has successfully identified genetic variants that are associated with some complex diseases/traits, such as early-onset obesity(Vaisse, et al., 2000) and HDL(Cohen, et al., 2004), this method requires prior biological knowledge to select the disease-associated genes for study and it largely ignores the non-coding regions.

GWAS became practical after the draft of the human genome was initially completed in 2001(International Human Genome Sequencing, 2001) followed by the characterization of genome-wide linkage-disequilibrium (LD) patterns by the HapMap project in 2003 (International Human Genome Sequencing, 2001). As the most common type of genome variation, single nucleotide polymorphisms (SNPs) were firstly discovered in 1998 and have been observed to occur every 300 bases on average in the human genome(The International HapMap, 2007). Several hundreds of thousands of SNPs selected based on LD have been shown to be able to cover most of the common genome variations(Hirschhorn and Daly, 2005) and could be genotyped using SNP arrays(Price, et al., 2015). Over the past decade, GWASs, which test the association between genetic variants, one at a time, and the disease/trait of interest, have proven successful at shedding light on the underlying genetic variations that affect the risk of human complex diseases. The elucidation of the genetic basis of diseases/traits can be translated to novel preventative and therapeutic strategies. For example, PCSK9 inhibitors were developed to reduce cholesterol and as a therapeutic alternative to statins after the gene was identified in a previous GWAS(Kathiresan, et al., 2009; Teslovich, et al., 2010).

The common disease-common variants hypothesis, proposed in late 1990s(Cargill, et al., 1999; Chakravarti, 1999; Lander, 1996), states that the genetic risk of common human diseases are mainly attributed to common genetic variants with low to modest effects, which are usually defined as variants with minor allele frequency (MAF) > 1%(Reich and Lander). Based on this hypothesis, early GWASs focused on identifying common variants using SNP chip arrays, which mainly detect common variants. However, the common disease-associated variants that have been identified in GWASs only explain a limited proportion of disease heritability(Manolio, et al., 2009). A possible explanation is that the missing disease heritability is due to the rare variants, which usually have MAF < 1%(Frazer, et al., 2009; Saint Pierre and Génin, 2014).

Rare variants became accessible as the further development of genotype imputation methods and next-generation sequencing technologies. Multiple GWASs start identifying association between rare variants and complex diseases/traits(Fritsche, et al., 2015; Long, et al., 2017; Marouli, et al., 2017; Sebastiani, et al., 2017). For example, the International Age-related macular degeneration (AMD) Genomics Consortium has identified seven rare variants (MAF<1%) significantly associated with AMD with odds ratios of 1.1-47.6(Fritsche, et al., 2015).

The power of a GWAS can be defined as the probability of successfully identifying the association between a genetic variant and the phenotype given a true association exists. The overall study power can be seen as a combination of the power to obtain the correct genotypes of the genetic variant (genotyping power) and the statistical power for the significance testing (statistical power). In a sequencing-based study, the genotyping power relies on the sample size, sequencing depth, sequencing errors, and allele frequency of the variant. In a study based on genotype imputation (described in 1.2), the genotyping power depends on the size of the reference panel, genetic similarity between samples in the reference panel and study samples, and the allele frequency. To well control the family-wise error rate at 0.05, $5x10^{-8}$ has been determined as the genome-wide significance threshold for GWAS for European samples by simulation studies using data on HapMap Encyclopedia of DNA Elements (ENCODE) regions(Pe'er, et al., 2008). With the genome-wide significance threshold, the statistical power is influenced by sample sizes, allele frequency, the effect sizes of the tested variant, as well as some factors that differ across diseases, such as the complexity of genetic architecture and phenotyping accuracy(Price, et al., 2015; Sham and Purcell, 2014).

In Chapter 2, we will present the first large-scale GWAS study to identify novel genetic variants that are associated with bicuspid aortic valve(BAV), the most common cardiovascular malformation in humans(Tutar, et al., 2005).

1.2 Genotype imputation

Genotype imputation is an approach to infer missing genotypes for untyped genetic variants in study samples from a sequenced reference panel with high-density haplotypes(Li, et al., 2009). More specifically, given that samples with similar ancestry backgrounds tend to share chromosome stretches inherited from common ancestors, the genetic markers that have been sparsely genotyped for study samples using a commercial array can be used to identify the chromosome stretches shared by study samples and the reference panel, thereby allowing statistical imputation of the missing genotypes(Li, et al., 2009). Although the cost of whole-genome sequencing(WGS) has substantially dropped recently, array based genotyping followed by imputation is still more cost-efficient than WGS (~ 20-fold lower) to uncover complete sets of genetic variants across the genome since computation and personnel effort is the only cost in the imputation process. Genotype imputation has several other advantages which makes it widely used by GWAS for different human complex diseases/traits. 1. Genotype imputation boosts the power of GWAS and fine-mapping to detect causal variants by increasing the resolution of the disease-associated genetic loci with additional genetic variants added, including potentially causal ones. Both simulation studies(Marchini and Howie, 2010; Spencer, et al., 2009) and GWAS on complex diseases/traits(Marchini, et al., 2007; Orho-Melander, et al., 2008) have suggested this contribution made by genotype imputation. 2. The association power of GWAS can be increased by genotype imputation because of the increase in the effective sample size. Genotypes that are partially missed in a subset of study samples will be filled in by imputation. 3. Genotype imputation makes it possible to meta-analyze GWAS studies that genotype their samples using different commercial arrays.

Multiple imputation tools have been developed within the last decade(Browning and Browning, 2016; Browning, 2008; Das, et al., 2016; Fuchsberger, et al., 2015; Howie, et al., 2012; Howie, et al., 2009; Li, et al., 2010; Marchini, et al., 2007; Scheet and Stephens, 2006). The most widely-used methods (fastPHASE(Scheet and Stephens, 2006), MaCH/minimac(Das, et al., 2016; Fuchsberger, et al., 2015; Howie, et al., 2012; Li, et al., 2010), Impute(Howie, et al., 2009; Marchini, et al., 2007), and Beagle4.1(Browning and Browning, 2016)) use the Li and Stephen Model, which is a framework using Hidden Markov Model to describe the genetic data structure(Li and Stephens, 2003). These methods take characteristics of human genomes, such as linkage and recombination rates between the sites, mutation rates and genotyping error rates, into account when imputing the missing genotypes, leading to higher imputation accuracy compared with other methods(Li and Stephens, 2003). The Michigan Imputation Server (https://imputationserver.sph.umich.edu) is a cloud-based server providing free genotype imputation using the imputation engine minimac3(Das, et al., 2016) with multiple reference panel options, including the Haplotype Reference Panel (HRC)(McCarthy, et al., 2016), 1000G Phase 1 and 3(The Genomes Project, 2015) and HapMap Phase 2(The International HapMap, 2007).

The genotype imputation accuracy mainly depends on the sample size of the reference panel(Browning and Browning, 2009; Howie, et al., 2009; Huang, et al., 2009; Li, et al., 2009; Roshyara and Scholz, 2015) and the genetic similarity between the reference panel and the target samples(Deelen, et al., 2014; Huang and Tseng, 2014; Huang, et al., 2015; Low-Kam, et al., 2016; Mitt, et al., 2017; Okada, et al., 2015; Pistis, et al., 2015; Roshyara and Scholz, 2015; Walter, et al., 2015). Thus both publicly available reference panels (e.g. HRC and 1000G) (McCarthy, et al., 2016; The Genomes Project, 2015) and population-specific reference panels generated through whole-genome sequencing (WGS) a subset of study samples by individual studies(e.g. The Genome of the Netherlands Consortium and the UK10K study) (Deelen, et al., 2014; Huang, et al., 2015) have their own advantages. In Chapter 3, we will address the question how to combine imputed genotypes from multiple reference panels to achieve higher power in subsequent GWASs.

1.3 Correct for sample relatedness and population stratification in GWAS

As Fisher's 1919 paper stated, polygenic quantitative traits could be explained by Mendelian inheritance, suggesting that samples with similar genetic backgrounds tend to have more correlated quantitative traits than distantly related samples(Fisher, 1919). This is also true for polygenic dichotomous traits since the genetic liability is continuous(Plomin, et al., 2009). Due to the violation of the independence assumption between samples, GWAS results based on linear or logistic regression can be biased by sample relatedness, familial or cryptic, and population stratification, leading to spurious associations.

Several methods have been developed to correct for sample substructure, including genomic control(Devlin and Roeder, 1999), principal component analysis(Patterson, et al., 2006), and mixed model-based methods. The genomic control factor λ_{GC} is defined as the median of the observed χ^2(with 1 degree of freedom) association test statistics of all tested genetic markers divided by the expected theoretical median of χ^2 under the null hypothesis(Devlin and Roeder, 1999). $\lambda_{GC} > 1$ indicates the presence of inflated type I error rates due to sample substructures or differential bias(Clayton, et al., 2005). Dividing the observed test statistics χ^2 by λ_{GC} is a simple approach to correct for the inflated type I errors but is generally thought to be inadequate and inappropriate to use such a single adjustment for all SNPs(Price, et al., 2010). Principal component analysis (PCA) is a well-developed statistical tool to infer population substructure(Patterson, et al., 2006). It can be used to detect sample outliers due to batch effects or population stratification. Top PCs are usually included in linear/logistic model as covariates to correct for population stratification, while they do not correct for sample relatedness(Patterson, et al., 2006).

Mixed model methods have long been used for selection of animal breeding(Henderson, 1984) before being applied to association mapping on humans with known pedigrees(Abney, et al., 2002; Chen and Abecasis, 2007) and unknown pedigree(Yu, et al., 2005). Models that accounts for fixed effects and random effects jointly are referred as mixed models(Eisenhart, 1947). Intuitively, mixed models incorporate the pairwise sample relatedness to capture confounders such as sample relatedness and population stratification. A substantial challenge to using mixed models for GWAS is their high computational burden from the iterative numeric optimization procedure, especially when the number of tested genetic markers and the sample size are large.

One of the main milestones in reducing the computational time is the **two-stage approach** that was proposed by Chen and Abecasis in 2007(Chen and Abecasis, 2007). This approach assumes that each genetic locus has a small effect on the phenotype. In step one, the variance parameters are estimated once in the null model, which does not include any fixed genetic effects. Then the estimated variance parameters are used for the association test for each genetic marker in step two. This approach avoids the computationally expensive task of estimating the variance component parameters for all genetic markers which substantially reduces the computational time for GWAS. The drawback is that the p-values at step two are approximated assuming the effect size of the tested genetic marker is small. The estimates may be biased, leading to power loss, if the assumption is violated. All mixed model methods using this approach are called approximate methods, while others are called exact methods. Figure 1-1 presents mixed model methods that have been developed since 2006 for population-based genetic association tests with no pedigree information since 2006. Below is a brief review for these methods.

Q+K is a unified mixed model method developed in 2006(Yu, et al., 2005). This method is the first to propose using the relative kinship (K) matrix to replace the pedigree-based co-ancestry matrix of the traditional mixed model to account for sample relatedness when the pedigree information is unknown. An additional variance matrix Q accounts for population structure and is included if the population structure is present. With both K and Q matrices, multiple levels of relatedness between samples are systematically corrected. This was the first study to use a linear mixed model for association mapping with unknown pedigree information has shown that the linear mixed model results in lower type I and II error rates and higher power than other previously used methods including genomic control and structured association. Q+K has been implemented in the software TASSEL(Bradbury, et al., 2007).

EMMA is a mixed-model association method developed by Kang, et al. in 2008 for model organism association mapping but a similar framework can be used for genetic association tests in human populations(Kang, et al., 2008). Compared to Q+K, EMMA has substantially increased the computation speed and reliability of the mixed-model association mapping using the following techniques. 1. Spectral decomposition is used to compute the likelihood which avoids the large number of matrix multiplications and inversions at each iteration. This allows optimization of likelihood function or restricted maximum-likelihood function (REML) in a single-dimensional parameter space. 2. The dramatic decrease in computation time makes it feasible to obtain the global optimization with high confidence by the combination of grid search and Newton-Raphson algorithm. 3. A simpler Identity-By-State allele-sharing genetic similarity matrix is used as a kinship matrix to account for sample relatedness. 4. REML takes the degrees of freedom of fixed effects into account and provides unbiased maximum likelihood variance component estimates.

EMMAX further expedited EMMA by using the two-stage approach. It was developed by the same author as EMMA in 2010 and applied to a human population(Kang, et al., 2010) EMMAX decreases the computation time for large GMAS with thousands of samples and hundreds of thousands of genetic markers to hours compared to years using EMMA.

Compressed MLM clusters samples into groups based on kinship and use the kinship between groups for random effects(Zhang, et al., 2010). This reduces the time complexity from the cubic of the sample size to the cubic of the number of groups. However, the maximum time complexity reduction was shown to be about 20 fold(Huang, et al., 2010). **P3D** also uses the two-stage approach, except that in step two, a fixed previously determined population parameters (P3D) is used(Zhang, et al., 2010). Compressed MLM and P3D were implemented in the software TASSEL(Bradbury, et al., 2007).

FaST-LMM stands for factored spectrally transformed LMM and was developed in 2011(Lippert, et al., 2011). It has a clever use of spectral decomposition for the genetic relationship matrix (GRM) to rotate the phenotypes, so that the rotated data become independent and a simple linear regression can be used for genetic association testing. Unlike the previous two-stage approach method, FaST-LMM does not need to assume all genetic markers have small effects and provides exact p-values.

GEMMA is another exact linear mixed model method. It is similar to EMMA, but is n (n = sample size) times faster(Zhou and Stephens, 2012). Instead of using the expensive spectral decomposition, GEMMA uses a few recursions to compute some induced quantities that are needed for the optimization of the likelihood function or REML. The recursions only involve in matrix-vector multiplications and have quadratic complexity of sample size.

GRAMMA-Gamma improves the computation efficiency of the second step of the two-stage approach by using a constant called GRAMMA-Gamma factor(Svishcheva, et al., 2012), the ratio of the GRAMMA score test statistic that ignore the random effects and the one that incorporate the random effects. This ratio has been shown to be constant empirically and analytically(Aulchenko, et al., 2007). The score test statistics for each genetic marker can be computed with no random effects and adjusted using GRAMMA-Gamma factor.

BOLT-LMM is the only existing linear mixed model method that can handle large sample sizes such as UK Biobank (n=~500,000) (Loh, et al., 2015). It utilizes several efficient strategies to improve the computational efficiency of the algorithm. To reduce the computational time, it uses retrospective mixed-model association statistics, similar to GRAMMA-Gamma. BOLT-LMM uses the following techniques to save on memory usage, 1. It replaces the spectral decomposition method with the conjugate gradient method to obtain the inverse of the GRM matrix. 2. Instead of storing the GRM matrix, it computes the elements of GRM as needed. 3. It stores the hard-call genotypes to be used for GRM elements computation in a binary vector. Besides the algorithm feasibility for large sample sizes, BOTL-LMM improves the test power by using a Gaussian mixture model as a Bayesian prior to modeling the effect sizes that are following the infinitesimal model, calibrating the test statistics using the LD Score regression technique, and adapting the leave-one-chromosome-out (LOCO) scheme to avoid proximal contamination.

GMMAT is the first logistic mixed model method proposed for association tests and was developed by Chen *et al.* in 2016(Chen, et al., 2016). Although linear mixed model methods have been widely used for sample relatedness correction, the homoscedasticity assumption that assumes constant residual variance regardless of the covariate values is usually violated by binary traits. The authors have shown that for genetic association tests for dichotomous traits, GMMAT successfully corrects the inflated type I error rates observed in linear mixed model methods. However, the high computational cost of GMMAT keeps it from being widely used in large-scale GWASs (sample size n > 20,000) for binary traits.

1.4 Statistical methods to account for case-control imbalance in GWAS

Phenotypes for human diseases are dichotomous (affected/unaffected). In population-based biobanks, binary phenotypes often have case-control ratios that are unbalanced (<1:10) or very unbalanced case-control ratio (<1:100). For example, more than 85% of 1603 binary phenotypes in UK Biobank have case-control ratio < 1:99(Bycroft, et al., 2017; Sudlow, et al., 2015). The parameter estimates based on the maximum likelihood function are biased or even infinite in presence of the unbalanced case-control sampling(Albert and Anderson, 1984; Ma, et al., 2013). To address this issue, Firth proposed a bias-corrected log-likelihood function that is penalized with an information matrix(Firth, 1993), based on which Heinze and Schemper described a likelihood ratio test, called Firth test(Heinze and Schemper, 2002). Despite its good performance on correcting the bias due to case-control imbalanced in logistic regression, the Firth test is computational inefficient because the maximum likelihood estimates(MLE) need to be obtained under both full and null models.

Most of the mixed model methods use score tests for genetic association because model parameters are only estimated under the null model, leading to higher computational efficiency than the likelihood-ratio tests and Wald tests. However, case-control imbalance also results in inflated type I error rates in score tests based on logistic regression, especially for low-frequency variant(Dey, et al., 2017; Ma, et al., 2013). This is because skewness in the tail exists when the case-control ratios are unbalanced but normality is incorrectly assumed in the traditional score tests. This is especially true for less frequent variants. The normal distribution does not capture the tail skewness using the first two cumulants, mean and the variance. Saddlepoint approximation(SPA) was firstly introduced by Daniels and it uses the entire cumulant generating function to approximate a distribution rather than only the first two used by normal approximation(Daniels, 1954). Rounak *et al*. proposed using SPA to approximate the score test distribution in logistic regression for unbalanced case-control phenotypes and implemented the SPA test in an R pacakage SPAtest(Dey, et al., 2017). SPA test accounts for case-control imbalance for dichotomous traits as well as Firth test does, and the SPA test is 100 to 300 times faster than the Firth test(Dey, et al., 2017). In Chapter 5, we implement a method that combines advantages of logistic mixed model approaches and the saddle-point approximation of score test distribution to efficiently and robustly analyze phenotypes with unbalanced case-control ratios and sample relatedness. We utilize state-of-art optimization strategies to make our method practical for very large sample sizes.

1.5 Challenges for phenome-wide GWAS in large cohorts

Recently, increasing numbers of biobanks are able to genotype or sequence all of their participants, such as the UK Biobank(Bycroft, et al., 2017; Sudlow, et al., 2015), the population-based Nord Trøndelag Health Study (HUNT) (Krokstad, et al., 2013) and Michigan Genomics Initiative (https://www.michigangenomics.org). Phenome-wide GWASs for thousands of phenotypes based on ICD codes have both statistical and computational challenges in these large cohorts. Sample relatedness and case-control ratio imbalance are the two common issues in such studies. Statistically, no method exists for GWAS that can address both issues simultaneously. Computationally, the time and memory cost for GWASs on tens to hundreds of millions of genetic markers for hundreds of thousands of samples for thousands of phenotypes is significantly high.

1.6 Dissertation Outline

My research aims at identifying novel disease-associated genetic variants through large-scale GWAS and developing computational and statistical pipelines and methods to improve power and accuracy of GWAS.

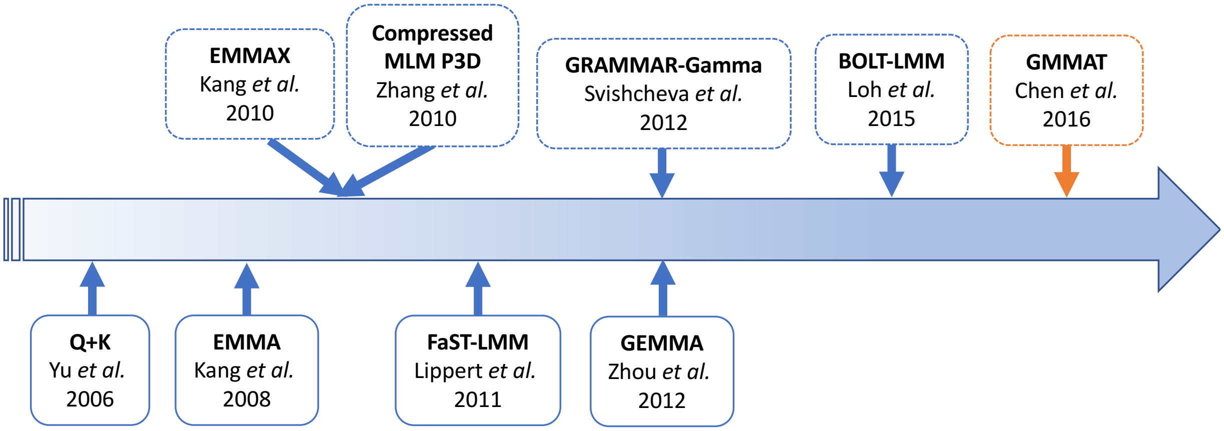
Bicuspid aortic valve (BAV) is a birth defect of the heart characterized by fusion of two of the normal three leaflets of the aortic valve. BAV is the most common cardiovascular malformation in humans(Hoffman and Kaplan, 2002; Tutar, et al., 2005), moderately heritable(Cripe, et al., 2004; Ellison, et al., 2007; Garg, 2006), is associated with serious consequences, such as dilated thoracic aorta, severe aortic valve stenosis, aortic valve incompetence and aortic dissection, which carries very high mortality (Losenno, et al., 2012; Michelena, et al., 2008; Michelena, et al., 2011; Siu and Silversides, 2010; Ward, 2000). In Chapter 2, we present the first large-scale GWAS study to identify novel genetic variants that are associated with BAV. We report association with a non-coding variant 151kb from the gene encoding the cardiac-specific transcription factor, GATA4, and near-significance for p.Ser377Gly in *GATA4*. We use multiple bioinformatics approaches to demonstrate that the *GATA4* gene is a plausible biological candidate, whose potential roles in heart valve development has been further investigated through functional follow-up.

Genotype imputation is widely used in GWAS to perform in silico genotyping, leading to higher power to identify novel genetic signals. When multiple reference panels are not consented to combine together, it is unclear how to combine the imputation results to optimize the power of genetic association tests. In Chapter 3, we compare the accuracy of 9,265 Norwegian genomes imputed from three reference panels – 1000 Genomes Phase 3 (1000G)(The Genomes Project, 2015), Haplotype Reference Consortium (HRC)(McCarthy, et al., 2016), and a reference panel containing 2,201 Norwegian participants from the HUNT study with low-pass genome sequencing(Krokstad, et al., 2013). We also evaluate different strategies to utilize multiple sets of imputed genotypes to increase the power of association studies. We propose that testing association for all variants imputed from any panel results in higher power to detect association than the alternative strategy of testing only the version of each genetic variant with the highest imputation quality metric.

In GWAS for thousands of phenotypes in large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, linear mixed model and the recently proposed logistic mixed model, perform poorly -- producing large type I error rates -- in the analysis of phenotypes with unbalanced case-control ratios. In Chapter 4, we propose a scalable and accurate generalized mixed model association test that uses the SPA to calibrate the distribution of score test statistics based on logistic mixed models. This method, SAIGE, provides accurate p-values even when case-control ratios are extremely unbalanced or when allele frequencies are low, two situations which all other approaches struggle with. It utilizes state-of-art optimization strategies to reduce computational time and memory cost of generalized mixed models. The computation cost linearly depends on sample size, and hence can be applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK Biobank data of 408,961 white British European-ancestry samples for 1,403 dichotomous phenotypes(Bycroft, et al., 2017; Sudlow, et al., 2015), we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.

Figure 1-1 Mixed model methods developed for population-based genetic association tests since 2006.

The methods in the top row are approximate methods and in the bottom row are exact methods. GMMAT(Chen, et al., 2016) is the only logistic mixed model method.



Chapter 2 Protein-altering and regulatory genetic variants near *GATA4* implicated in bicuspid aortic valve

2.1 Abstract

Bicuspid aortic valve (BAV) is a heritable congenital heart defect and an important risk factor for valvulopathy and aortopathy. Here we report a genome-wide association scan of 466 BAV cases and 4,660 age, sex, and ethnicity-matched controls with replication in up to 1,326 cases and 8,103 controls. We identify association with a non-coding variant 151kb from the gene encoding the cardiac-specific transcription factor, GATA4, and near-significance for p.Ser377Gly in *GATA4*. GATA4 was interrupted by CRISPR-Cas9 in induced pluripotent stem cells from healthy donors. The disruption of *GATA4* significantly impaired the transition from endothelial cells into mesenchymal cells, a critical step in heart valve development.

2.1 Introduction

Bicuspid aortic valve (BAV) is a congenital aortic valve defect characterized by fusion of two of the normal three leaflets. With a prevalence of ~1% in the population and a feature of some rare connective-tissue syndromes, BAV is the most common cardiovascular malformation in humans(Hoffman and Kaplan, 2002; Tutar, et al., 2005). BAV is associated with serious consequences: 30-70% of those with BAV will develop dilated thoracic aorta(Losenno, et al., 2012); 15 – 71% of BAV patients develop aortic valve stenosis depending on age group and individuals with BAV have a 50-fold higher risk of severe aortic valve stenosis(Ward, 2000); and up to 47% of BAV patients develop aortic valve incompetence(Siu and Silversides, 2010). The presence of a BAV confers an eight-fold increased risk of aortic dissection, which carries very high mortality(Michelena, et al., 2011). 27% of BAV patients will require surgical intervention to either replace their aortic valve or aorta for aortic aneurysm and dissection(Michelena, et al., 2008). BAV accounts for ~40% of the >50,000 aortic valve replacements (AVR) performed in the US each year(Roberts and Ko, 2005).

BAV is moderately heritable, with estimates ranging from 20 – 89%(Cripe, et al., 2004; Ellison, et al., 2007; Garg, 2006). Despite the prevalence, importance, and heritability of BAV, its genetic origins remain elusive. Previous genetic studies of BAV have focused primarily on linkage analysis in families(Ellison, et al., 2007; Martin, et al., 2007) or sequencing candidate genes in cases(Foffa, et al., 2013) under a hypothesis of Mendelian inheritance. Only one previous GWAS for BAV has been published in a limited number of cases (N=68)(Wooten, et al., 2010), which did not identify any genome-wide significant results. The only gene in which variants have been identified to cause BAV in multiple families is *NOTCH1*, but <6% of BAV cases are accounted for by *NOTCH1* variation(Foffa, et al., 2013). It is clear that BAV is not a simple Mendelian trait(Ellison, et al., 2007; McBride, et al., 2008), but is indeed heritable, and therefore, we applied genetic association methods typically used for complex traits.

With a goal of identifying genetic variants associated with BAV, leading to biological insight of the underlying causes, here we perform an unbiased genome scan in a large study of BAV cases (N=466) and controls (N=4,660), with replication in additional samples of up to 1,326 cases and 8,103 controls). We identify two genetic variants that reached or were near genome-wide significance levels ($P < 5x10^{-8}$). One is a low-frequency intergenic variant rs6601627 (odds ratio

(OR) = 2.38, $P_{after\text{-}replication}$=3x10^{-15}) with a substantially higher frequency in BAV cases (8.3%) than in controls (4.2%) and the other one is an independent association signal at a common protein-altering variant p.Ser377Gly (rs3729856) in *GATA4,* which encodes a cardiac-specific transcription factor that is 151 kilobases(kb) away from the first variant ($P_{after\text{-}replication}$ = 8.8x10^{-8}). Induced pluripotent stem cells (iPSCs) with *GATA4* disrupted by CRISPR-Cas9 demonstrate impaired transition of endothelial into mesenchymal cells (EndoMT), a critical step in valve formation (Lin, et al., 2012).

2.2 Results

To discover the underlying genetic basis of BAV, we successfully genotyped 498,075 genetic variants with enrichment of protein-altering variants (43.8% of variants examined) for 466 BAV cases and 4,660 controls. Imputation from the Haplotype Reference Consortium (HRC) panel(McCarthy, et al., 2016) enabled examination of a total of 12,320,487 variants. Clinical characteristics for BAV cases are summarized in Table S2-1. Following a genome-wide association scan(Figure S2-1 and Figure S2-2), we examined three variants in replication cohorts with a combined total of up to 1,326 additional cases and 8,103 controls.

2.2.1 Variants near *GATA4*

The strongest result from the genome-wide discovery for BAV was observed for a genotyped low-frequency variant, rs6601627, in an intergenic region of chromosome 8 (rs6601627, MAF = 4.1%, OR = 1.9, $P_{combined}$ = 3.0x10^{-15}) (Table 2-1, Figure 2-1, and Figure S2-2). 97 imputed variants in this region also reached genome-wide significance ($p < 5$ x10^{-8}). The two nearest genes are not obvious functional candidates (*CTSB* and *DEFB135*), however, this variant is 151kb from the 3' end of the *GATA4* gene (Figure 2-1). We also observed a common missense variant p.Ser377Gly in *GATA4* (rs3729856) that was also associated with BAV ($P_{discovery} = 3.2x10^{-4}$, MAF = 14.5%) (Table 2-1 and Figure 2-1). We selected *GATA4* p.Ser377Gly for *in silico* replication because it is a protein-altering variant and because it was located in the local genomic region of the most significant variant (**Error! Reference source not found.**). The p.Ser377Gly variant reached near genome-wide significance after including *in silico* replication data (OR = 1.31, P = $8.8x10^{-8}$) (Figure S2-4) and exceeded the typical significance level used for exome-wide studies of coding variation (typically $p < 2 \times 10^{-7}$)(Sveinbjornsson, et al., 2014). This suggests that *GATA4* may be the functional gene at this GWAS locus, but further experiments will be needed to demonstrate which gene(s) causes BAV. The two variants at 8p23.1 (rs6601627 and rs3729856) appear to be independent from each other, since they were not in linkage disequilibrium (LD $r^2 = 0.013$) and reciprocal conditional association analysis maintained nominal significance for both (P_{cond} rs6601627 = $8.92x10^{-9}$, P_{cond} rs3729856 = 0.012). After including *in silico* replication data, the reciprocal conditional association analysis still maintained nominal significance (P_{meta} rs6601627 = $1.52x10^{-9}$, P_{meta} rs3729856 = $8.17x10^{-3}$) (Figure S2-5). The non-additive association tests showed that both variants appear to have dominant effects on risk of BAV (Table S2-2).

The ExAC database characterizes protein-altering variants in 60,706 multi-ethnic individuals with whole exome sequences(Lek, et al., 2016). ExAC lists 96 missense variants in *GATA4* (95 of them have MAF < 1%), a deficit compared to the 140 variants predicted based on gene size. Additionally, 9.4 loss-of-function (LoF) variants are predicted and only 1 was observed (p.Lys365Ter) out of 60,706 individuals with deep exome sequences. The probability that the gene is intolerant to LoF, a measure of the relative importance of gene function, is high (pLI=0.8 where > 0.9 is considered extremely intolerant). Moreover, the missense *GATA4* variant rs3729856 is predicted as benign or tolerated by PolyPhen-2(Adzhubei, et al., 2013) and SIFT(Kumar, et al.,

2009) and it has a CADD score 9.418 (in top 11% of deleterious variants in the human genome)(Kircher, et al., 2014). These results suggest the importance of the *GATA4* gene's function, although the missense variant rs3729856 itself may not be significantly deleterious.

We hypothesized that the functional BAV gene at this *CTSB/GATA4* locus would demonstrate high expression in heart or vascular tissue. Using the GTEx portal(2013), we examined mRNA expression levels of all genes within the 200kb surrounding the non-coding associated variant rs6601627 and found that *GATA4* showed strong expression in heart (atrial appendage and left ventricle) and coronary artery, and also ovary, testis, pancreas, and liver (Figure S2-6). The other genes in the region (*NEIL2*, *FDFT1*, and *CTSB*) showed ubiquitous expression levels across all tissues (Figure S2-6). Examination of all GTEx association results did not identify any significant expression quantitative trait locus (eQTL) with the noncoding rs6601627 ($P < 10^{-5}$)((Bahcall, 2015)). We propose that this noncoding variant, or a variant tagged by it, influences *GATA4* expression in a manner not detectable by GTEx – either exerting an influence on gene expression levels only in the developing fetal heart or with a relatively modest effect that was not detectable in the current GTEx sample size.

After detecting association with both coding and non-coding variants at the *GATA4* locus, we sought to examine the role of GATA4 in the development of the aortic valve. In the primitive heart tube, heart valves develop from endocardial cushions, which are formed by mesenchymal cells derived from endothelial cells (ECs) through a process called EndoMT(Lin, et al., 2012). Despite the critical role in heart valve development, the mechanism of EndoMT is not well understood(de Lange, et al., 2004; Lincoln, et al., 2004; Wirrig and Yutzey, 2014). GATA4 was previously shown to be essential for heart formation (Kuo 1997 and Molkentin 1997) and for endocardial cushion development in mice(Rivera-Feliciano, et al., 2006). Here we evaluated the impact of disruption of GATA4 on human iPSCs differentiation into mesenchymal cells through EndoMT to examine the role of GATA4 in the development of aortic valves in humans.

The *GATA4* knock-out mouse is embryonic lethal between embryonic day (E) 7.0 and E9.5 and lacks a primitive heart tube(Molkentin, et al., 1997). Deletions of *GATA4* in humans have been associated with congenital heart defects (CHD)(Kennedy, et al., 2001; Pehlivan, et al., 1999) and a missense variant p.Gly296Ser was identified in a family with atrial and ventricular septal defects(Garg, et al., 2003). A mouse model of the p.Gly296Ser missense change is also embryonic lethal by E11.5 but a subset of these mice demonstrate semilunar valve stenosis and small defects of the atrial septum, thought to be resultant from defects in cardiomyocyte proliferation during embryogenesis(Sarkozy, et al., 2005). Previous studies observed missense variants in *GATA4* in patients with septal defects (Tomita-Mitchell, et al., 2007), congenital heart defects(Zhang, et al., 2008), and Tetralogy of Fallot (ToF)(Zhang, et al., 2008), but have not been tested in case-control models. The frequency of *GATA4* variants in healthy controls is not clear from these studies and their pathogenicity is unknown. The co-appearance of congenital heart disease and testicular anomalies was found in a family with a *GATA4* p.Gly221Arg mutation, thought to disrupt interaction with FOG2 and/or NR5A1, important factors for gonadal development(Lourenco, et al., 2011). *GATA5* has 46% homology with *GATA4*. *GATA5* sequence variants have been identified in humans with BAV(Bonachea, et al., 2014; Padang, et al., 2012) and *GATA5* knock-out mice and zebrafish demonstrate high rates of cardiac abnormalities(Laforest, et al., 2011).

2.2.2 Chromatin conformation at the *GATA4* locus

We attempted to evaluate the hypothesis that non-coding variants in LD with rs6601627 impact expression of *GATA4* during a critical stage of development. This is supported by prior evidence that GATA4 dosage impacts cardiac formation(Pu, et al., 2004). We first identified potentially functional variants using RegulomeDB and HaploReg in the region near rs6601627 or variants in high linkage disequilibrium (LD) ($r^2 > 0.6$) (ranging from rs112197605 to rs117851931; hg19:chr8:11774952-11838697)(Boyle, et al., 2012; Ward and Kellis, 2012). After examining local chromatin states, DNase hypersensitive regions and transcription factor binding sites, we identified rs118065347 as a variant likely to be functional because it co-localizes with binding regions for multiple transcription factors (including KAP1, CCNT2, CJUN, CMYC, GATA2, HDAC2, HMGN3, JUND, MAX, SP1, TAL1, YY1, ZBTB7A) and is in a known enhancer active in fetal heart, left and right ventricle, right atrium, as well as other tissues(Bernstein, et al., 2012; Boyle, et al., 2012). This variant disrupts the binding motif for a variety of transcription factors including PAX6(Piper, et al., 2013). There are other candidate functional variants in high LD with the index variant, and molecular experiments will be required to definitively identify the functional variant(s) and the mechanism of action on aortic valve development.

We next asked which genes in the locus may interact with this candidate enhancer region. We identified chromatin interaction loops in K562 and GM12878 cells using chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and high-throughput sequencing (Hi-C) data (Figure 2-2)(Phanstiel, et al., 2015; Rao, et al., 2014). rs118065347 falls near the edge of a topologically associated domain spanning from hg19:chr8:11250000-11825000 defined by Hi-C in both cell lines. The variant falls inside a ChIA-PET loop connecting to a region also annotated as an enhancer 3' of *GATA4* and C8orf49 and 5' of *NEIL2*. These data indicate that this distal region is brought in close proximity to *GATA4* and disruption of this region may have direct impact on *GATA4* expression. Further molecular experiments will be needed to clarify the gene(s) that impacts BAV.

2.2.3 Phenotypic characteristics of BAV cases in discovery sample

Among our 466 non-syndromic BAV cases, 93 (20%) reported one or more family members also having BAV (Table S2-1). This suggests a high recurrence risk and supports the hypothesis of large-effect variants, but not necessarily Mendelian inheritance(Pasta, et al., 2013). Majority of BAV cases were recruited from cardiac surgery clinic at the University of Michigan Frankel Cardiovascular Center (FCVC) where patients are referred to cardiac surgery for aneurysm repair or valve replacement, thus we found a high proportion of patients with thoracic aortic aneurysm (TAA) (83%). However, at these two loci, we saw no evidence for heterogeneity between BAV cases with or without TAA (Table S2-3) and between BAV cases with or without a positive family history of BAV and/or TAA (Table S2-4), suggesting that BAV probably impact the risk of TAA due to altered hemodynamic blood flow and aortopathy from different mechanisms instead of sharing molecular mechanisms with TAA that impact both aorta and valve tissue(Pasta, et al., 2013). Additionally, we did not find evidence for heterogeneity in the association results at *GATA4* and BAV subtypes (Table S2-5) and among males and females (Table S2-6).

2.2.4 Implication of rs6601627 and *GATA4* p.Ser377Gly in other CHD

To investigate whether the two variants at *GATA4* that we report are involved in development of other and more severe congenital heart defects, we tested for association with 806 cases of ToF along with 5,029 matched controls and performed association tests for the two variants as described previously(Cordell, et al., 2013). In an additive genetic model, we did not find evidence for association between the non-coding rs6601627 and ToF (MAF = 0.03, OR = 0.89, 95% CI 0.67-1.20, P=0.46), however, for *GATA4* p.Ser377Gly, the association was nominally significant (MAF = 0.11, OR = 1.24, 95% CI 1.06-1.45. P = 0.007). This suggests that the regulatory variant associated with BAV may act in a highly tissue or developmentally controlled manner to cause only BAV and not other congenital heart defects, whereas *GATA4* missense changes may have a more broad impact on other CHDs.

2.2.5 Missense variant in *DHX38*

The GWAS highlighted a rare missense variant (0.14% frequency in controls) in *DHX38* (p.Thr1221Met) with a strong association with BAV (OR=13.14, 95% CI 5.39-32.04, P = $1.5x10^{-8}$) in the discovery sample (Figure S2-2). After genotyping of this variant in 720 cases and 5,831 controls, only 22 copies of the rare allele were identified (5 in cases and 17 in controls), providing a replication p-value of 0.05. Additional large studies will be needed to confirm this rare variant association with BAV.

2.2.6 GATA4 deficiency impairs EndoMT in iPSC-derived cells

We investigated the biological impact of GATA4 in the EndoMT process required for human valve formation. Human induced pluripotent stem cells (iPSCs) were generated from peripheral blood mononuclear cells of a donor with normal tri-leaflet aortic valve, using non-integrated DNA vectors containing *OCT4*, *SOX2*, *C-MYC*, and *KLF4*(Su, et al., 2013). The pluripotency of iPSCs was confirmed by expression of OCT4, SOX2, NANOG, and SSEA4, TRA-1-60, and TRA-1-81 (Figure S2-7A-B). Additionally, iPSCs generated teratoma containing tissues from three germ layers, demonstrating their pluripotency in vivo (Figure S2-7C). In a previous study, wildtype GATA4 localized completely in the nucleus, whereas GATA4 mutant p.Ser377Gly (a C-terminal mutant) was shown to be partially distributed to the cytoplasm, indicating a loss-of-function mutation(Wang, et al., 2013). To evaluate whether disruption of *GATA4* may result in a loss-of-function phenotype, iPSCs were electrotransfected with plasmid containing Cas9, *GATA4* single guide RNA (sgRNA), and green fluorescent protein (GFP) as an indicator for transfection(Ran, et al., 2013). As control, iPSCs were transfected with plasmid containing Cas9 and GFP. Transfected cells were enriched by flow cytometry sorting based on GFP positivity (Figure S2-8A-C). iPSCs were differentiated into ECs with efficiency above 90% (Figure S2-8D). GATA4 level was significantly lower in ECs from the *GATA4* sgRNA transfected group than control (Figure 2-3A), indicating successful targeting to *GATA4*. When EndoMT was induced by TGF2 and BMP2 in ECs, smooth muscle actin (SMA), a mesenchymal marker gene was upregulated in control cells (Figure 2-3B). Noticeably, the *GATA4* sgRNA group showed significantly lower SMA levels (Figure 2-3B). ECs were also explanted to collagen gel to induced EndoMT(Rivera-Feliciano, et al., 2006). The *GATA4* sgRNA group showed significantly fewer mesenchymal cells migrating out after 3 days than control cells (Figure 2-3C). Cells undergoing EndoMT express SMA and CD31 simultaneously at a certain point(Rivera-Feliciano, et al., 2006). Immunofluoresence staining of SMA and CD31, markers of EndoMT(Rivera-Feliciano, et al., 2006), also showed significantly less SMA and CD31 double positive cells in *GATA4* sgRNA group (Figure 2-3D).These results indicate EndoMT was impaired by disruption of *GATA4* with *GATA4* sgRNA.

2.3 Discussion

In this study we find variants associated with bicuspid aortic valve (BAV) that reach genome-wide significance. We identified association with a low frequency non-coding variant 151kb from *GATA4*, as well as a common missense variant in *GATA4.* Although we cannot yet confirm the mechanism of action of the non-coding variant(s) on chromosome 8 on aortic valve development, chromatin conformation experiments suggest that the region near the associated variants appears to loop and physically interact with regions intronic to *GATA4*. This hypothesis could be tested in future functional experiments to investigate whether the non-coding BAV-associated variants identified here affect expression of *GATA4* at a critical time in heart development. This could possibly disrupt EndoMT, a process important for normal trileaflet aortic valve formation.

GATA4, a zinc finger transcription factor, is one of three major transcription factors, together with Nkx2.5 and TBX5, that are critical for heart differentiation(Huang, et al., 1995). Although not previously associated with BAV, the *GATA4* gene is a plausible biological candidate. The missense *GATA4* mutation G296S disrupts the transcriptional cooperativity between GATA4 and TBX5, resulting in abnormal cellular functions related to morphogenetic defects(Ang, et al., 2016). Many mutations in *GATA4* have been previously reported to be found in different kinds of CHD: atrial septal defect(Garg, et al., 2003; Hirayama-Yamada, et al., 2005; LaHaye, et al., 2016; Mattapally, et al., 2015; Posch, et al., 2008; Sarkozy, et al., 2005; Tomita-Mitchell, et al., 2007; Yang, et al., 2013), ventricular septal defect(Garg, et al., 2003; Mattapally, et al., 2015; Tomita-Mitchell, et al., 2007; Wang, et al., 2011; Yang, et al., 2012), and ToF(Mattapally, et al., 2015; Tomita-Mitchell, et al., 2007; Yang, et al., 2013), although mostly tested in family studies. Furthermore, the *GATA4* mutations have been identified in CHD patients with various ancestries: European(Garg, et al., 2003; Posch, et al., 2008; Sarkozy, et al., 2005; Tomita-Mitchell, et al., 2007), Asian(Garg, et al., 2003; Hirayama-Yamada, et al., 2005; Mattapally, et al., 2015; Wang, et al., 2011; Yang, et al., 2013; Yang, et al., 2013; Yang, et al., 2012), and Native and Hispanic American(Tomita-Mitchell, et al., 2007). Additionally, *GATA4* knockout mice are embryonic lethal with heart defects(Kuo, et al., 1997). Mice that are missing *GATA5* also develop BAV(Laforest, et al., 2011) and rare *GATA5* mutations have been identified in humans with BAV (Bonachea, et al., 2014).

We have provided evidence for the complexity of the BAV phenotype, with multiple genetic variants of incomplete penetrance contributing to susceptibility. To assess whether the two variants that we report are specific for BAV or whether they are also implicated in other congenital heart defects, we studied cases of ToF, characterized by several cardiac malformations including an overriding aorta, pulmonic stenosis, ventricular septal defect, and right ventricular hypertrophy. We found that the common coding variant in *GATA4* (p.Ser377Gly) was associated with increased risk of ToF whereas the low frequency non-coding variant (rs6601627) was not associated. We speculate that the low frequency non-coding variant disrupts a regulatory element that plays a critical role in regulating *GATA4* expression in a precise time of cardiac embryogenesis that may impact the valve more specifically, whereas the common *GATA4* missense variant might disrupt GATA4 function more generally and increase the risk of several cardiac malformations, including ToF. The frequencies of the associated variants at the *GATA4* locus (variants with $r^2 > 0.6$ in EUR samples of 1000G) vary among different populations(Auton, et al., 2015). For example, among the non-coding variant rs6601627 and its 115 correlated variants, 108 variants have MAF < 0.01 and 73 are monomorphic in East Asians(Auton, et al., 2015). Association studies in other populations will be critical for determining if the association exists in other populations and may be helpful at narrowing the associated interval.

To investigate the possible role of GATA4 in aortic valve development, we used sgRNA guided Cas9 to disrupt *GATA4* in iPSCs from a healthy human donor with normal tricuspid aortic valves. The iPSCs were differentiated into endothelial cells and then induced to mesenchymal cells through EndoMT. We demonstrated that deficiency of GATA4 impaired the transition of endothelial into mesenchymal cells, a critical step in valve formation(Lin, et al., 2012) (Figure 2-3). This indicates that GATA4 is required for aortic valve formation and that disruption of the *GATA4* gene, either by non-coding or protein-altering variants, may affect aortic valve formation.

2.4 Methods

2.4.1 GWAS genotyping and genotype imputation

We performed genotyping of a combined set of 498,075 genome-wide association scan (GWAS) variants, including 217,957 protein-altering variants, using a GWAS+exome chip array (Illumina Human CoreExome). To avoid any potential batch effects, cases and controls were genotyped using the same array in the same genotyping center (Sequencing and Genotyping core at the University of Michigan). Genotype calling was performed using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina) using identical cluster files for cases and controls. Samples with <98% genotype calls, evidence of gender discrepancy, duplicates as well as individuals with non-European ancestry identified by plotting the first 10 genotype-driven principal components were excluded from further analysis. We performed variant-level quality control by excluding 22,983 variants that met any of the following criteria; variants with a cluster separation score < 0.3, < 98% genotype call rate, or deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-5}$). We phased the autosomal genotype data using SHAPEIT2(Delaneau, et al., 2012) and imputed variants from the HRC v1 reference panel(McCarthy, et al., 2016) using minimac3(Fuchsberger, et al., 2015). We excluded poorly imputed variants with imputation $R^2 < 0.3$ and then merged the genotyped variants and the successfully imputed variants to a combined data set, which contains 12,320,487 variants in total.

2.4.2 Description of cases in discovery cohort

We collected DNA from consented individuals with bicuspid aortic valve from the Frankel Cardiovascular Center at the University of Michigan as part of the University of Michigan BAV registry or the Cardiovascular Health Improvement Project (CHIP). All repository projects utilized for this study are approved by the University of Michigan, Medical School, Institutional Review Board (IRBMED), and informed consent was obtained from study participants. Patients were typically seen in clinic for aortic valve replacement or aortic aneurysm. Diagnoses of bicuspid aortic valve were made by cardiac surgeons upon visual inspection of the aortic valve during open surgery for aneurysm repair or valve replacement. BAV cases with major syndromic connective-tissue disorders (e.g. Marfan syndrome) were excluded. DNA was isolated from peripheral blood lymphocytes.

2.4.3 Description and selection of controls in discovery cohort

We identified potential controls from a surgical-based biobank, the Michigan Genomics Initiative (MGI), that were genotyped with the same GWAS array (Illumina Human CoreExome). After excluding those with possible aortic disease (N=1,586, Table S2-7), we were left with 15,642 potential controls with GWAS data. We performed age matching by requiring controls to have a birth year within -5 and +10 years of the case. From the available controls in the appropriate age and sex category, we selected the best ethnic match for each case and repeated the greedy algorithm until a control was selected for each case. We repeated the entire process so that 10 controls were selected for each case. We opted for this approach to provide the best ancestry matching between cases and controls, to reduce the potential for false positives due to ethnicity mismatch, and to also provide the most power for rare variants that increase risk of BAV by including the highest number of matching controls. All MGI research subjects provided informed consent.

2.4.4 Statistical analyses

In the discovery cohort, we performed association testing for BAV status using logistic regression with single genetic variants (295,759 with MAF > 1%), with age, sex and the first four principal components as covariates using PLINK for hard call genotypes(Purcell, et al., 2007) and the EPACTS software (URL: http://csg.sph.umich.edu//kang/epacts/) for imputed dosages. We identified two genetic variants that were directly genotyped and reached genome-wide significance levels ($P < 5x10^{-8}$). We observed no evidence for inflation due to population stratification ($\lambda = 1.033$, Figure S2-1). We observed a genotyped missense variant within 200kb (rs3729856) of one of the significant non-coding variants (rs6601627) and selected this third variant for follow-up in additional samples.

2.4.5 Association with Tetralogy of Fallot (ToF)

A total of 835 unrelated ToF cases and 5159 controls were genotyped and imputed from 1000 Genomes Phase 3 for the region 11MB-12MB on chromosome 8 using IMPUTE2(Cordell, et al., 2013; Howie, et al., 2009). The association tests were performed using logistic regression of the "best-guess" genotypes for all imputed SNPs with IMPUTE2 info score ≥ 0.5 and with MAF ≥ 0.01 in controls using SNPTEST(Marchini, et al., 2007). This study has been approved by Newcastle and North Tyneside NHS Research Ethics Committee.

2.4.6 Gene expression, chromatin conformation and epigenetics data

We assessed expression levels of relevant genes using the GTEx server (http://www.gtexportal.org/home/)(2013). We obtained the Hi-C interaction calls from Rao *et al.*(Phanstiel, et al., 2015; Rao, et al., 2014) and ChIA-PET interactions from Phanstiel *et al.*(Phanstiel, et al., 2015; Rao, et al., 2014) available through the ENCODE DCC accessions ENCSR000FDB and ENCSR752QCX (https://www.encodeproject.org/). ChromHMM data are displayed from the K562 Genome Segmentation by ChromHMM from ENCODE/Analysis available at https://genome.ucsc.edu/ and is created through chromatin segmentation using 8 histone modifications, CTCF, Pol2, and open chromatin annotations(Ernst and Kellis, 2012).

2.4.7 IPS cells generation and culture

The procedure of iPSC derivation was performed according to methods we described (Su, et al., 2013). PBMCs were separated from human peripheral blood with LSM (MP Biomedicals LLC.), cultured in medium containing IMDM (Life technologies Corp.), 10% FBS (Life technologies Corp.), TPO， SCF， FLT-3 at final concentration 100ng per mL, G-CSF， IL-3 at final concentration 10ng per mL (Peprotech Inc.), penicillin-streptomycin (Life technologies Corp.) and electrotransfected using Nucleofector 2 device (Lonza Corp.) with episomal DNA plasmids containing *OCT4, SOX2, KLF4*, and *C-MYC*. At around day 30 post-infection, the colonies became compact. The colonies were mechanically picked up from the culture dishes and firstly cultured with mouse embryonic fibroblasts for 3 passages(Jiao, et al., 2013) and transited to TesRE8 medium (Stemcells Inc.) on matrigel-coated (BD Corp.) dishes. iPSCs were passaged every 4 to 6 days with Versene (Life Technologies Corp.). And iPSCs from passage 25 to passage 35 were used in experiments.

2.4.8 Teratoma formation in immune-deficient mice

Conduction of Animal experiments was in compliance to regulations of the Unit for Laboratory Animal Medicine (ULAM) at the University of Michigan. Two million iPSCs were injected subcutaneously into each flank of the recipient male, 6-to-8-weeks-old NOD-SCID mice (Jackson Laboratory, Bar Harbor, Maine). 3–5 weeks after injection, teratomas were harvested from the mouse flanks and fixed with formalin (Thermo Corp.) for 2 days. Then the tumors were imbedded in paraffin and sections were prepared with microtome (Leica Corp.) and stained by H&E staining solutions from Thermo Corp. The slides were examined and photos were taken under brightfield with microscope (Nikon Corp.).

2.4.9 *GATA4* sgRNA design and electrotransfection of iPSCs

sgRNA were designed to target *GATA4* exon2 (the first coding exon) with sgRNA design tool (http://www.genome-engineering.org) developed by Dr. Feng Zhang group(Ran, et al., 2013). Sequence of *GATA4* sgRNA was: CGCGCCGTGCATGAAGGCGCCGG. Target site was: chr8:-11565888. Quality score was 93. Minimal number of mismatch nucleotides in offsite targets was 3. SgRNA were cloned into PX458, which contains SpCas9-2A-EGFP using AgeI and EcoRI at 5' and 3' cloning sites(Ran, et al., 2013). One million iPSCs were electrotransfected with constructed 5μg PX458 containing *GATA4* sgRNA, using Lonza Human Stem Cell Nucleofector® Kit 2 with program U-023 on Nuclefector 2 device (Lonza Ltd.). Another one million iPSCs were electrotransfected with PX458 vector as control under the same conditions.

2.4.10 Endothelial cell differentiation from iPSCs

To differentiation iPSCs into ECs, iPSCs were dissociated with Versene (Life Technologies Corp.) into single cells and seeded at 2x10^4 cells per cm^2 with TesRE8 (Stemcell technology Inc.) medium supplemented with Rocki (Y27632, Stemgent Inc.). When the cells reached a confluence of 20% to 30%, medium was changed into differentiation medium, which contained DMEM-F12 (Life technologies Corp.), B27 supplement without vitamin A (Life technologies Corp.), L-glutamine (Life technologies Corp.), penicillin-streptomycin (Life technologies Corp.), 400 µM 1-thioglycerol (Sigma Corp.). 50 µg per mL Ascorbic acid (Sigma Corp.), 25 ng per mL BMP4 (R&D Systems Corp.), and 6 µM GSK3 inhibitor CHIR99021 (Sigma Corp.). Differentiation medium was refreshed daily for 3 days. Then cells were dissociated with Accutase (Life technologies Corp.) and seeded at 1x10^4 cells/cm^2 on Matrigel (BD Corp.) coated dishes with endothelial cell medium containing Stempro34(Life technologies Corp.), Stempro34 supplement (Life technologies Corp.), L-glutamine (Life technologies Corp.), penicillin-streptomycin(Life technologies Corp.), and 50 ng per mL VEGF (Peprotech Inc.). Medium was refreshed every two days for 13 days.

2.4.11 Immunofluorescence staining and flow cytometry

Immunofluorescence staining and flow cytometry was performed as follows, firstly cells were fixed in 4% formaldehyde (Thermo Corp.) for 1 hour at room temperature, then the cells were washed with DPBS (Thermo Corp.) once and incubated with primary antibodies for 2 hours at room temperature(Jiao, et al., 2013). The following primary antibodies were used: anti-OCT4 (mouse IgG, dilute 500 hundred times upon usage, sc-5279, Santa Cruz Biotechnology Inc.), anti-SOX2 (mouse IgG, dilute 500 hundred times upon usage, sc-365964, Santa Cruz Biotechnology Inc.), anti-NANOG (rabbit polyclonal, dilute 500 hundred times upon usage, REC-RCAB004PF, Cosmo Inc.), anti-SSEA4 (mouse IgG, dilute 100 hundred times upon usage, 60062, Stemcell technology Inc.), anti-TRA-1-60 (mouse IgM, dilute 100 hundred times upon usage, 60064, Stemcell technology Inc.), anti-TRA-1-81 (mouse IgM, dilute 100 hundred times upon usage, 60065, Stemcell technology Inc.), anti-CD31 (rabbit polyclonal, dilute 500 hundred times upon usage, ab28364,Abcam Inc.), anti-SMA (mouse IgG, dilute 1000 hundred times upon usage, A5228, Sigma, Corp.). Cells were washed three time with DPBS (Thermo Corp.), then incubated with secondary antibodys for 1 hour at room temperature. The following fluorochrome-conjugated secondary antibodies were used: Alexa Fluor 488 goat anti-rabbit IgG (goat, dilute 1000 times upon usage, A11034, Thermo Corp.), and Alexa Fluor 488 goat anti-mouse IgG (goat, dilute 1000 times upon usage, A32723, Thermo Corp.), Alexa Fluor 594 goat anti-mouse IgG (goat, dilute 1000 times upon usage, A11032, Thermo Corp.) Slides were mounted with anti-fade mounting media containing DAPI (Prolong gold, Life technologies Corp.), and were observed on a Nikon A1 confocol microscope (Nikon Corp.). In flow cytometry study, electrotransfected iPSCs were dissociated into single cells with Accutase (Stemcell technology Inc.), and applied to MoFlo Astrios (Beckman Coulter Inc.) flow cytometry machine.

2.4.12 Western blot analysis

Whole cell extracts were prepared using RIPA buffer (1% NP-40, 1% sodium deoxycholate, 0.1% SDS, 0.15 M NaCl, 0.01 M sodium phosphate, 2 mM EDTA, 50 mM sodium fluoride, 0.2 mM $Na_3VO_4.2H_2O$, 100 U per mL protease inhibitor), resolved on SDS-PAGE gels, and transferred to acetate cellulose membranes. Primary antibodies used were anti-GATA4 (rabbit IgG, diluted 500 times upon usage, 36968, Cell Signaling technology Inc.), anti-SMA (mouse IgG, diluted 300 times upon usage, A5228 Sigma), anti-GAPDH (rabbit IgG, diluted 2000 times upon usage, sc25778, Santa Cruz Inc.). Secondary antibodies used were IRDye800CW Donkey anti-Mouse

(92532212), IRDye680LT Donkey anti-Rabbit (92568023), IRDye800CW Donkey anti-Rabbit (92532213) (All secondary antibodies were diluted 500 times upon usage and purchased from Licor Inc.). Licor western blot detection system was used for the dual-color imaging. Uncropped version of western blot is presented in Figure S2-9 and Figure S2-10. Image was used for quantification of bands. Each band was normalized by GAPDH. Experiments were repeated three times. Average value and standard derivation were plotted.

2.4.13 Endothelial-to-mesenchymal transition and collagen gel assay

EndoMT was induced by changing medium to EndoMT inducing medium which contained stempro34 medium with stempro34 supplement (Life science technology Corp.), L-glutamine (Life technologies Corp.), penicillin-streptomycin (Life technologies Corp.), 200ng per mL BMP2 (Peprotech Inc.) and 50ng/mL TGFβ2 (Peprotech Inc.). Non EndoMT control group were kept in EC medium containing Stempro34 (Life technologies Corp.), Stempro34 supplement (Life technologies Corp.), L-glutamine (Life technologies Corp.), penicillin/streptomycin (Life technologies Corp.), and 50 ng/mL VEGF (Peprotech Inc.). Cells were harvested three days after induction.

Type I collagen (Sigma Corp.) at 1mg/mL (final concentration) were mixed with stempro34 medium, stempro34 supplement (Life science technology Corp.), and 50mM NaOH (Sigma Corp.). The mixture was poured into 24-well tissue culture plates (0.5 mL per well) and allowed to gel in 5% CO_2 incubator at 37°C for 30 minutes. And then 0.5 mL EndoMT inducing medium was added. After 3 days, pictures of cells were taken with 100 times magnificence under Eclipse Ti-U inverted research microscope (Nikon Corp.). Mesenchymal cells which migrated out in three pictures from different field were counted. Experiments were repeated three times. Average value and standard derivation were plotted.

2.4.14 Genetic association replication cohorts

CHIP: In the CHIP replication cohort, an additional 140 BAV cases from the University of Michigan FCVC biobank were collected. These samples were genotyped using the same GWAS array as the discovery cohort but only examined for the three variants described here. The association was tested using PLINK(Purcell, et al., 2007) with 1,400 age, sex, and ancestry-matched controls from the MGI study, which were independent samples from previously used controls. Informed consent was obtained from all participants and approval was obtained from the Institutional Review Board of the University of Michigan Medical School.

MHI: In the Montreal Heart Institute (MHI) biobank, 305 BAV cases and 2746 controls were collected and genotyped on the Illumina Core Exome array at the MHI Pharmacogenomic Centre. Controls were selected by excluding those with MI, PCI, Angina, CHF, valve defects, heart surgeries, heart arrest, atrial fibrillation and sudden cardiac death. Genotyping was performed with the Illumina HumanExome array. Association analysis was performed in PLINK(Purcell, et al., 2007) using a logistic regression model correcting for sex, age, and principal components of ancestry 1-10. The project has been approved by the Ethics Committee of the Montreal Heart Institute and informed consent was obtained from study participants.

Partners HealthCare: In the Partners HealthCare cohort, 452 Caucasian BAV cases were identified from the electronic medical records of Partners HealthCare (Boston, MA). Individual echocardiographic images were reviewed to confirm BAV diagnosis. Whole blood DNA was genotyped using the Illumina Omni2.5 Beadchip. The Framingham Heart Study dbGaP cohort, genotyped using the Illumina Omni5.0 Beadchip, was used as controls. Quality control and population stratification of the genotype data were performed in PLINK(Purcell, et al., 2007). SNPs with MAF less than 1%, without physical map reference, not in Hardy-Weinberg equilibrium ($P<10^{-4}$), differential missingness ($P<10^{-5}$), were removed. Related individuals (PI_HAT > 0.25) were excluded. Genome wide IBD & IBS were used to detect outliers and clusters. After merging case and control genotypes, additional genotypes have been imputed against the 1000 Genome reference (phase3) and HRC (Michigan University) panels using SHAPEIT2(Delaneau, et al., 2012) and IMPUTE2(Howie, et al., 2009). After QC, 452 cases and 1634 controls (1094 males + 992 females) with 7.5 million markers were analyzed using an additive logistic regression model accounting for gender, age and principal components. This study has been approved by Partner's HealthCare Human Research Committee and informed consent was obtained from study participants.

University of Texas Health Science Center: In the UTHSC cohort, 765 patients with sporadic thoracic aortic aneurysms or aortic dissections (TAAD) were collected and genotyped. 874 genotypes from dbGAP (NINDS Neurologically Normal control collection) were used as controls. QC and population stratification of the genotype data were performed in PLINK (Purcell, et al., 2007). SNPs with MAF less than 1% or missing more than 1% of genotypes were excluded. Multidimensional scaling was used to detect and exclude population outliers. We imputed additional genotypes against 1000 Genomes Phase3 using SHAPEIT2(Delaneau, et al., 2012) and IMPUTE2(Howie, et al., 2009). After QC, a total of 152 BAV cases and 633 tricuspid aortic valve (TAV) cases or 874 controls were analyzed using an additive logistic regression model accounting for gender and principal components. This study has been approved by the Committee for the Protection of Human Subjects at UT Health Science Center at Houston and informed consent was obtained from study participants.

ASAP-Artist-Polca-Olivia cohort: Three cohorts ASAP, ARTIST and POLCA were included in this replication group with a total of 275 BAV cases and 1,686 controls used for analysis. The POLCA/Olivia cohort is a merged cohort with a total of 1,295 individuals. The ASAP cohort consists of 429 patients genotyped on Illumina 610wQuad beadchips. Approximately 588,400 SNPs were provided after quality control. The Artist cohort consists of 406 samples genotyped with Omni-2.5 Quad beadchips on 2,443,180 SNPs. In POLCA, 625 control samples were genotyped on Illumina 610kwQuad and in Olivia, 670 control-samples were genotyped on illumina 1M genotyping arrays. The vast majority of included samples are of Scandinavian ancestry. For the ASAP database, where ancestry is specifically registered, this corresponds to >95% of the individuals, supported by PCA plots of genotype clustering. Imputation was performed using Impute2 from 1000G phase1 v3(Howie, et al., 2009). Analysis was performed using SNPTEST(Marchini, et al., 2007), with age, sex and first 10 principal components as covariates. This study was approved by Regional ethical committee of Stockholm and informed consent was obtained from study participants.

Bio*Me:* The Mount Sinai Bio*Me* Biobank (Bio*Me*) is an ongoing, prospective, hospital- and outpatient- based population research program operated by The Charles Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai and has enrolled over 33,000 participants since September 2007. Bio*Me* is an Electronic Medical Record (EMR)-linked biobank that integrates research data and clinical care information for consented patients at The Mount Sinai Medical Center, which serves diverse local communities of upper Manhattan with broad health disparities. Bio*Me* populations include 25% of African ancestry (AA), 36% of Hispanic Latino ancestry (HL), 30% of white European ancestry (EA), and 9% of other ancestry. The Bio*Me* disease burden is reflective of health disparities in the local communities. Bio*Me* operations are fully integrated in clinical care processes, including direct recruitment from clinical sites waiting areas and phlebotomy stations by dedicated recruiters independent of clinical care providers, prior to or following a clinician standard of care visit. Recruitment currently occurs at a broad spectrum of over 30 clinical care sites. Information on BAV status, age, and sex was derived from participants' EMRs. BAV cases were defined as Bio*Me* participants with the ICD-9 code 746.4 (Congenital insufficiency of aortic valve). In total, there were 41 BAV cases with available genotyping data (8 AA and 13 HL BAV cases genotyped on the Infinium Multi-Ethnic Global (MEGA) BeadChip from Illumina as well as 13 additional HL and 7 EA BAV cases genotyped on the Illumina HumanOmniExpressExome-8 v1.0 BeadChip. For each case, three controls were selected by genetically matching using the first two genetic principal components and stratification by age and sex. Logistic regression was performed in PLINK for the 3 SNPs in the 4 groups(Purcell, et al., 2007). We performed analyses both including and excluding the BioME non-European samples and results were highly similar. We present results in this study excluding the non-European samples since there were few cases. This study has been approved by Icahn School of Medicine IRB and informed consent was obtained from study participants.

2.4.15 Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. Acknowledgements

We thank all participants of the Cardiovascular Health Improvement Project (CHIP) and Bicuspid Aortic Valve registry at the University of Michigan for their contribution to research. We appreciate the valuable efforts from the CHIP and BAV registry collection team (WH, AD, AL, MT, LF, Jennifer McNamara, Joyvina Evans, Daniel Ferman, Neelima Goyal, Joanna Hider, Tom Jones, John Kruszewski, Brian Kulick, Jamie Love-Nichols, Joe Luciano, Alison Lunau, Phillip Nulty, Kevin Packard, Ian Patterson, Bailey Pearce, Mike Ranella, Alexander Shevchik, Carolyn Sommer, Christina White). The University of Michigan Health System – Cardiovascular Health Improvement Project was supported by the Frankel Cardiovascular Center. We appreciate the Aikens Fund for Aortic Research (CJW, BY) and McKay research award for supporting this project (BY). BY is supported by American Association of Thoracic Surgery (AATS) Graham Foundation and Thoracic Surgery Foundation of Research and Education (TSFRE). CJW is supported by HL109946, HL130705 and HL127564. SKG is supported by Doris Duke Charitable Foundation, grant # 2013104 and HL122684. JBN is supported by grants from The Danish Heart Foundation, the Lundbeck Foundation, and the A.P Møller Foundation for the Advancement of Medical Science.

We thank all participants and staff of the André and France Desmarais Montreal Heart Institute (MHI) Biobank, and in particular John D. Rioux and Simon de Denus who participated in the design of the study. The MHI replication study was funded by the MHI Foundation and the Canada Research Chair Program.

We thank all participants and physician investigators of the International Bicuspid Aortic Valve Consortium (BAVCon) for their scientific collaboration. Investigators of the International Bicuspid Aortic Valve Consortium (BAVCon): Eduardo Bossone, Azienda Ospedaliera Universitaria Salerno; Rodolfo Citro, Azienda Ospedaliera Universitaria Salerno; Stefano Nistri, CMSR Veneto Medica, Vicenza; Dan Gilon, Hadassah-Hebrew University Medical Center; Ronen Durst, Hadassah-Hebrew University Medical Center; Simon Body, Harvard University;

Thoraf M. Sundt, Harvard University; J. Daniel Muehlschlegel, Harvard University; Carlo de Vincentis, IRCCS Policlinico San Donato; Francesca R. Pluchinotta, IRCCS Policlinico San Donato; Hector I. Michelena, Mayo Clinic; Maurice Sarano, Mayo Clinic; Bo Yang, University of Michigan; Kim Eagle, University of Michigan; Cristen J. Willer, University of Michigan; Giuseppe Limongelli, Monaldi Hospital; Malenka M. Bissell, Oxford University; Alessandro DellaCorte, Second University of Naples; Amalia Forte, Second University of Naples; Gordon Huggins, Tufts University; Victor Dayan, Universidad de la República, Uruguay; Yohan Bossé, Université Laval; Evaldas Girdsaukas, University of Hamburg; Ashutosh Hardikar, University of Tasmania; Thomas Marwick, University of Tasmania; Joseph Bavaria, University of Pennsylvania; Rita C. Milewski, University of Pennsylvania; Dianna M. Milewicz, University of Texas Health Science Center at Houston; Siddarth K. Prakash, University of Texas Health Science Center at Houston; Arturo Evangelista, Vall d'Hebron University Hospital; Joshua C Denny, Vanderbilt University; Edward Hulten, Walter Reed National Military Medical Center. The BAVCon efforts at Partners HealthCare (Boston, MA) are supported by HL114823 to SCB.

The ASAP study was supported by the Leducq Foundation (MIBAVA).

The Mount Sinai Bio*Me* Biobank is supported by The Andrea and Charles Bronfman Philanthropies. Bio*Me* is partially funded by a grant from NIH (NHGRI U01HG007417). Analyses of Bio*Me* data was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the

NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Additional funding was provided to the Broad Institute (HHSN268201000029C), the van Andel Institute (10ST1035, HHSN261200800001E), University of Miami (DA006227, DA033684, N01MH000028), the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 11/16/2015.

2.5 Author contributions

CJW and BY designed the study. CJW, BY, WZ, and JJ drafted the manuscript. BY, JBN, MM, MH, GL, LF, SP, CS, LF, ML, WH, AD, AL, MT, LF, M-PD, EMI, AF-C, D-CG, EPB, GMD, AB, SK, HC, BK, JG, SG, and KE characterized phenotypes of samples. MO, JK, YEC, and BY contributed to the discussion of biological mechanisms. WZ, JBN, MH, GL, LF, SP, CS, LF, ML, HC, and HMK performed the statistical analysis of association data. GAF and APB performed the chromatin conformation analysis. JJ and BY performed experiments of CRISPR-Cas9 on iPSCs and experiments of EndoMT. RJFL, PE, J-C T, CMB, DM, SCB, CJW, and BY are Principal Investigators of cohorts.

Figure 2-1 Regional association plot of the chr8 association region near *GATA4*, as observed in the discovery cohort

Genome-wide single variant association tests were performed on 466 BAV cases and 4,660 controls. The upper panel shows all variants that were directly genotyped in the chip array in this region. A missense variant (rs3729856, p.S377G) within *GATA4* was observed to be associated with BAV with P = 3.2x10^{-4}, that reached P = 8.8x10^{-8} following replication in 1,326 BAV cases and 8,103 controls. The bottom panel shows results after genotypes imputed from the HRC reference(McCarthy, et al., 2016). Coding variants are represented by triangles and noncoding variants are represented by squares.

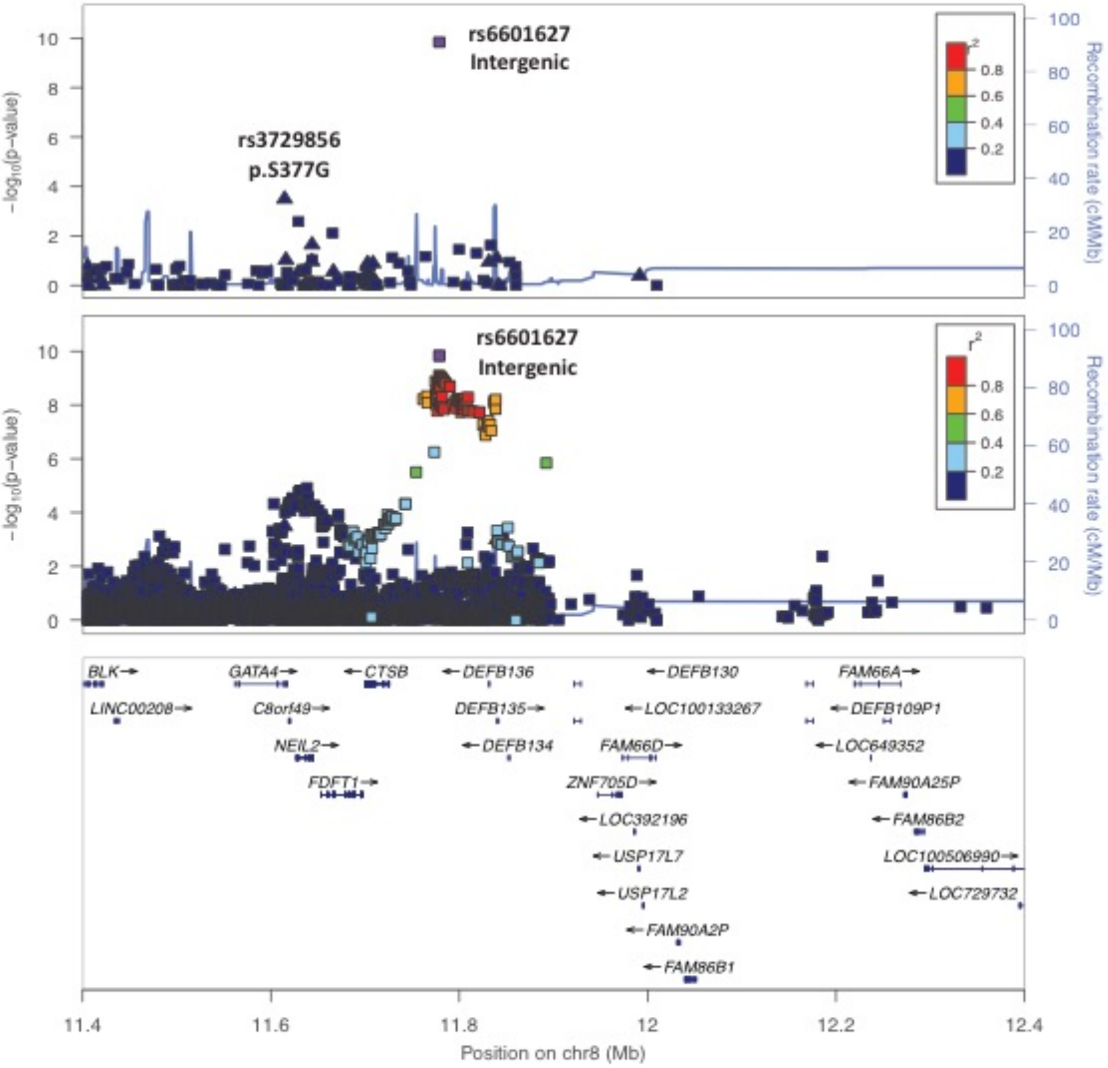


Figure 2-2 Chromatin interactions between associated variants and *GATA4*

The topological domain region containing associated variants (orange lines), genes (see inset for annotation descriptions), chromatin interactions by Hi-C (blue) and ChIA-PET (purple), and chromatin state (outer ring and standard colors from(Ernst and Kellis, 2012) but of significance yellow as enhancers, red as promoters, green as transcribed, blue as CTCF, and grey as inactive. All data are from K562 cells. rs3729856 is indicated as falling within a coding exon of *GATA4*. rs6601627 was identified as the associated variant to BAV and rs118065347 is the putative functional variant in linkage. rs11865347 overlaps an annotated enhancer as well as a ChIA-PET loop connecting to a region 3’ of *GATA4*.

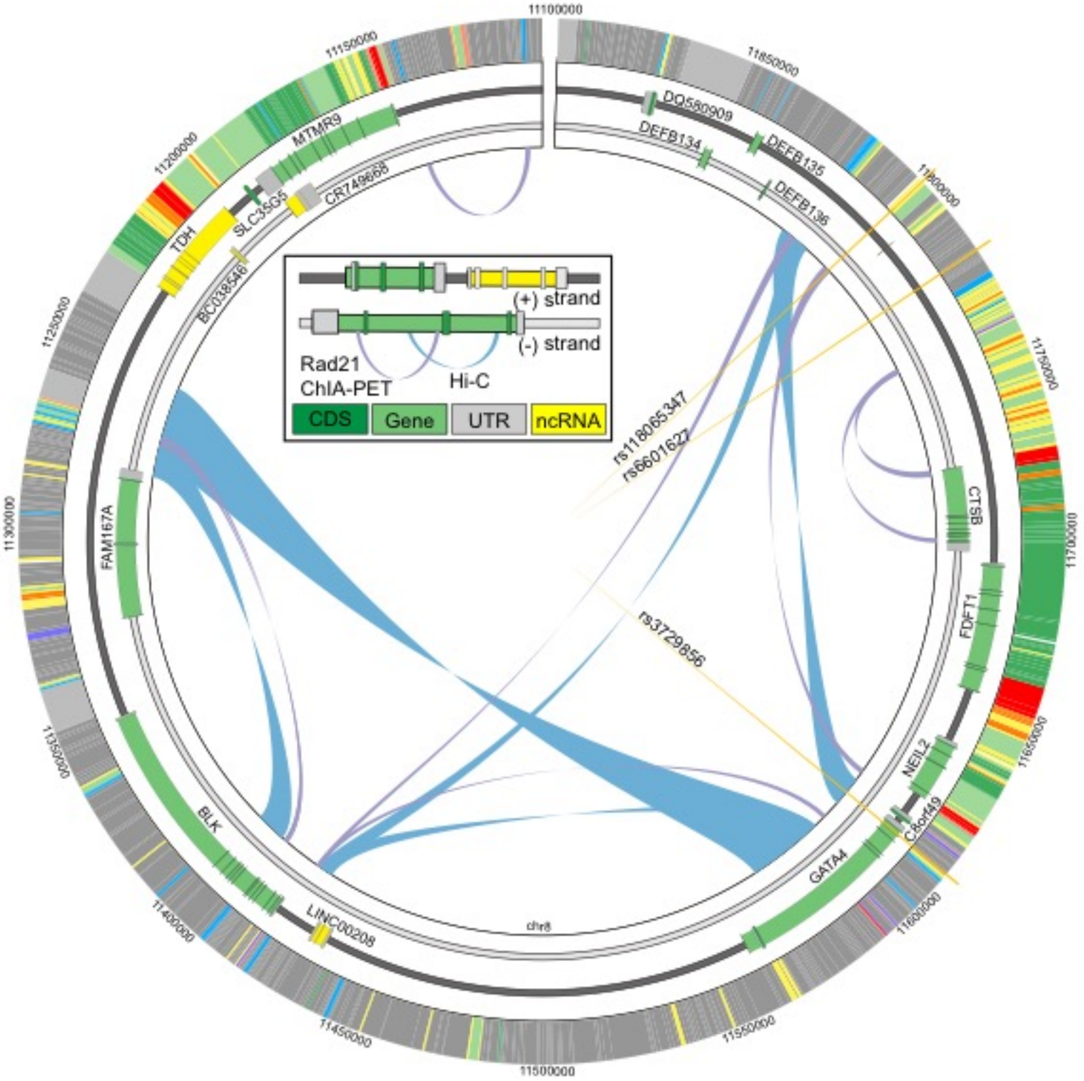


Figure 2-3 EndoMT, a key process in valve development, is impaired by GATA4 deficiency

(A) Western blot of GATA4 and GAPDH from control and *GATA4* sgRNA ECs. *GATA4* sgRNA ECs were differentiated from iPSCs transfected with px458 with *GATA4* sgRNA and enriched by GFP. Control ECs were derived from iPSCs with px458 and enriched by GFP. Uncropped version is presented in Figure S2-9.

Lower panel: quantification of western blot data. The data was normalized to control ECs. Experiments were repeated three times, averages and standard derivations were plotted.

(B) Western blot of SMA and GAPDH from control ECs, control ECs undergoing EndoMT, *GATA4* sgRNA ECs, and *GATA4* sgRNA ECs undergoing EndoMT. Uncropped version is presented in Figure S2-10.

Lower panel: quantification of western blot data. The data was normalized to control ECs undergoing EndoMT. Experiments were repeated three times, averages and standard derivations were plotted.

(C) Numbers of mesenchymal cells from control and *GATA4* sgRNA in collagen gel assay. The data was normalized to control. Experiments were repeated three times, averages and standard derivations were plotted.

(D) Immunofluorescence staining of SMA and CD31 of the control and *GATA4* sgRNA undergoing EndoMT. The scale bars represent 50 µm. Abbreviations: iPSCs: induced pluripotent stem cells. EC: endothelial cells. EndoMT: endothelial-to-mesenchymal transition. MW: molecular weight. kDa: kilodalton. * indicates P<0.05. ** indicates P<0.01.

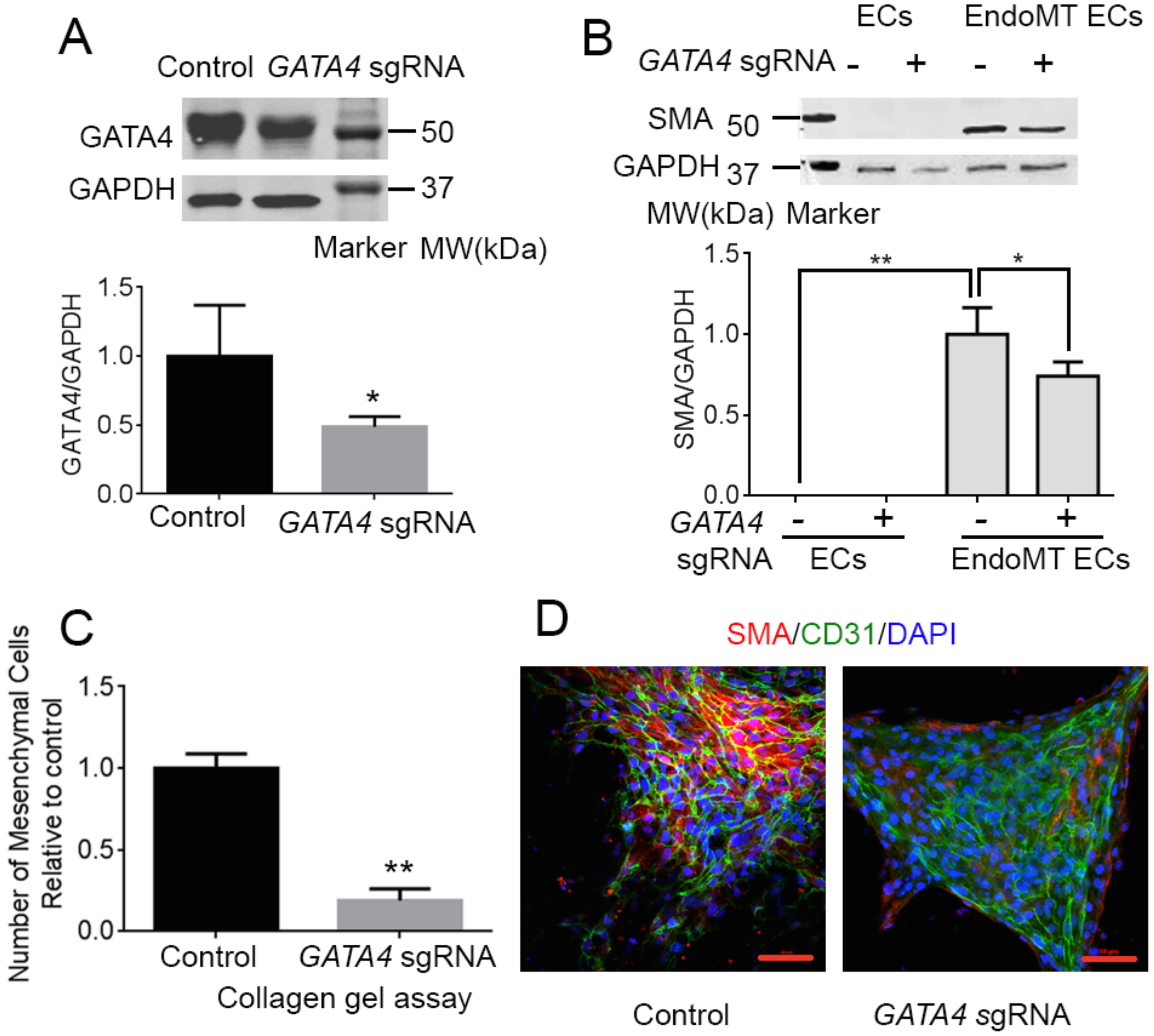

A
Control
GATA4 sgRNA
GATA4
GAPDH
50
37
Marker MW(kDa)
GATA4/GAPDH
1.5
1.0
0.5
0.0
*
Control
GATA4 sgRNA
B
ECs
EndoMT ECs
GATA4 sgRNA - + - +
SMA 50
GAPDH 37
MW(kDa) Marker
SMA/GAPDH
1.5
1.0
0.5
0.0
**
*
GATA4 sgRNA
- + - +
ECs
EndoMT ECs
C
Number of Mesenchymal Cells Relative to control
1.5
1.0
0.5
0.0
**
Control
GATA4 sgRNA
Collagen gel assay
D
SMA/CD31/DAPI
Control
GATA4 sgRNA

Table 2-1 Genetic variants associated with BAV

Variants			**Discovery**			**Replication**				**Combined**		
Chr:pos rsid	**Protein Change**	**Freq Case/ Ctrl (%)**	**N Case/ Ctrl**	**OR**	**P**	**N Case/ Ctrl**	**Freq Case/ Ctrl (%)**	**OR**	**P**	**N Case/ Ctrl**	**OR**	**P**
8:11778803 rs6601627	Intergenic	8.2/3.7	466/ 4660	2.38 (1.81-3.13)	1.5x10^{-10}	1,021/ 5,357	7.2/ 4.2	1.73 (1.42-2.12)	1.1x10^{-7}	1,487/ 10,017	1.93 (1.64-2.27)	3.0x10^{-15}
8:11614575 rs3729856	p.S377G GATA4	18.2/14.1	466/ 4660	1.39 (1.17-1.66)	3.2x10^{-4}	1,326/ 8,103	15.3/ 12.7	1.28 (1.14-1.45)	5.3x10^{-5}	1,792/ 12,763	1.31 (1.19-1.45)	8.8x10^{-8}
16:72146374 rs137867582	p.T1221M DHX38	0.9/0.1	466/ 4660	13.14 (5.39-32.04)	1.5x10^{-8}	720/ 5,831	0.37/ 0.15	2.87 (1-8.22)	5.0x10^{-2}	1,186/ 10,491	7.13 (3.63-14)	1.2x10^{-8}

Figure S2-1 Quantile-quantile plot for single-variant analysis results of BAV in the discovery cohort.

Variants in this plot include all those directly genotyped using the chip array and those successfully imputed from Haplotype Reference Consortium (HRC)(McCarthy, et al., 2016)

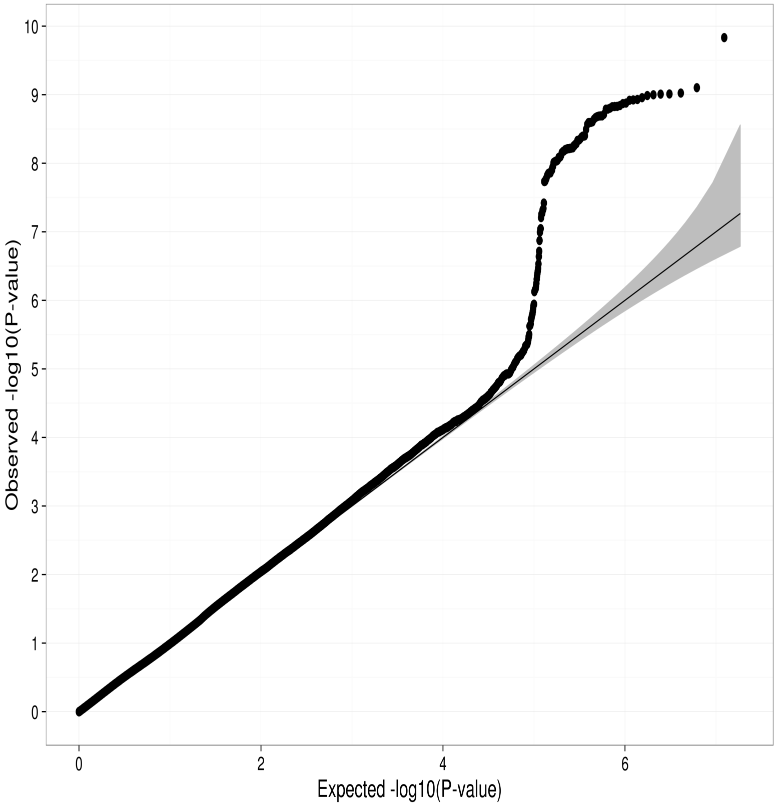


Figure S2-2 Manhattan plots for single variant association tests with BAV in the discovery cohort.

The red line indicates the genome-wide significance threshold (P = 5x10^{-8}).

(A) Before genotype imputation

(B) After genotype imputation from the HRC(McCarthy, et al., 2016)

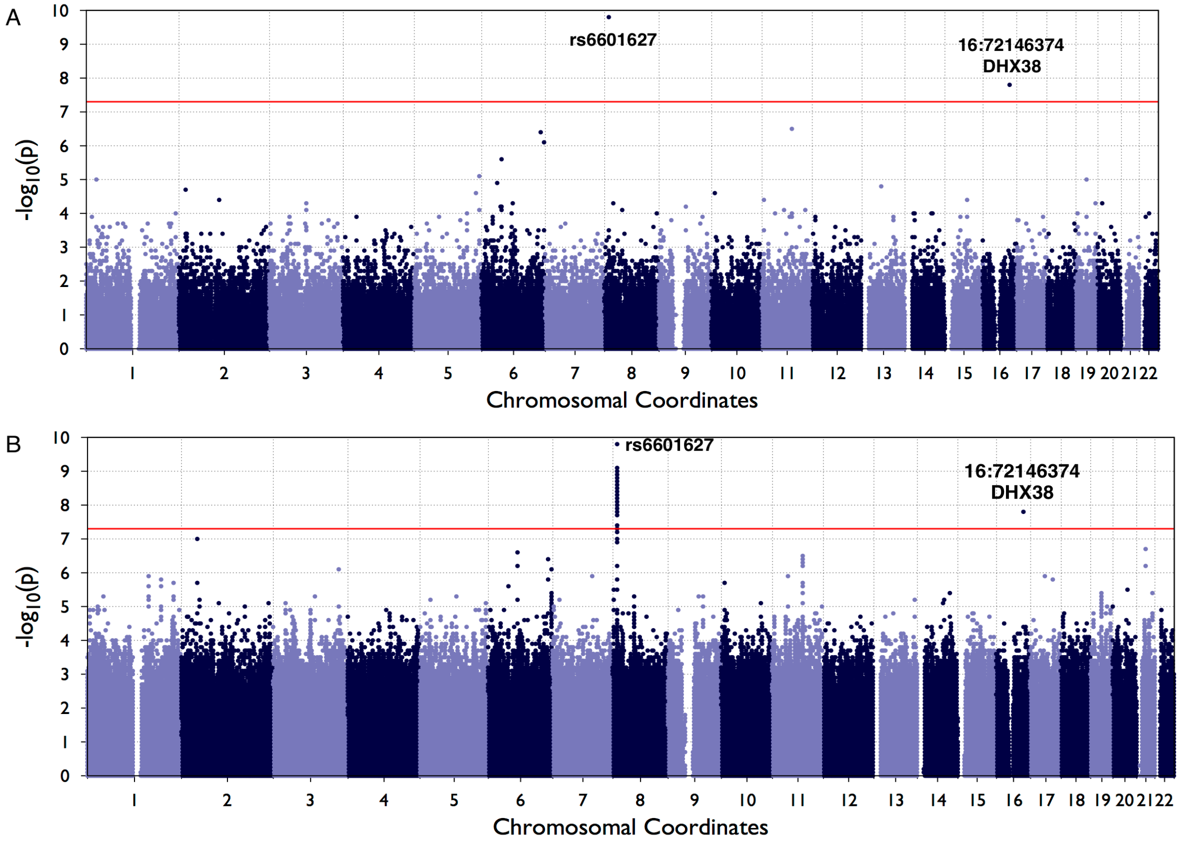


Figure S2-3 Regional association plot for all coding variants of the chr8 association region near *GATA4*, as observed in the discovery cohort (N=466 BAV cases, 4,660 controls).

The upper panel shows all 59 coding variants that were directly genotyped in the chip array in this region. A missense variant rs3729856 within *GATA4* was observed with p = 3.2x10^{-4}, that reached P = 8.8x10^{-8} following replication. The bottom panel includes additional 11 coding variants whose genotypes are imputed to the HRC reference(McCarthy, et al., 2016). Coding variants observed in this region contain missense variants (represented by triangles) and stop gain variants (represented by squares).

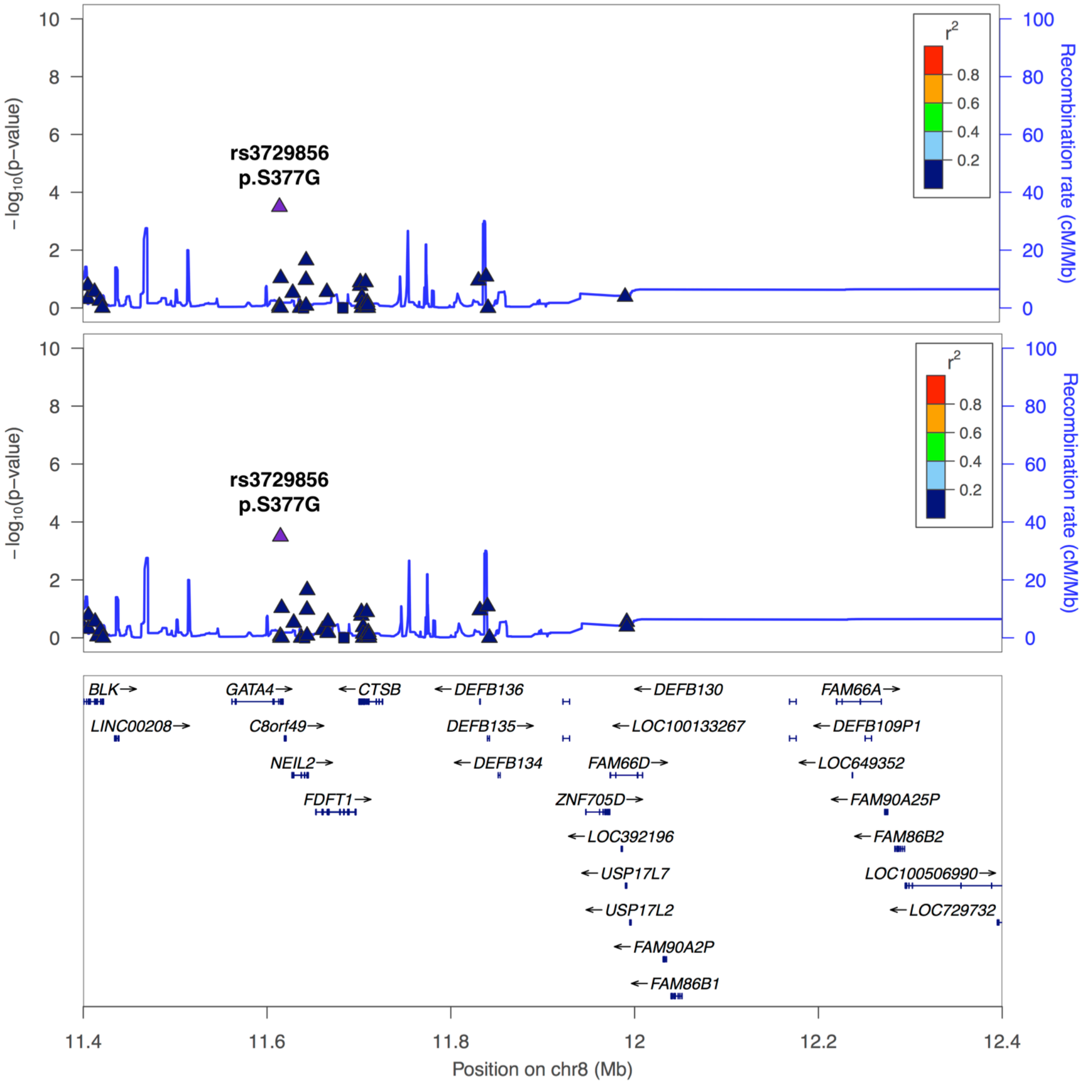
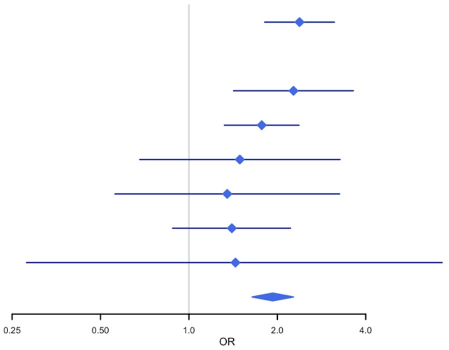


Figure S2-4 Forest plots of the BAV hits near *GATA4* by stage and study.

The combined results are for the meta-analysis of the discovery study and all replication studies.

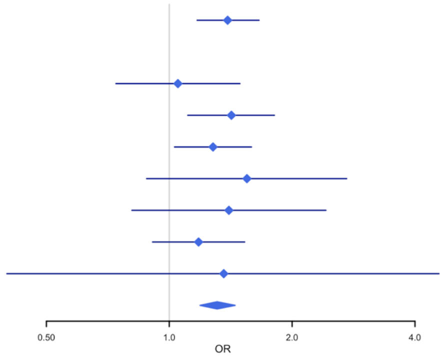
rs6601627

	OR	95% CI	Case/Control
Discovery Study			
CHIP	2.38	1.81-3.13	466/4,660
Replication Study			
CHIP replication	2.27	1.42-3.63	140/1,400
Partners HealthCare	1.77	1.32-2.37	452/1,634
UTHSC _BCM	1.49	0.68-3.27	62/337
UTHSC _UT	1.35	0.56-3.26	85/279
ASAP-Artist-Polca-Olivia	1.4	0.88-2.22	275/1,686
BioMe-EA-Omni	1.44	0.28-7.28	7/21
Combined	**1.93**	**1.64-2.27**	**1,487/10,017**



rs3729856

	OR	95% CI	Case/Control
Discovery Study			
CHIP	1.39	1.17-1.66	466/4,660
Replication Study			
CHIP replication	1.05	0.74-1.49	140/1,400
MHI	1.42	1.11-1.81	305/2,746
Partners HealthCare	1.28	1.03-1.59	452/1,634
UTHSC _BCM	1.55	0.88-2.72	62/337
UTHSC _UT	1.4	0.81-2.42	85/279
ASAP-Artist-Polca-Olivia	1.18	0.91-1.53	275/1,686
BioMe-EA-Omni	1.36	0.40-4.58	7/21
Combined	**1.31**	**1.19-1.45**	**1,792/12,763**



rs137867582

	OR	95% CI	Case/Control
Discovery Study			
CHIP	13.14	5.39-32.04	466/4,660
Replication Study			
CHIP replication	11.96	1.64-87.25	140/1,399
MHI	2.02	0.54-7.53	305/2,746
ASAP-Artist-Polca-Olivia	0.50	0.02-14.2	275/1,686
Combined	**7.13**	**3.63-14**	**1,186/10,491**

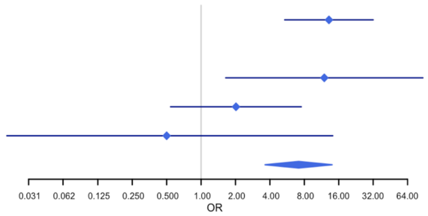
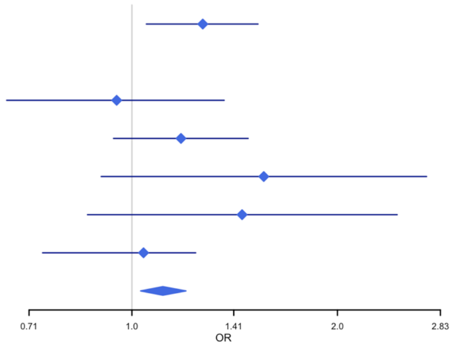


Figure S2-5 Forest plots of the reciprocal conditional analysis of the two BAV hits near *GATA4* by stage and study.

The combined results are for the meta-analysis of the discovery study and all replication studies.

rs3729856

	OR	95% CI	Case/Control
Discovery Study			
CHIP	1.27	1.05-1.53	466/4,660
Replication Study			
CHIP replication	0.95	0.66-1.37	140/1,400
Partners HealthCare	1.18	0.94-1.48	452/1,634
UTHSC _BCM	1.56	0.90-2.70	62/337
UTHSC _UT	1.45	0.86-2.45	85/279
ASAP-Artist-Polca-Olivia	1.04	0.74-1.24	275/1,686
Combined	**1.11**	**1.03-1.20**	**1,480/9,996**



rs6601627

	OR	95% CI	Case/Control
Discovery Study			
CHIP	2.217	1.69-2.91	466/4,660
Replication Study			
CHIP replication	2.3	1.42-3.74	140/1,400
Partners HealthCare	1.68	1.24-2.27	452/1,634
UTHSC _BCM	1.65	0.73-3.73	62/337
UTHSC _UT	1.4	0.60-3.26	85/279
ASAP-Artist-Polca-Olivia	0.91	0.56-1.4	275/1,686
Combined	**1.4**	**1.26-1.56**	**1,480/9,996**

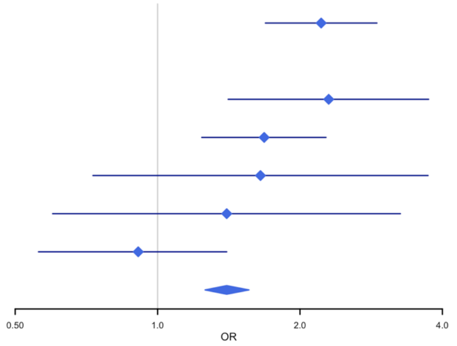


Figure S2-6 The mRNA expression levels of genes surrounding the non-coding associated variant rs6601627 from the GTEx portal(2013).

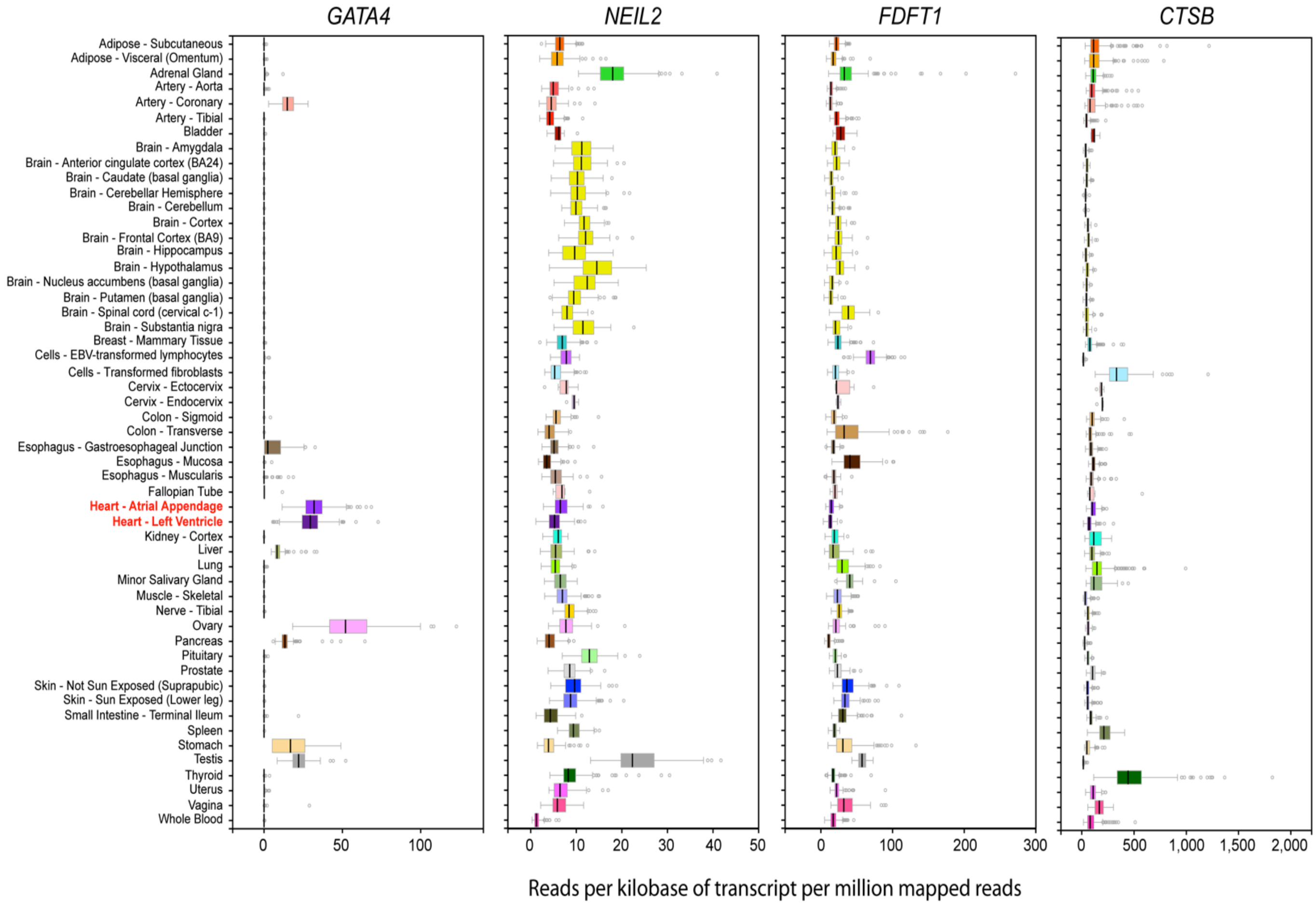


Figure S2-7 iPSCs from Control patient are pluripotent.

(A) Immunofluorescence staining of OCT4, SOX2, NANOG of the iPSC colonies. The scale bars represent 50μm.

(B) Immunofluorescence staining of SSEA4, TRA-1-60 and TRA-1-81 of the iPSC colonies. The scale bars represent 50μm.

(C) H&E staining of teratomas. The scale bars represent 50μm. DAPI marks the nucleus. Abbreviations: iPSCs: induced pluripotent stem cells.

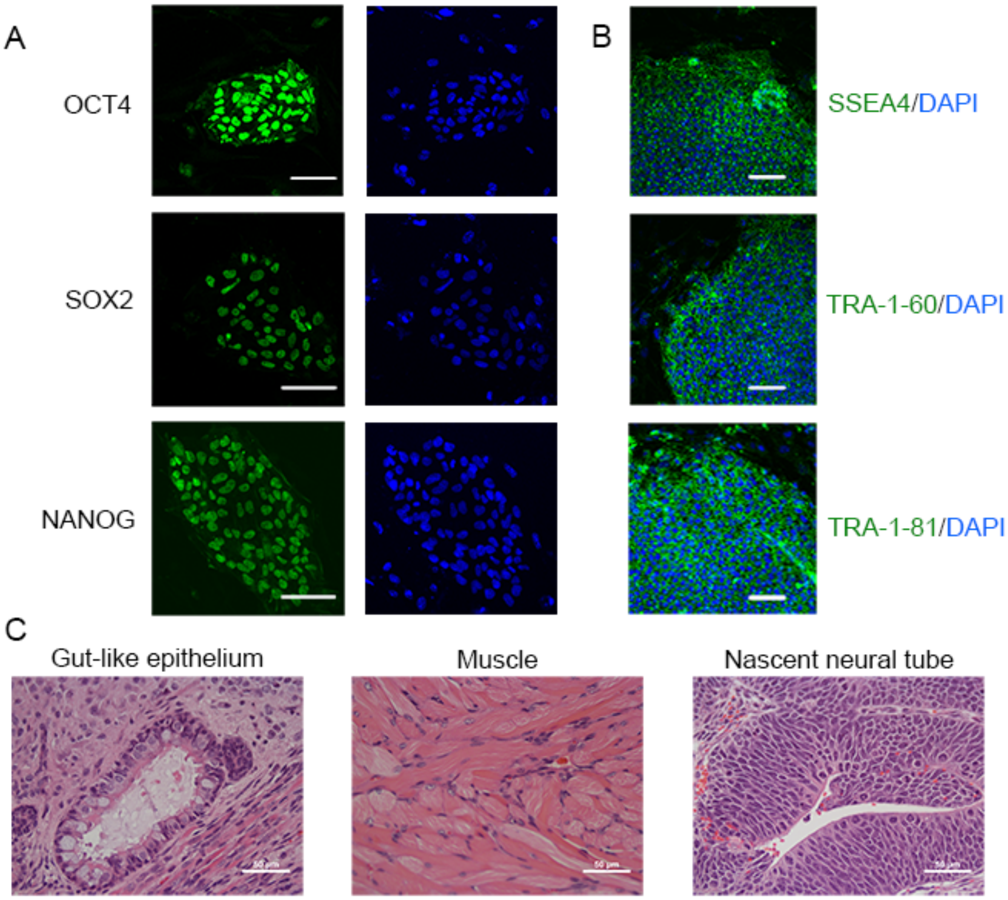


Figure S2-8 *GATA4* sgRNA/cas9 electrotransfection of iPSCs.

(A) Diagram of experimental process.

(B) Illustration of *GATA4* sgRNA target site.

(C) Flow cytometry of electrotransfected iPSCs. IPSCs in "Control" group were transfected with PX458 plasmids containing Cas9 and GFP. IPSCs in *GATA4* sgRNA group were transfected with PX458 plasmids containing Cas9, GFP and *GATA4* sgRNA. Successfully transfected cells were GFP positive. GFP positive cells within the inside area were selected for further experiments.

(D) Immunofluorescence staining of CD31 on ECs differentiated from iPSCs. DAPI marks the nucleus. Scale bars represent 100μm. Abbreviations: iPSCs: induced pluripotent stem cells. EC: endothelial cells. GFP: green fluorescent protein. EndoMT: endothelial-to-mesenchymal transition.

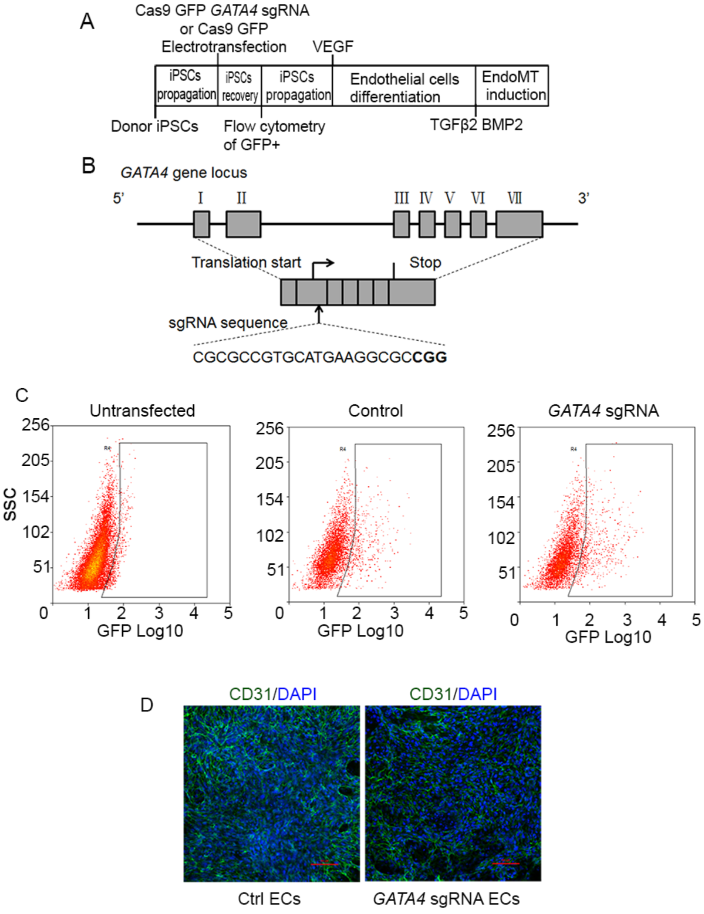


Figure S2-9 Uncropped version of GATA4 and GAPDH western blot.

This is uncropped version of Figure 2-3A. Abbreviations: Ctrl: control. MW: molecular weight. kDa: kilodalton.

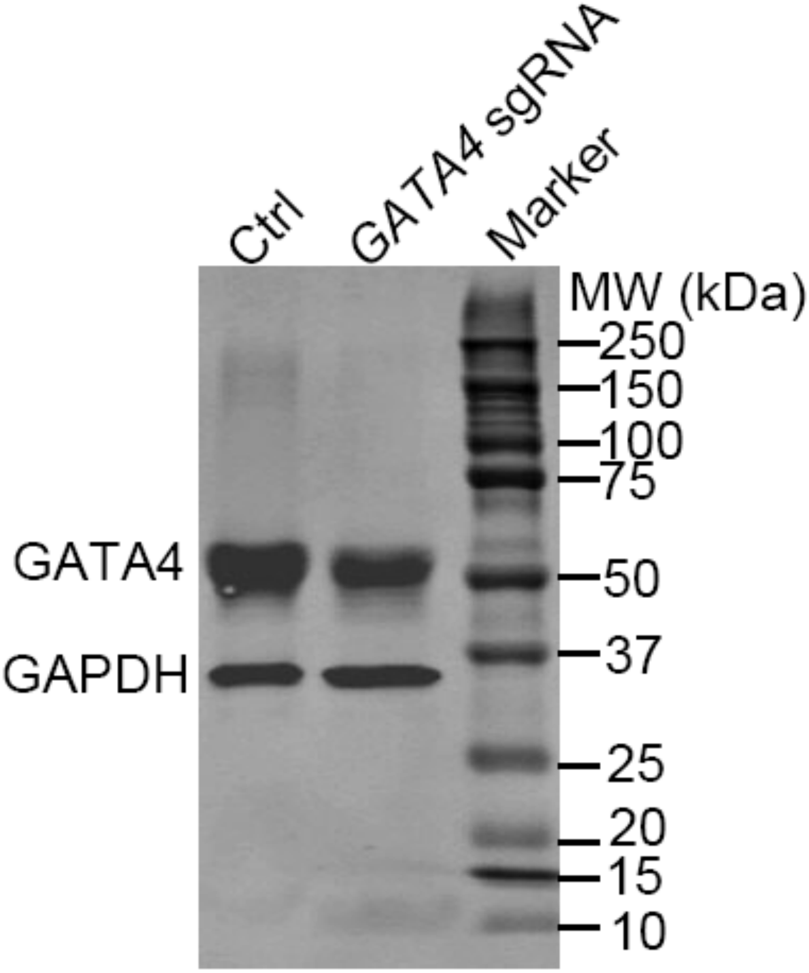


Figure S2-10 Uncropped version of SMA and GAPDH western blot.

This is uncropped version of Figure 2-3B. Abbreviations: EC: endothelial cells. EndoMT: endothelial-to-mesenchymal transition. MW: molecular weight. kDa: kilodalton.

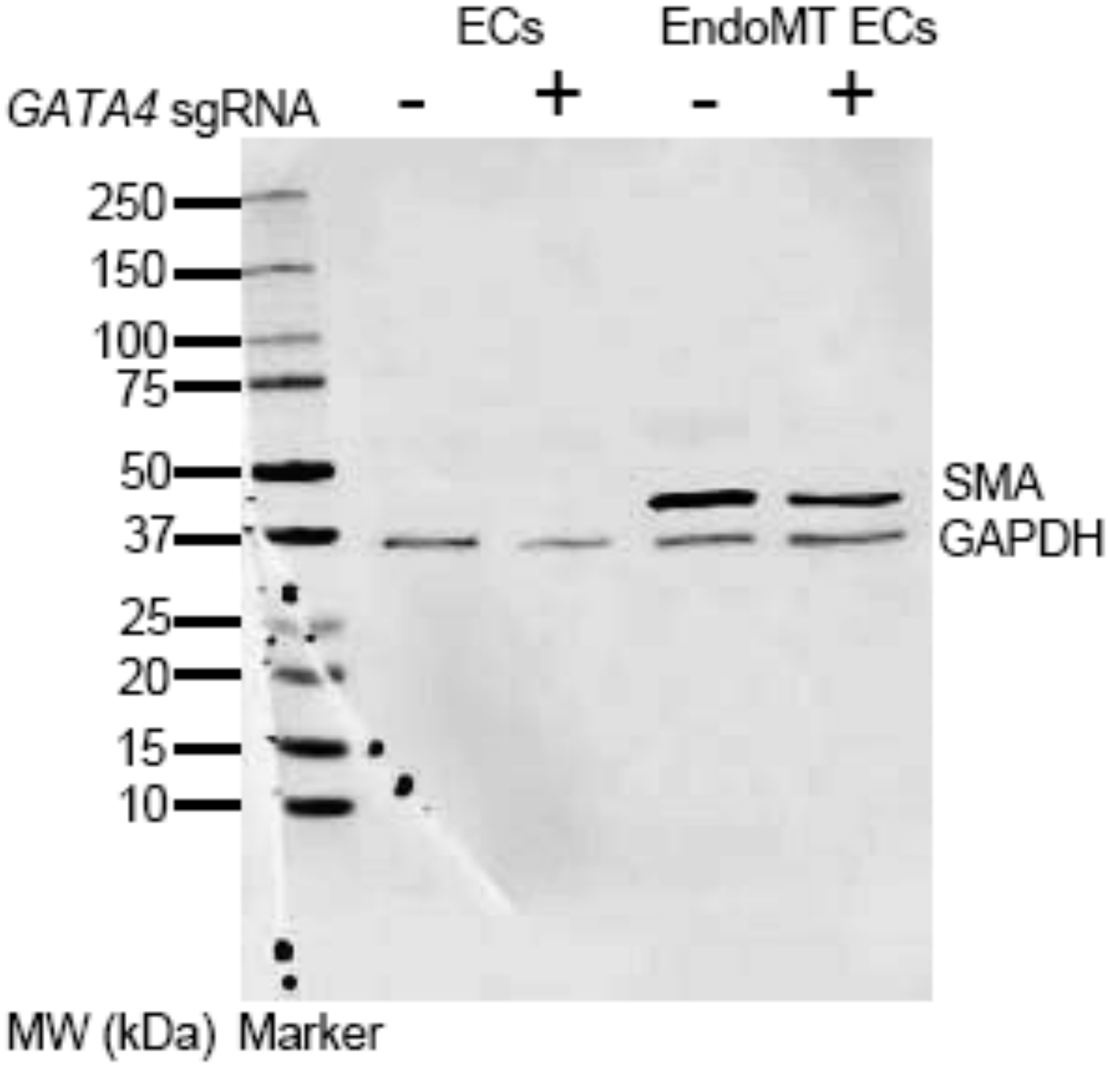


Table S2-1 Clinical characteristics of the BAV cases in the discovery cohort (n = 466)

	BAV Cases in Discovery Cohort
Age at inclusion, median (IQR)	39.0 (31.0-46.0)
Male sex, n (%)	345 (74)
Hypertension, n (%)	253 (54)
Dyslipidemia, n (%)	203 (44)
Smoking - ever, n (%)	198 (43)
BAV subtype, n (%)	
Type 0 anterior-posterior	6 (1.3)
Type 0 lateral	10 (2.1)
Type 1a	202 (43)
Type 1b	45 (9.7)
Type 1c	11 (2.4)
Type 2a	16 (3.4)
Type 2b	1 (0.2)
Type 2c	8 (1.7)
Type 3	5 (1.1)
No information on subtype	162 (35)
Thoracic aortic aneurysm, n (%)	
Arch	40 (8.6)
Ascending	316 (68)
Descending	10 (2.1)
Root	21 (4.5)
None	79 (17)
Aortic stenosis, n (%)	259 (56)
Aortic insufficiency, n (%)	246 (53)
Other congenital heart defects, n (%)	4 (0.9)
BAV in family*, n (%)	93 (20)

*Number (%) of cases reporting one or more family member with BAV.

Table S2-2 Non-additive association results for the BAV hits in the discovery study (466 BAV cases and 4,660 controls)

Variants			**Dominant Tests**		**Recessive Tests**	
Chr:pos	**rsID**	**Protein Change**	**OR**	**P**	**OR**	**P**
8:11778803	rs6601627 A/G	Intergenic	2.39 (1.81-3.15)	$6.4x10^{-10}$	8.04 (1.78-36.26)	$6.7x10^{-3}$
8:11614575	exm682536 rs3729856 A/G	p.S377G GATA4	1.48 (1.21-1.82)	$1.5x10^{-4}$	1.24 (0.63-2.41)	0.5

Table S2-3 Association results with thoracic aortic aneurysm (TAA) of the two BAV hits in the discovery study.

Variants			BAV with TAA				BAV without TAA				TAA without BAV				Heterogeneity P*
Chr:pos	rsID	Protein Change	Freq (%) Case/ Ctrl	N Case/ Ctrl	OR	P	Freq (%) Cases/ Ctrl	N Case/ Ctrl	OR	P	Freq (%) Cases/ Ctrl	N Case/ Ctrl	OR	P	
8:11778803	rs6601627 A/G	Intergenic	8.1/ 3.8	387/ 3870	2.29 (1.72-3.06)	$2.0x10^{-8}$	8.2/ 3.4	79/ 790	2.76 (1.41-5.40)	$3.0x10^{-3}$	4.8/ 3.6	414/ 4140	1.38 (0.98-1.95)	$6.0x10^{-2}$	0.62
8:11614575	exm682536 rs3729856 A/G	p.S377G GATA4	18.0/ 14.3	387/ 3870	1.35 (1.11-1.65)	$3.0x10^{-3}$	19.6/ 13.2	79/ 790	1.63 (1.05-2.54)	$3.0x10^{-2}$	13.7/ 13.3	414/ 4140	1.04 (0.84-1.28)	$7.2x10^{-1}$	0.44

*Heterogeneity tests were performed to compare the tests of BAV patients with TAA and the tests of BAV patients without TAA.

Table S2-4 Association results with BAV cases with and without family members with BAV and/or TAA of the two BAV hits in the discovery study.

Variants			BAV cases without family members who have BAV and/or TAA				BAV cases with family members who have BAV and TAA				Heterogeneity P
Chr:pos	rsID	Protein Change	Freq (%) Cases/Ctrl%)	N Case/Ctrl	OR	P	Freq (%) Cases/Ctrl	N Case/Ctrl	OR	P	
8:11778803	rs6601627 A/G	Intergenic	8.5/3.6	371 / 3710	2.54 (1.90-3.41)	4.8x10^{-10}	6.8/4.0	95/ 950	1.84 (0.98-3.44)	5.8x10^{-2}	0.36
8:11614575	exm682536 rs3729856 A/G	p.S377G GATA4	18.2/14.3	371 / 3710	1.37 (1.11-1.67)	2.7x10^{-3}	18.4/13.6	95/ 950	1.50 (1.00-2.26)	5.3x10^{-2}	0.69

Table S2-5 Association results with BAV subtypes of the two BAV hits in the discovery study.

Variants			BAV type 1a				BAV non-type 1a (patients without available subtype information are excluded)				Heterogeneity P
Chr:pos	**rsID**	**Protein Change**	**Freq (%) Cases/Ctrl**	**N Case/Ctrl**	**OR**	**P**	**Freq (%) Cases/Ctrl**	**N Case/Ctrl**	**OR**	**P**	
8:11778803	rs6601627 A/G	Intergenic	6.9/3.5	202/ 2020	2.07 (1.35-3.17)	$8.3x10^{-4}$	10.8/3.6	102/ 1020	3.42 (2.00-5.87)	$8.7x10^{-6}$	0.15
8:11614575	exm682536 rs3729856 A/G	p.S377G GATA4	17.3/13.4	202/ 2020	1.40 (1.06-1.84)	$1.9x10^{-2}$	15.2/14.6	102/ 1020	1.09 (0.72-1.66)	$6.8x10^{-1}$	0.34

Table S2-6 Association results with BAV cases in males and females of the two BAV hits in the discovery study.

Variants			In males				In females				Heterogeneity P
Chr:pos	rsID	Protein Change	Freq (%) Cases/Ctrl	N Case/Ctrl	OR	P	Freq (%) Cases/Ctrl	N Case/Ctrl	OR	P	
8:11778803	rs6601627 A/G	Intergenic	9.2/3.8	344/3440	2.74 (2.03-3.69)	4.0x10^{-11}	5.3/3.5	122/1220	1.53 (0.84-2.80)	1.6x10^{-1}	0.09
8:11614575	exm682536 rs3729856 A/G	p.S377G GATA4	18.0/14.1	344/3440	1.36 (1.10-1.68)	4.2x10^{-3}	18.9/14.1	122/1220	1.48 (1.03-2.11)	3.2x10^{-2}	0.70

Table S2-7 ICD-9 Diagnoses codes used to exclude MGI controls with aortic diseases

ICD-9	**Inclusion Diagnoses for Aortic Disease MGI Controls**
441	Aortic Disease - Major Classes
Aortic dissection	
441.00	Unspecified site
441.01	Thoracic
441.02	Abdominal
441.03	Thoracoabdominal
Aortic Aneurysm	
441.2	Ascending
441.1	Ascending, if ruptured
441.2	Arch
441.1	Arch, if ruptured
441.9	Descending, not otherwise specified (NOS)
441.5	Descending, if ruptured
441.2	Thoracic descending
441.1	Thoracic descending, if ruptured
441.4	Abdominal descending
441.3	Abdominal descending, if ruptured
441.7	Thoracoabdominal
441.6	Thoracoabdominal, if ruptured
441.4	Abdominal
441.3	Abdominal, if ruptured

Chapter 3 Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels

3.1 Abstract

The accuracy of genotype imputation depends upon two factors: the sample size of the reference panel and the genetic similarity between the reference panel and the target samples. When multiple reference panels are not consented to combine together, it is unclear how to combine the imputation results to optimize the power of genetic association studies. We compared the accuracy of 9,265 Norwegian genomes imputed from three reference panels – 1000 Genomes Phase 3 (1000G), Haplotype Reference Consortium (HRC), and a reference panel containing 2,201 Norwegian participants from the population-based Nord Trøndelag Health Study (HUNT) from low-pass genome sequencing. We observed that the population-matched reference panel allowed for imputation of more population-specific variants with lower frequency (minor allele frequency (MAF) between 0.05% and 0.5%). The overall imputation accuracy from the population-specific panel was substantially higher than 1000G and was comparable with HRC, despite HRC being 15-fold larger. These results recapitulate the value of the population-specific reference panels for genotype imputation. We also evaluated different strategies to utilize multiple sets of imputed genotypes to increase the power of association studies. We observed that testing association for all variants imputed from any panel results in higher power to detect association than the alternative strategy of including only one version of each genetic variant, selected for having the highest

imputation quality metric. This was particularly true for lower-frequency variants (MAF < 1%), even after adjusting for the additional multiple testing burden.

3.2 Introduction

Many novel disease-associated signals for a wide variety of diseases and traits have been successfully identified using imputation-based meta-analyses(Cheng and Thompson, 2016; Cooper, et al., 2008; De Jager, et al., 2009; Ge, et al., 2016; Horikoshi, et al., 2015; Houlston, et al., 2008; Jin, et al., 2016; Loos, et al., 2008; Ruth, et al., 2015; Zeggini, et al., 2008; Zeggini, et al., 2007). Genotype imputation is the process of inferring missing genotypes in study samples using a reference panel of high-density haplotypes(Li, et al., 2009). Imputation allows variants that are not directly genotyped to be studied without other costs than computation. Previous simulations showed that imputation substantially increases the power of association studies to detect causal loci(Marchini and Howie, 2010; Spencer, et al., 2009). Imputation-based genome-wide association studies (GWAS) have successfully identified novel signals that were undetected in chip-based studies. For example, two disease-associated signals were detected in the 1000G-based imputation(Auton, et al., 2015) for the Wellcome Trust Case Control Consortium phase 1 Data (WTCCC), which were missed in the original WTCCC GWAS study that was performed four years before(Burton, et al., 2007; Huang, et al., 2012). Imputation also facilitates fine-mapping studies by allowing most polymorphic variants, including causative ones, to be tested in known disease associated loci. For example, the strongest association signal, observed at the imputed variant rs7903146 of the *TCF7L2* locus in the WTCCC type 2 diabetes scan, is suggested to be causal association in the locus(Mahajan, et al., 2014; Marchini, et al., 2007). Furthermore, imputation allows for meta-analysis between samples that have genotyped using different arrays, increasing power.

However, for studies that have access to population-matched genome sequenced individuals, there is uncertainty in deciding between a smaller, ancestry-matched reference panel and a larger publicly-available cosmopolitan reference panel. An ideal reference panel is expected to have closely matched ancestry to study samples because the genetic similarity increases the accuracy of imputation(Deelen, et al., 2014; Huang and Tseng, 2014; Huang, et al., 2015; Low-Kam, et al., 2016; Mitt, et al., 2017; Okada, et al., 2015; Pistis, et al., 2015; Roshyara and Scholz, 2015; Walter, et al., 2015). On the other hand, the imputation accuracy increases when larger reference panels are used, especially for lower-frequency variants(Browning and Browning, 2009; Howie, et al., 2009; Huang, et al., 2009; Li, et al., 2009; Roshyara and Scholz, 2015).

Furthermore, different whole-genome reference panels may generate discordant imputed genotypes for the same variants in the same study samples. This brings in challenges for the follow-up association tests. The optimal strategy to perform association tests using genotypes imputed by different reference panels remains unclear. IMPUTE2 provides one possible approach to merge all reference panels to a single larger panel for genotype imputation when multiple reference panels are available (Howie, et al., 2009), which may avoid the problem that different versions of genotypes are imputed for the same variants. The Genome of the Netherlands Consortium and the UK10K study have further shown that the combined reference panel of 1000G and the population-specific reference resulted in better imputation results compared to the two individual panels for rare variants(Deelen, et al., 2014; Huang, et al., 2015). However, this approach is not feasible when individual-level haplotypes within the reference panel are not accessible, as is the case with the Haplotype Reference Consortium (HRC)(McCarthy, et al., 2016), primarily due to ethical issues surrounding sharing of individual-level genetic data(McCarthy, et al., 2016).

Here we genotyped 9,265 Norwegian participants from the HUNT study(Krokstad, et al., 2013) for 350,270 polymorphic autosomal variants using the Illumina Human CoreExome array with approximately 240,000 GWAS tagging markers. We created a population-matched reference panel by whole-genome sequencing (WGS) 2,021 individuals from the HUNT study to a mean depth of 5x. We imputed variants from the HUNT WGS reference panel as our ethnically matched panel. We also performed imputation with two additional imputation reference panels: the HRC(McCarthy, et al., 2016) and 1000G Phase 3(Auton, et al., 2015). First, we systematically evaluated and compared the imputation results from the three reference panels, including the number of successfully imputed variants as well as the imputation accuracy. Next, we evaluated and compared the power of association tests between two approaches to incorporate multiple versions of imputed genotypes. First is the "best Rsq" approach, which retains imputed genotypes only from the panel with highest imputation quality metrics for each variant. Second is the "best p-value" approach that tests association with all imputed genotypes and uses the most significant association p value, adjusting for the additional variants tested.

3.1 Materials and Methods

3.1.1 Array-based genotyping

9,265 samples from the HUNT Biobank in Norway were genotyped at 350,270 polymorphism autosomal variants using an Exome + GWAS chip array (HumanCoreExome-12 v1.0, Illumina). Genotype calling was performed using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina). Samples with <98% genotype calls (N = 37), evidence of gender discrepancy (N = 21), duplicates (N = 66) as well as individuals with non-Norwegian ancestry identified by plotting the first 10 genotype-driven principal components(Springer-Verlag, 1986) (N = 7) were excluded from further analysis (N = 131, 1.19%). As Figure S3-1 shows, the HUNT GWAS samples have similar ancestry to the samples in the HUNT WGS reference panel. All HUNT research subjects provided informed written consent and IRB approval was obtained for genetic studies.

Relatedness was evaluated based on the estimation of the proportion of identity by descent (IBD) by PLINK(Purcell, et al., 2007). We excluded 1,644 samples from the HUNT GWAS sample due to 1st or 2nd degree relatedness to samples in HUNT WGS, defined as IBD ≥ 0.25. We excluded samples that were related to samples within the reference panel to avoid inflating imputation statistics for regions inherited IBD. We performed variant-level quality control by excluding 19,872 variants that met any of the following criteria; variants with a cluster separation score < 0.3 reported by GenomeStudio V2011.1 (Illumina), < 95% genotype call rate, or deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-5}$).

3.1.2 Genotype imputation

Genotype imputation with the 1000G Phase 3(Auton, et al., 2015) and the HRC(McCarthy, et al., 2016) reference panels was conducted using the Michigan Imputation Server(Das, et al., 2016) and imputation with the HUNT WGS reference panel was conducted using a local server. The study samples were phased using SHAPEIT2(v2.r790)(Delaneau, et al., 2013) followed by imputation using minimac3(v2.0.1)(Fuchsberger, et al., 2015; Howie, et al., 2012). Two imputation metrics output by minimac3 were used for evaluating the imputation quality: ImpRsq and EmpRsq. ImpRsq is previously known as $\hat{r}^2$ in different versions of the MaCH/minimac(Fuchsberger, et al., 2015; Howie, et al., 2012; Li, et al., 2010). ImpRsq is defined for both genotyped and ungenotyped variants in the chip array as an estimate of the squared correlation between imputed dosages and true, unobserved genotypes, calculated as the observed variance over the expected variance. EmpRsq is defined only for genotyped variants in the chip array as the squared correlation between leave-one-out imputed dosages and the true, observed genotypes (See "Estimated Imputation Accuracy" section at http://genome.sph.umich.edu/wiki/Minimac_Diagnostics for details).

3.1.3 Reference panels

The HUNT WGS reference panel contains 1,101 earliest onset cases with myocardial infarction and 1,100 age and sex matched controls that were selected from the HUNT study(Krokstad, et al., 2013). Whole genome sequencing to ~5x depth was performed on either Illumina HiSeq 2000 or 2500. We followed the GotCloud SNP calling pipeline to process the whole genome sequencing data(Jun, et al., 2015). The variant sites and genotype likelihood were called using SAMtools(Li, et al., 2009) and the genotypes for SNPs were refined and phased using Beagle v4(Browning and Browning, 2013). After quality control, 20.2 million single nucleotide variants were retained in 2,201 samples, of which 4 million were unique to our study; not observed in dbSNP 144(Sherry, et al., 2001), 1000 Genomes Phase 3(Auton, et al., 2015), UK10K(Walter, et al., 2015), ESP6500(2013), or ExAC.r0.3(Lek, et al., 2016) (Table 3-1). The individuals in the HUNT WGS panel have similar ancestry to the HUNT study samples (Figure S3-1) and are from the same geographic region, although we excluded in the genotyped samples any 1st or 2nd degree relatives of the sequenced samples to avoid biased estimates of the accuracy of imputation. Additionally, there were no close relatives within the sequenced samples. The other two reference panels that we used for genotype imputation are the 1000 Genomes Phase 3 (1000G)(Auton, et al., 2015) and the HRC release 1(McCarthy, et al., 2016) containing 32,488 individuals, both of which are pre-stored in the Michigan Imputation Server(Das, et al., 2016) (Table 3-2). The HUNT cohort contributed an early freeze of whole genome sequencing data consisting of 1,023 samples to the HRC consortium. Thus, the HUNT WGS and the HRC reference panels have 1,023 samples in common. Variants with minor allele counts (MAC) less than or equal to 5 were excluded from HRC(McCarthy, et al., 2016).

3.1.4 Permutation test

To determine the genome-wide significance thresholds for association tests using the two approaches to incorporate imputed genotypes, we performed permutation tests. The measurements of the high-density lipoprotein (HDL) cholesterol for the study samples were permuted 1,000 times. Each permutation was followed by a genome-wide association test (GWAS) using the permuted phenotypes. The most significant p-values from each of the 1,000 GWAS were ranked. And the significance threshold with family-wise error rate (FWER) n/1000 equals to the nth smallest p-value. Because the "best p-value" approach tests more variants, it will be a more stringent significance threshold than the "best Rsq" approach.

3.1.5 Power estimation

In order to estimate the power to detect association under the two approaches to incorporate imputed genotypes from multiple reference panels, we considered directly genotyped variants as causal variants, and used multiple sets of imputed genotypes to evaluate the power. First, we obtained the leave-one-variant-out imputed dosages for those directly genotyped variants. The official release of minimac3 performs leave-one-out hidden Markov model (*HMM*) calculation internally to calculate leave-one-out Rsq summary statistics, but does not output individual dosages (Fuchsberger, et al., 2015; Howie, et al., 2012). We modified minimac3 to include the individual leave-out-out dosages in the output VCF for the genotyped variants. Second, we simulated phenotypes based on the genotypes obtained by the chip array. Finally, we evaluated power of the two approaches by performing association tests between the simulated phenotypes and the imputed dosages based on either “best Rsq” or “best p-value” approaches.

The details of simulation follow the steps described below:

1. Select the non-centrality parameter corresponding to the association test p-value p_t. We calculate the non-centrality parameter Nr^2 as a chi-square statistic corresponding to the upper-tail probability p_t, where *N* is the total number of study subjects. This ensures that the median p-value is p_t when the true phenotypic variance explained by the genotype is r^2.

2. For each variant, we randomly draw ε from the normal distribution with mean 0 and standard deviation $\sqrt{1-r^2}$. We calculate the effect size β as $\sqrt{r^2/2f(1-f)}$, where f is the minor allele frequency (MAF) estimated using the chip genotypes of the variant. The phenotype value y is then calculated as $G\beta + \varepsilon$, where the chip genotypes G is 0, 1, or 2. The phenotypic variance explained by G and ε will be r^2 and *1*-r^2, respectively.

3. We perform the linear regression using the leave-one-variant-out dosages for this variant, which were imputed using the three different reference panels respectively, and the phenotype y.

4. For the “best p-value” approach, the final association p value equals to most significant one among the three p values associated with the three different versions of imputed dosages. With the “best Rsq” approach, the final p value equals to the one corresponding to the reference panel with the highest imputation quality (ImpRsq), an estimated value for the correlation between imputed genotypes and true, unobserved genotypes.

5. The power to detect association signals equals to the percentage of final p values exceeding the genome-wide significance threshold determined for each approach by the permutation tests described above.

We performed linkage disequilibrium(LD) based variant pruning for the 289,376 directly genotyped variants that were found by all three reference panels using PLINK(Purcell, et al., 2007) and obtained 132,183 variants with LD $r^2 < 0.2$ among each other. Then we randomly selected 3,000 variants for each of the MAF categories: MAF ≤ 0.001, MAF > 0.001 and ≤ 0.01, MAF > 0.1 and ≤ 0.05, and MAF > 0.05. We applied ImpRsq > 0.3, 0.5 and 0.8 to remove poorly imputed genotypes and variants that were successfully imputed from at least two references were used for this simulation study. All 5 steps above were repeated given different p_t's ranging from $5x10^{-8}$ to $1x10^{-13}$. Additionally, the entire process was repeated 5 times across the selected variants to average power.

3.1.6 Partial correlation estimation

To quantify the net gain of imputation accuracy obtained by including another reference panel on top of an existing panel, we estimated the partial correlation between the leave-one-out imputed dosages from the additional panel and the chip genotypes, conditioned on the leave-one-out imputed dosages from the existing panel. The correlation has been estimated for every pair of reference panels among the three on each of the 289,376 genotyped variants that were found in all three panels. For example, to estimate the net gain of including 1000G panel on top of HUNT panel (PartialRsq [1000G,Chip | HUNT]), we first obtained the leave-one-out dosages based on 1000G and HUNT WGS (details described in the Power estimation subsection). Secondly, for each variant, we performed three linear regressions on the chip genotypes: the first one has the imputed dosages from 1000G and HUNT WGS as covariates (model 1), the second one has the imputed dosages from HUNT WGS only as a covariate (model 2), and the third one does not have any other covariate except for the intercept (model 3). Lastly, we obtained sum of squared residuals (SSR) for the three linear regressions and calculated the partial correlation (partial Rsq) as $\frac{SSR_{model2}-SSR_{model1}}{SSR_{model3}}$. In a similar notation, the EmpRsq is equivalent to $\frac{SSR_{model3}-SSR_{model2}}{SSR_{model3}}$, and their sum should be equivalent to the proportion of explained variance by both sets of imputed dosages. Our intuition is that the more extra information the additional reference panel provides, the higher the partial correlation will be.

3.4 Results

3.4.1 Evaluating successfully imputed variants using different reference panels

In total, ~23.8 million variants were successfully imputed using minimac 3(Fuchsberger, et al., 2015; Howie, et al., 2012) from at least one of the three reference panels and exceeded the threshold of estimated imputation quality (ImpRsq) ≥ 0.3 (Figure 3-1). The three reference panels yielded roughly equal number of SNPs with MAF more than 1%, but the 1000G uncovered more unique variants; approximately 75.3% (1,068,228 out of 1,418,417) that were uniquely imputed from 1000G are indels or structural variants, a category of variation that is not available in the other two reference panels. We observed that imputation from the HRC panel resulted in more extremely rare variants (MAF less than 0.05%) than from HUNT WGS and 1000G. Imputation from the HUNT WGS panel uncovered more variants with MAF between 0.05% and 1% than the other two reference panels (Table 3-3). Approximately 3.6 million variants were uniquely imputed by the HUNT WGS panel (Figure 3-1) and the majority of them have MAF less than or equal to 0.05% (Figure 3-2). A threshold ≥ 0.3 for ImpRsq was applied as recommended to remove most of poorly imputed variants while retaining the vast majority of well imputed SNPs(Li, et al., 2009).

We observed that the average EmpRsq remained above 0.6 for all MAF categories from all three reference panels when the ImpRsq ≥ 0.3 threshold was applied (Figure S3-2).

3.1.7 Comparing imputation accuracy from different reference panels

To compare the imputation accuracy across the three reference panels, we examined all 289,376 variants that were directly genotyped by the chip array and available in all three reference panels. "Leave-one-variant-out" imputation results were used for these directly genotyped variants, meaning that one-by-one, each genotyped variant was masked, imputed, and then compared to the directly genotyped calls. The EmpRsq was estimated for each genotyped variant from each panel, which is the squared Pearson correlation between the imputed allele dosages and the genotypes called by direct genotyping. Figure 3-3a compares the average EmpRsq for all genotyped variants categorized by MAF among different reference panels. The MAF is estimated using the genotypes called by the chip array. Imputation from HRC has higher imputation accuracy for rare variants with MAF < 0.5% than the other two reference panels, which is expected because the number of samples available in HRC is much larger than the other two panels and the imputation accuracy for extremely rare variants depends on the number of copies of alternate alleles(Roshyara and Scholz, 2015). What is unexpected is that for variants with MAF ≥ 0.5%, HRC and HUNT WGS panels show comparable imputation accuracy, even though the size of the HUNT WGS panel is 15 times smaller than HRC. Consistent to previous studies, this result demonstrated the value of whole-genome sequencing for ancestry matched samples as a reference panel for genotype imputation(Deelen, et al., 2014; Huang and Tseng, 2014; Huang, et al., 2015; Low-Kam, et al., 2016; Okada, et al., 2015; Pistis, et al., 2015; Roshyara and Scholz, 2015; Walter, et al., 2015). It is also noticed that imputation from 1000G has lower average ImpRsq than the other two reference panels (Figure 3-3b-d), which is consistent to the lower proportion of variants passing the various ImpRsq thresholds in 1000G observed in Figure S3-2.

To further evaluate the impact of the sample size of the HUNT WGS panel on the imputation accuracy, we have randomly drawn 500, 1000, and 1500 samples from the original HUNT reference panel for imputation. Figure S3-3 shows the comparison of the average EmpRsq for all genotyped variants categorized by MAF among the target samples, across all reference panels. As expected, increases in the sample size of the HUNT WGS reference panels resulted in higher imputation accuracy, particularly for less frequent variants with MAF < 0.5%. Interestingly, we observed that the HUNT WGS with 500 samples outperforms 1000G(Auton, et al., 2015) for variants with MAF > 0.5%. These results are consistent with other studies with population specific reference panels(Mitt, et al., 2017; Pistis, et al., 2015). The subset of 1000 samples provides better imputation accuracy than 1000G(Auton, et al., 2015) even for variants with MAF as low as 0.1% and comparable imputation accuracy to HRC(McCarthy, et al., 2016) for variants with MAF > 0.5%.

We examined whether our evaluation of imputation accuracy is biased in favor of HUNT WGS due to relatedness. Previous studies have shown that the relatedness between study samples and reference samples increases genotype imputation efficiency since related individuals tends to share longer haplotype stretches than unrelated ones(Huang and Tseng, 2014). To avoid the bias of imputation accuracy due to the relatedness between our study samples and the samples in the HUNT WGS reference panel, we excluded 1,644 study samples who are up to 2nd degree relatives of HUNT WGS samples. Relatedness was based on the estimation of the proportion of IBD by PLINK(Purcell, et al., 2007). We observed that excluding these study samples did not affect the imputation accuracy except causing a slight decrease of the imputation accuracy for those very rare variants with MAF < 0.05% (Figure S3-4).

3.1.8 Evaluating two possible association test strategies to use multiple sets of imputed genotypes

As Figure 3-1 shows, approximately 60% of all successfully imputed variants were imputed from more than one reference panel, which makes it unclear how to perform downstream association tests. We compared two possible strategies: the "best p-value" and the "best Rsq" approaches. The "best p-value" approach uses each version of imputed genotypes to choose the lowest association p-value, thereby increasing the burden of adjusting for multiple hypothesis testing. The "best Rsq" approach selects the imputed variant with the highest estimated imputation quality ImpRsq, which is expected to be a reasonable approximation of the association between imputed and true genotypes, especially for common variants (Figure S3-5).

We have compared the power of the two approaches to detect association signals accounting for the fact that the "best p-value" approach needs adjusting for the additional variants tested. To determine the significant thresholds for association tests with a family-wise error rate (FWER) 0.05, we estimated the number of independent tests using 1,000 permutations. For the "best Rsq" approach, where fewer 'variants' are analyzed, the significance threshold is 4.69×10^{-9} (2.10×10^{-9} with a Bonferroni correction) and for the best p-value approach, it is 2.53×10^{-9} (1.05×10^{-9} with a Bonferroni correction).

Using the permutation-derived significance thresholds above, we evaluated the power of the two approaches for association tests with quantitative traits through a simulation study (details described in methods). Our results indicated that the "best p-value" approach has more power to detect association signals than the "best Rsq" approach, particularly for rare variants with MAF <

1%, no matter how stringent the ImpRsq threshold was used for filtering out the poorly imputed genotypes (Figure 3-4, Figure S3-6 and Table S3-1). This is probably because the estimated imputation quality ImpRsq does not always agree with empirical imputation quality EmpRsq especially for rare variants (Figure S3-5), resulting in loss of variants with highest empirical imputation quality when selecting the "best Rsq" strategy. In addition, the distributions of the ImpRsq are quite different from different panels. Notably, from 1000G(Auton, et al., 2015), the ImpRsq and EmpRsq were substantially lower for low-frequency variants (0.5% < MAF < 5%), and ImpRsq tends to underestimate EmpRsq (Figure S3-5). The two approaches have comparable association power for variants with MAF ≥ 1%, where estimated and empirical imputation qualities highly agree with each other (Figure S3-5). Our observation suggests that the inaccurate prediction of imputation quality have a higher impact than increased burden of multiple testing in association test with rare variants.

3.1.9 Evaluating net gain of imputation accuracy by including an additional reference panel

Finally, we quantified the net gain of imputation accuracy by including an additional reference panel as a "partial Rsq" conditioned on the imputed genotypes from an existing reference panel (See Materials and Methods for details). Intuitively, this represents the difference between the "optimal EmpRsq" linearly combined between two sets of imputed genotypes and the EmpRsq from the original imputed genotypes. 289,376 genotyped variants that were found in all three panels were used to evaluate the additional information that were gained from one reference panel given imputed dosages based on another panel. As Figure S3-7 presents, each reference panel is able to provide additional information to improve imputation accuracy. However, relatively less information could be gained by including 1000G(Auton, et al., 2015) panel on top of HRC across all MAF categories. This is expected since 1000G samples are included in the HRC panel, with the caveat that only single nucleotide variants with minor allele count ≥ 5 were retained. Note that evaluation of indels and structural variants absent in HRC were not included in this experiment. In contrast, given the imputed dosages from 1000G, both HUNT WGS and HRC provide substantial net gain of imputation accuracy, which is consistent to our observations. Furthermore, HUNT WGS and HRC provide additional information conditional on each other. More specifically, more extra information was obtained from HRC given HUNT WGS than those were obtained from HUNT WGS given HRC for these genotyped variants, which is also consistent to our observations in Figure 3-3.

3.4 Discussion

Many studies have performed whole genome sequencing of a subset of samples followed by imputation into samples with GWAS data(Holm, et al.; Lane, et al., 2016; Nalls, et al., 2014; van Leeuwen, et al., 2016). However, the trade-offs between the panel size, imputable variant types, and population specificity across different reference panels make it challenging to decide on the optimal strategy for imputation and downstream association analysis. We evaluated methods for genotype imputation when different reference panels are available. Our findings have demonstrated the benefits of uncovering novel variants with low frequency by using population-specific reference panels as has been reported by previous studies(Huang, et al., 2015). Since the population-specific HUNT panel shared 1,023 samples with HRC(McCarthy, et al., 2016), we expect to see an even bigger advantage in the number of novel low frequency variants imputed by the population-specific panel if there were no overlap between the two reference panels.

We have also observed that large-scale publicly available reference panels, as exemplified by HRC (McCarthy, et al., 2016) and 1000G(Auton, et al., 2015), contribute a large number of variants that are not captured by population-specific reference panels. More specifically, HRC(McCarthy, et al., 2016), which has much larger sample size and contains more general European populations, contributes 3.5 million variants that could not be imputed by the other two panels. Since 1000G(Auton, et al., 2015) has additional advantages that indels and structural variants are comprehensively detected and genotyped, 1.3 million non-SNP variants have only been imputed by 1000G(Auton, et al., 2015). Furthermore, each reference panel may provide additional information to improve imputation accuracy. Therefore, to increase the variant coverage and imputation accuracy as much as possible, we recommend using all three reference panels for imputation if available. If a single panel has to be chosen, each option will have different advantages and disadvantages. We have shown that imputation from population-specific reference panels provides comparable imputation accuracy for variants with $\mathrm{MAF} > 0.1\%$. as using reference panels with 15 times larger sample size with only broad ancestry-matching (i.e. European). Although panel sizes are similar, the population-specific reference panel results in higher imputation accuracy than the mixed-ancestry 1000G panel (Auton, et al., 2015) for variants with MAF ≥ 0.05%. This has also been observed by a recently published study on Estonians(Mitt, et al., 2017).

To address the issue of imputing different versions of the same variant from different reference panels, we propose the "best p-value" approach, which analyzes all versions of each imputed variant and accounts for the multiple testing. Our simulation study demonstrated that this approach has higher power for detecting association signals than selecting the imputed variant with highest imputation quality given the distributions of the imputation quality metrics from different reference panels may be quite different, even adjusting for additional variants tested.

The UK10K study and the Genome of the Netherlands (GoNL) Consortium suggested that merging multiple reference panels to a larger reference panel would improve imputation performance, especially for less frequent variants(Deelen, et al., 2014; Huang, et al., 2015). Compared to this approach, our "best p-value" approach does not require access to all reference panes and is feasible even if not all reference panel haplotypes are directly accessible. If large imputation reference panels, such as the HRC(McCarthy, et al., 2016), are not directly accessible, conducting association tests for all imputed versions of genotype with slightly higher computational cost will be an effective strategy.

In summary, we recommend creating a small size ancestry-matched reference panel using whole genome sequencing to allow for improved imputation of low frequency variants that may be enriched in that ancestral group, performing genotype imputation using the ancestry-matched reference panel and other large publicly available databases, and analyzing all versions of imputed variants in downstream association testing.

Figure 3-1 Number of variants that are imputed by different reference panels.

The corresponding percentage is the variants number out of all 23.8 million variants that are successfully imputed by any of the three reference panels.

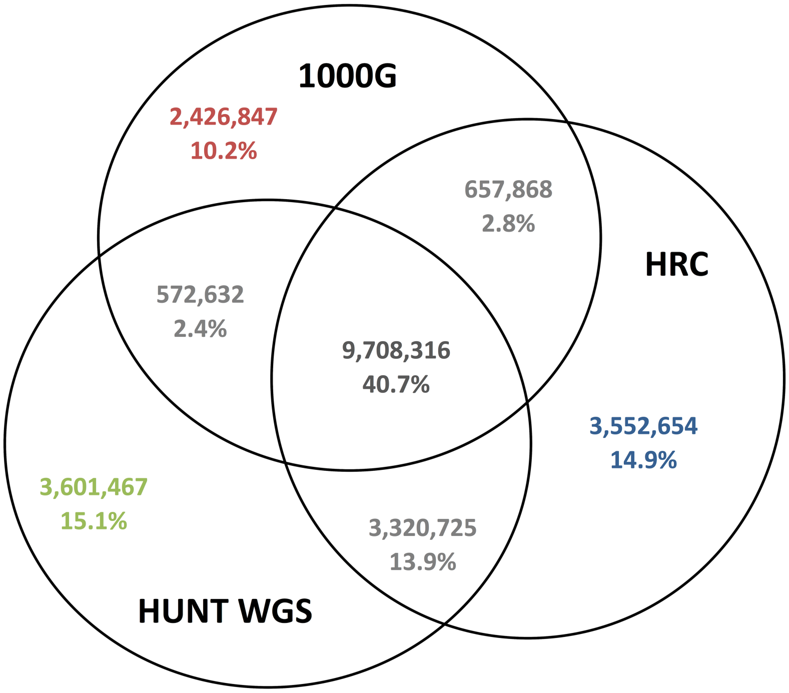


Figure 3-2 Distribution of numbers of variants that are imputed from only one reference panel or from multiple reference panels in different MAF categories.

Variants that are imputed by 1000G only are categorized as SNPs and non-SNP variants, including indels, deletions, complex short substitutions and other structural variant classes. 1000G, 1000 Genomes Phase 3; WGS, whole-genome sequencing; HRC, Haplotype Reference Consortium; MAF, minor allele frequency

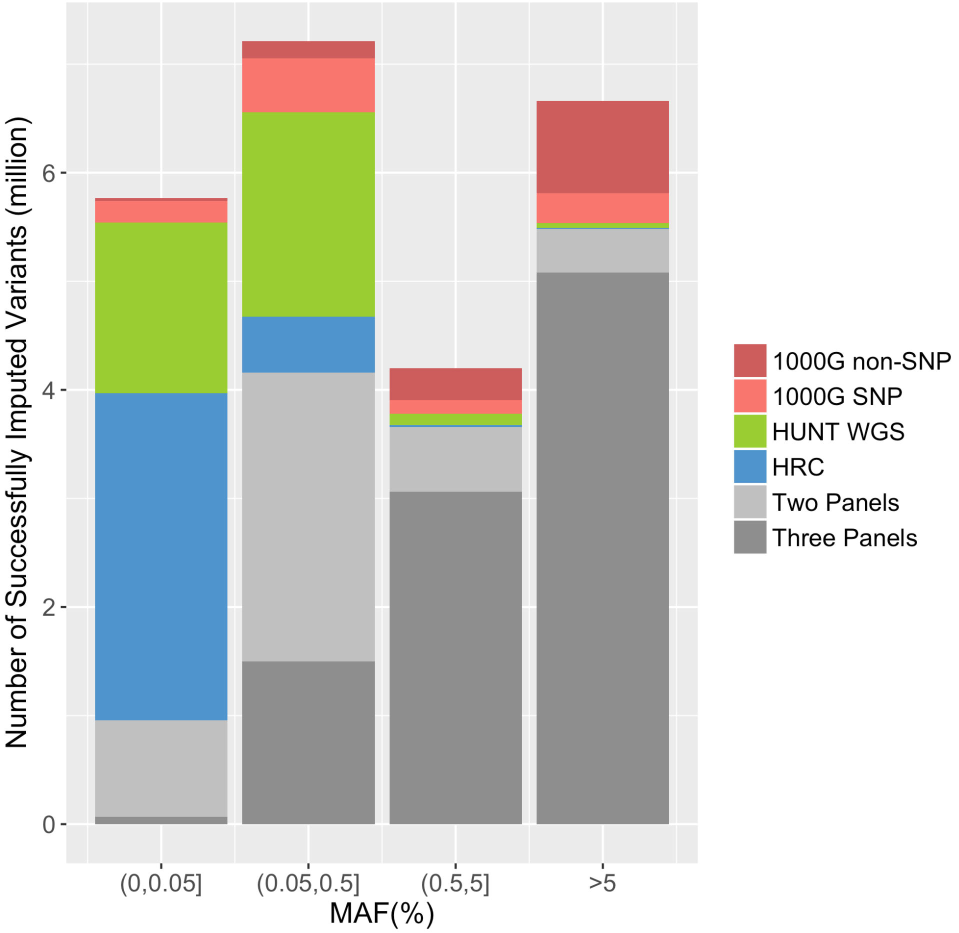


Figure 3-3 HRC and HUNT WGS panels show comparable imputation quality.

a. comparing the mean empirical R^2 (y axis) reported by different reference panels for variants that are directly genotyped categorized by the MAF (x axis) without any ImpRsq threshold applied.

b. comparing the mean Imputation R^2 (y axis) reported by different reference panels for variants that are directly genotyped categorized by the MAF (x axis) without any ImpRsq threshold applied.

c. comparing the mean Imputation R^2 (y axis) reported by different reference panels for all imputed variants (ImpRsq > 0.3) by the MAF (x axis).

d. comparing the mean Imputation R^2 (y axis) reported by different reference panels for all imputed variants by the MAF (x axis) without any ImpRsq threshold applied.

1000G, 1000 Genomes Phase 3; WGS, whole-genome sequencing; HRC, Haplotype Reference Consortium; MAF, minor allele frequency; ImpRsq, imputation quality metric R^2

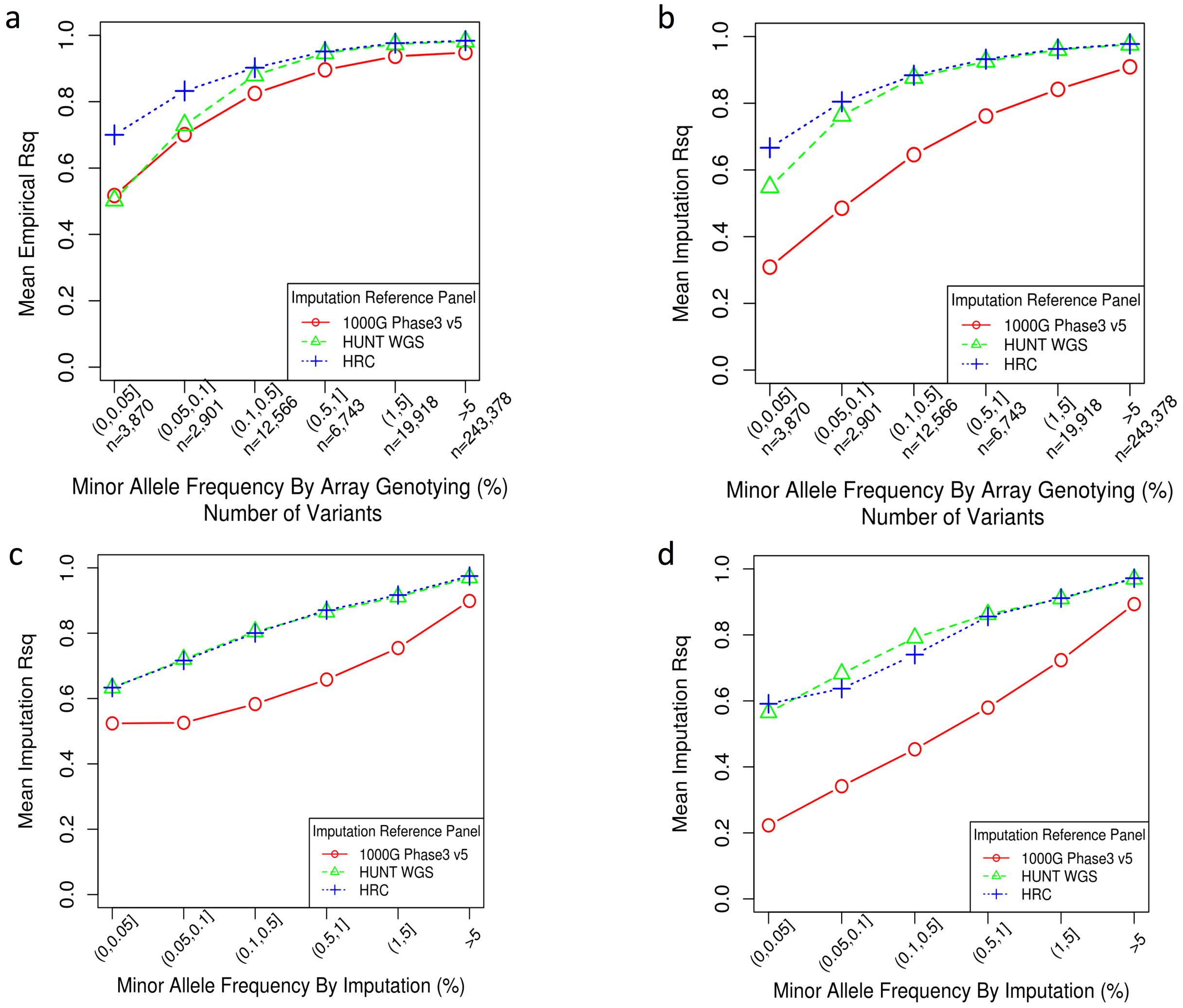


Figure 3-4 Comparison of power to detect true associations between best p-value and best Rsq approaches via simulation studies.

For each MAF category, 3,000 directly genotyped variants were randomly selected based on their MAF estimated with genotypes obtained from the chip array to estimate the power. The power is calculated as the proportion of significantly associated variants across three imputed panels based on each strategy given the corresponding significance threshold. ImpRsq > 0.3 was applied to remove poorly imputed genotypes. The numbers of variants that were successfully imputed from at least two reference panels and used in the simulation studies are: 2,513 with MAF > 0 and ≤ 0.001; 2,989 with MAF > 0.001 and ≤ 0.01; 3,000 with MAF > 0.01 and ≤ 0.05; and 3,000 with MAF > 0.05. MAF, minor allele frequency; ImpRsq, imputation quality metric R^2

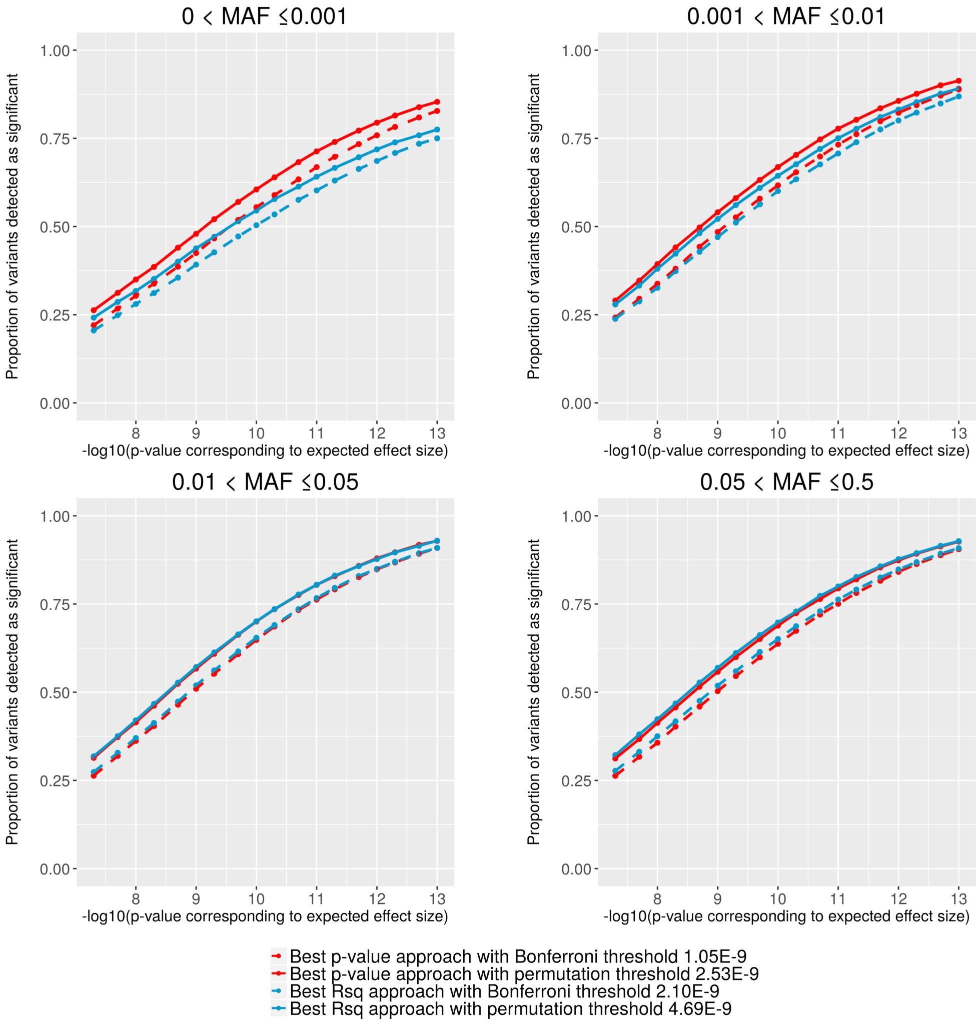


Table 3-1 Summary of the variants in the HUNT whole-genome sequencing reference panel containing 2,201 individuals with average sequencing depth 5x.

Variant Type	**Total number of variants**	**Mean number of variants per individual (SD)**	**Mean number of unique variants per individual (SD)**	**% in 1000 Genomes**	**Number of novel variants***
Splice	1,265	71.5(4.6)	0.2(0.47)	36.6	355
Nonsense	2,432	71.5(6)	0.43(0.74)	36.6	585
Missense	113,576	9,480(113)	13.8(13.6)	56.3	13,927
Synonymous	77,699	10,707(100)	7.1(7.5)	68.5	5,935
Noncoding	20,050,237	3,342,839(15,415)	1531(906)	68.7	4,030,199
Total	20,245,209	3,363,168(15,522)	1,552(919)	68.6	4,051,001

*Novel: not reported in dbSNP 144(Sherry, et al., 2001), 1000 Genomes Phase 3(Auton, et al., 2015), UK10K(Walter, et al., 2015), ESP6500(2013), or ExAC.r0.3(Lek, et al., 2016)

Table 3-2 Reference panels used for genotype imputation

MAC: minor allele count

Reference Panels	Variants	Sample Size	Population
Haplotype Reference Consortium(McCarthy, et al., 2016) (HRC)	39 million SNPs (MAC ≥ 5)	32,488[a]	Cosmopolitan (mostly European)
1000 Genomes Phase 3 Version 5(Auton, et al., 2015) (mean depth < 8x)	81 million Biallelic SNPs, indels, deletions, complex short substitutions and other structural variant classes (MAC ≥ 2)	2,504	Cosmopolitan
HUNT Whole Genome Sequencing (HUNT WGS) (mean depth ~ 5x)	20 million SNPs	2,201[a]	Norwegian

[a] HRC and HUNT whole-genome sequencing data set have 1,023 samples in overlap.

Table 3-3 Numbers of imputed variants contributed by each reference panel categorized by MAF.

The threshold ImpRsq > 0.3 was applied. Each reference panel contributed uniquely imputed variants. The greatest number of the uniquely imputed variants among the three reference panels for variants in each MAF category is highlighted in red. MAF: minor allele frequency; ImpRsq, imputation quality metric R^2

	HRC Release 1 (39.2M SNPs, 32,488 samples including 1,203 HUNT samples)			1000G Phase3 v5 (81.2M markers, 2,504 samples)			HUNT 5x WGS (20.2M SNPs, 2,201 samples)		
MAF	Number of Passed Variants	Percent of Passed Variants	Number of Uniquely Imputed Variants	Number of Passed Variants	Percent of Passed Variants	Number of Uniquely Imputed Variants	Number of Passed Variants	Percent of Passed Variants	Number of Uniquely imputed Variants
(0, 0.0005)	4,337,138	23.9%	3,009,729	567,481	2.4%	230,186	2,291,216	50.6%	1,570,259
(0.0005, 0.001)	1,339,096	91.1%	373,964	501,248	11.4%	176,252	1,668,837	94.4%	901,106
(0.001, 0.005)	2,964,988	97.5%	140,318	2,119,956	33.6%	475,376	3,917,801	98.0%	982,320
(0.005, 0.01)	1,125,181	99.2%	7,426	1,074,885	68.9%	126,616	1,279,200	98.6%	47,426
(0.01, 0.05)	2,314,490	99.6%	10,525	2,554,206	89.2%	295,991	2,538,140	99.1%	55,490
> 0.05	5,158,670	99.8%	10,692	6,547,887	98.1%	1,122,426	5,507,946	99.6%	44,866
Total	17,239,563	55.1%	3,552,654	13,365,663	29.5%	2,426,847	17,203,140	87.4%	3,601,467

Figure S3-1 The principle component plots for the individuals in the HUNT WGS reference panel and the HUNT study samples.

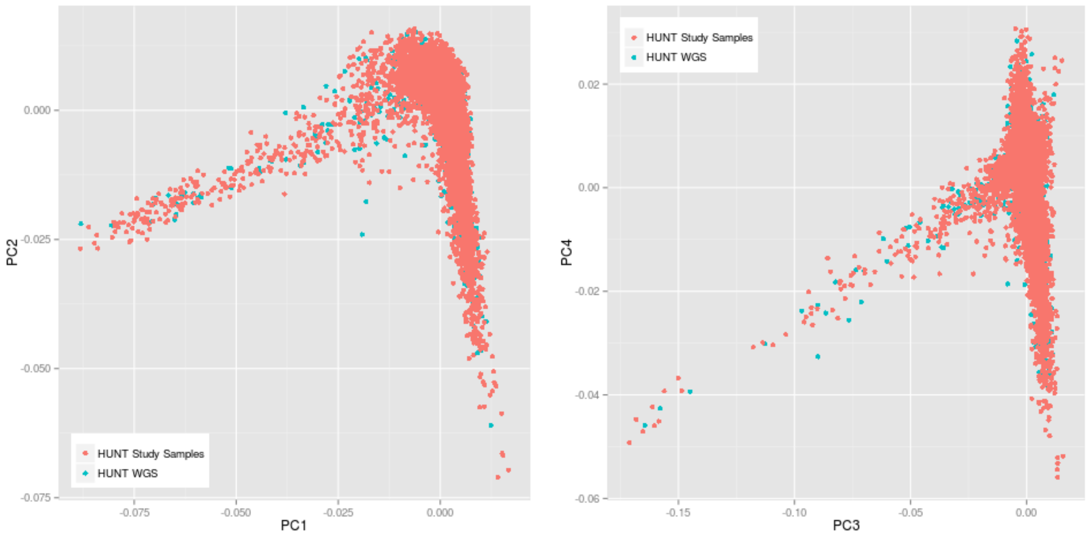


Figure S3-2 Average EmpRsq (top panel) and the proportion of variants that passed the ImpRsq thresholds (bottom panel) are plotted against ImpRsq thresholds for post-imputation QC for variants in different MAF categories.

Reference panels include HRC(McCarthy, et al., 2016), 1000 Genomes Phase 3(Auton, et al., 2015). and the HUNT WGS. The dashed line is for the ImpRsq threshold ≥ 0.3.

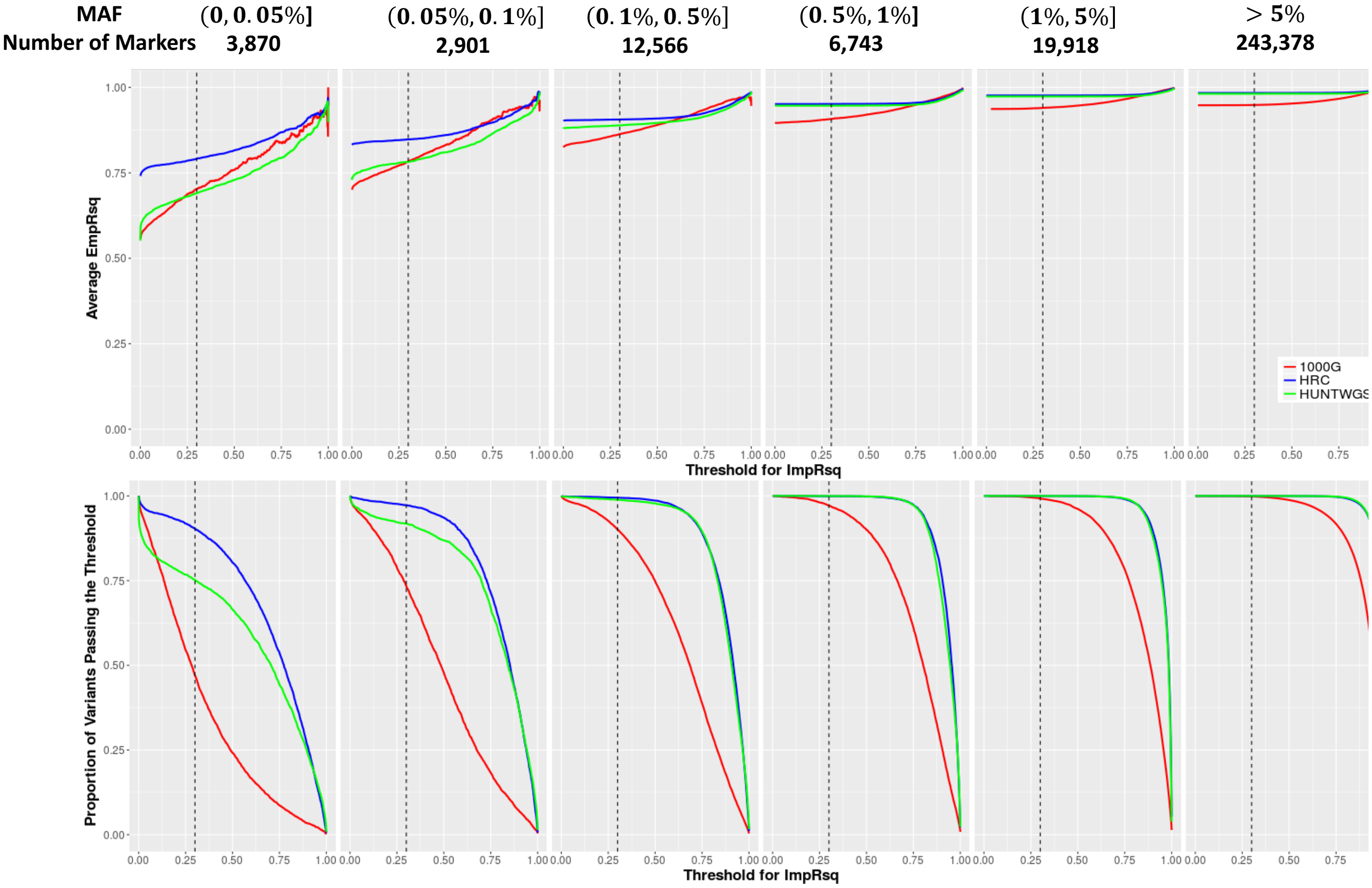


Figure S3-3 Comparing the mean empirical R^2 (y axis) reported by all four HUNT WGS reference panels, HRC(McCarthy, et al., 2016) and 1000G(Auton, et al., 2015) for all 289,376 variants that were directly genotyped categorized by the MAF (x axis) of the target samples.

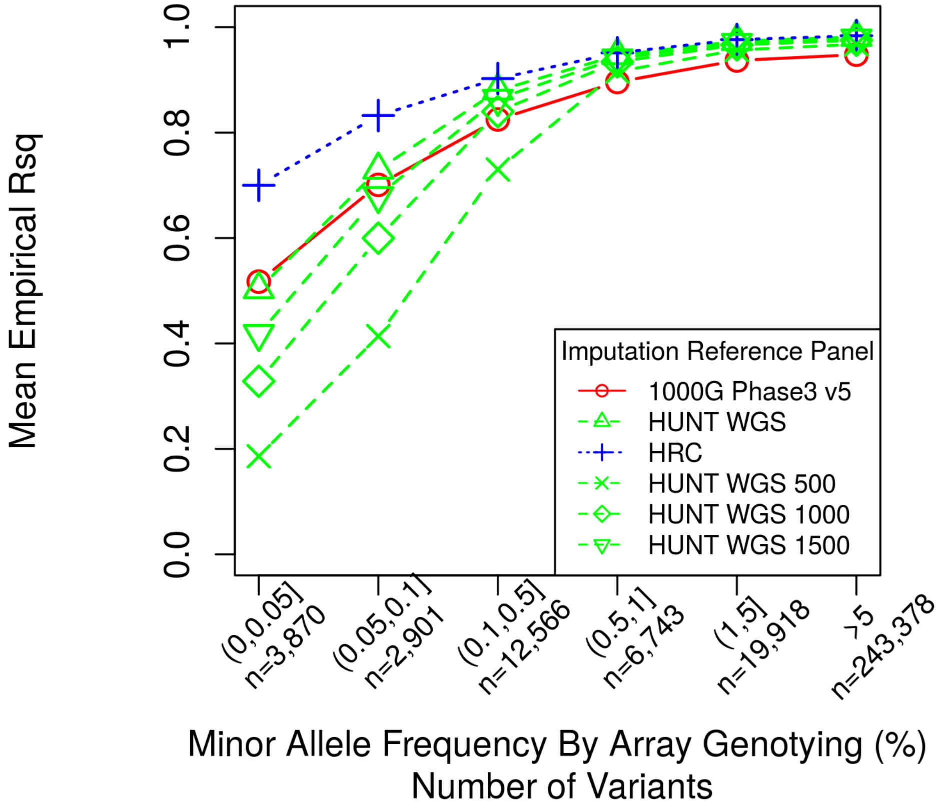


Figure S3-4 Imputation of the study data set with and without the 1,644 samples, who are closely related to samples in the HUNT WGS, has been performed.

For variants that were directly genotyped by the chip array, average EmpRsq are plotted by different MAF categories.

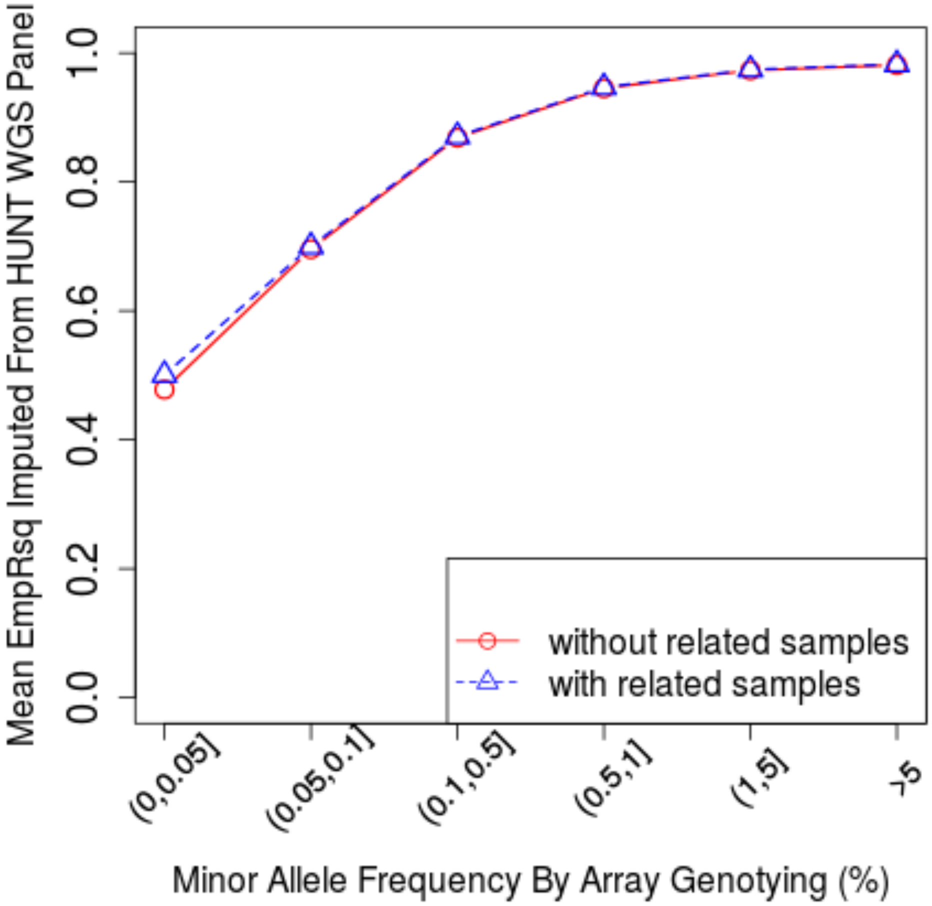


Figure S3-5 Relationships between ImpRsq (imputation quality metric estimated by minimac3) and EmpRsq (correlation between imputed and true genotypes).

The plots are based on 289,376 directly genotyped variants that were found in all three panels. Reference panels include A. HRC(McCarthy, et al., 2016).. B. 1000G Phase 3(Auton, et al., 2015). C. HUNT WGS. MSE: mean square errors.

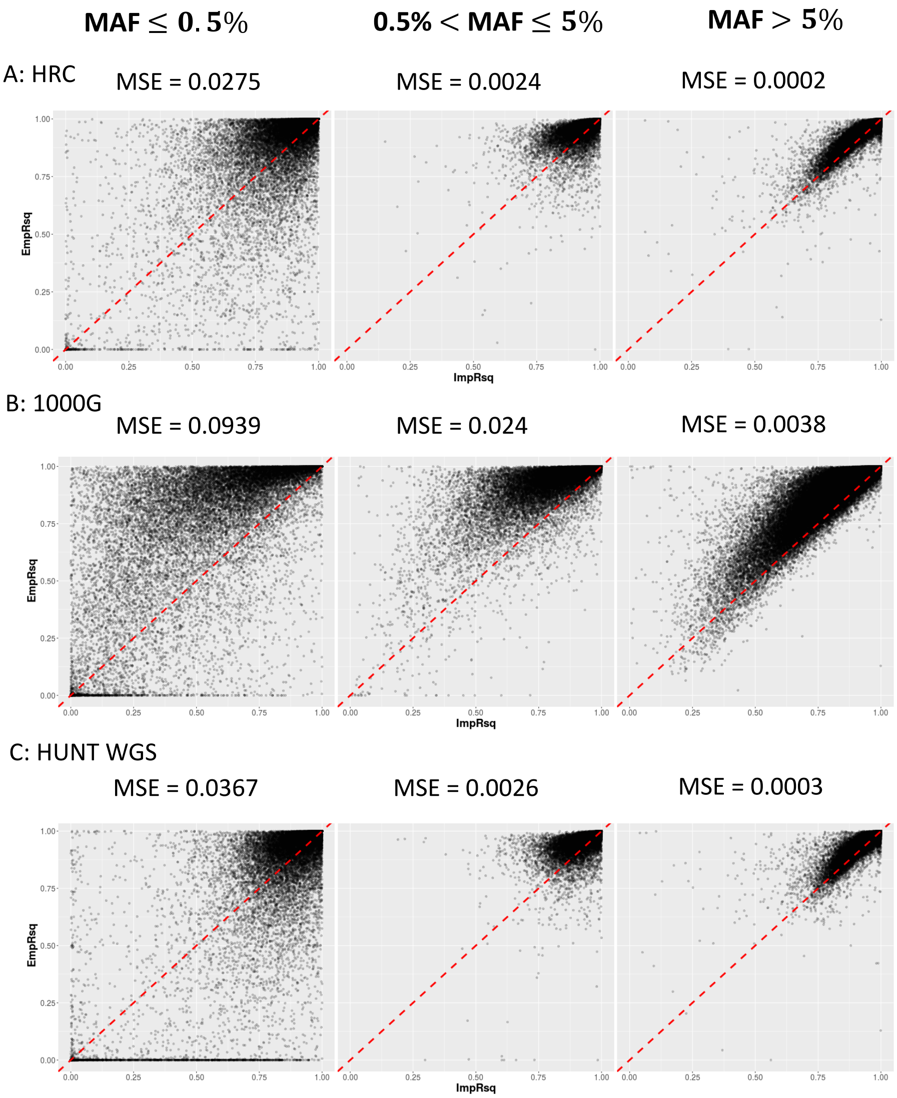


Figure S3-6 Comparison of power to detect true associations between best p-value and best Rsq approaches via simulation studies for variants with MAF ≤ 0.1%.

Different ImpRsq cutoffs were used to remove poorly imputed genotypes. Only variants with at least two versions of successfully imputed genotypes were used in the simulation study. 2,513 variants passed the threshold ImpRsq > 0.3, 2,095 variants passed the threshold ImpRsq > 0.5, and 899 variants passed the threshold ImpRsq > 0.8.

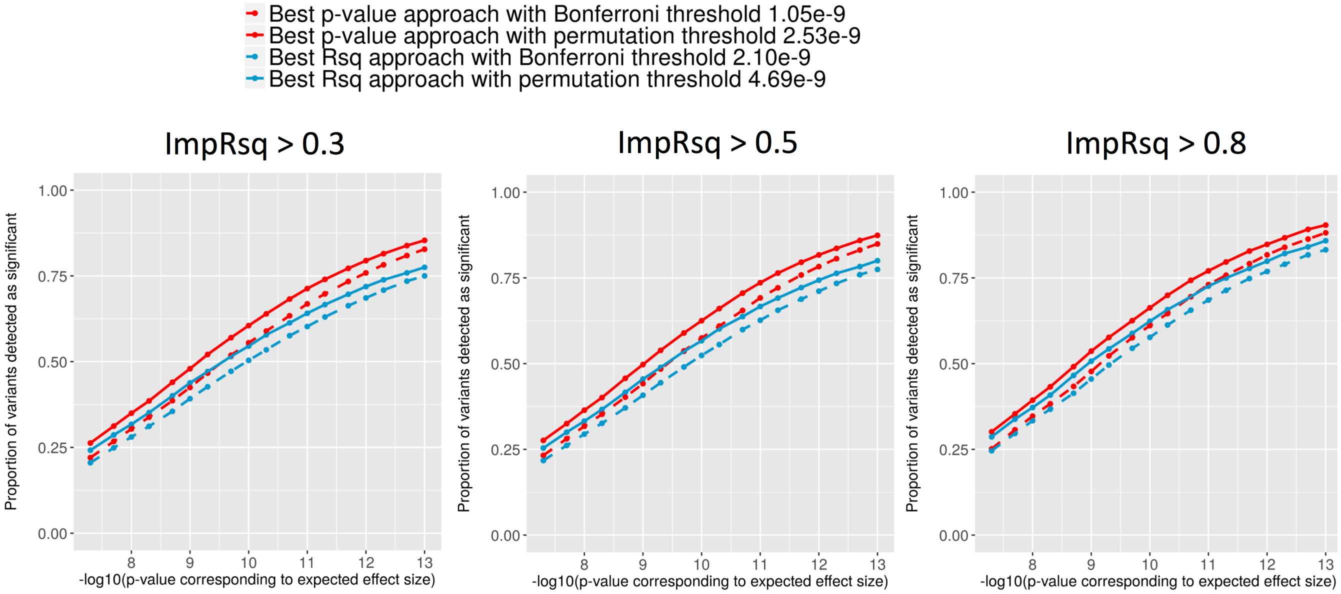
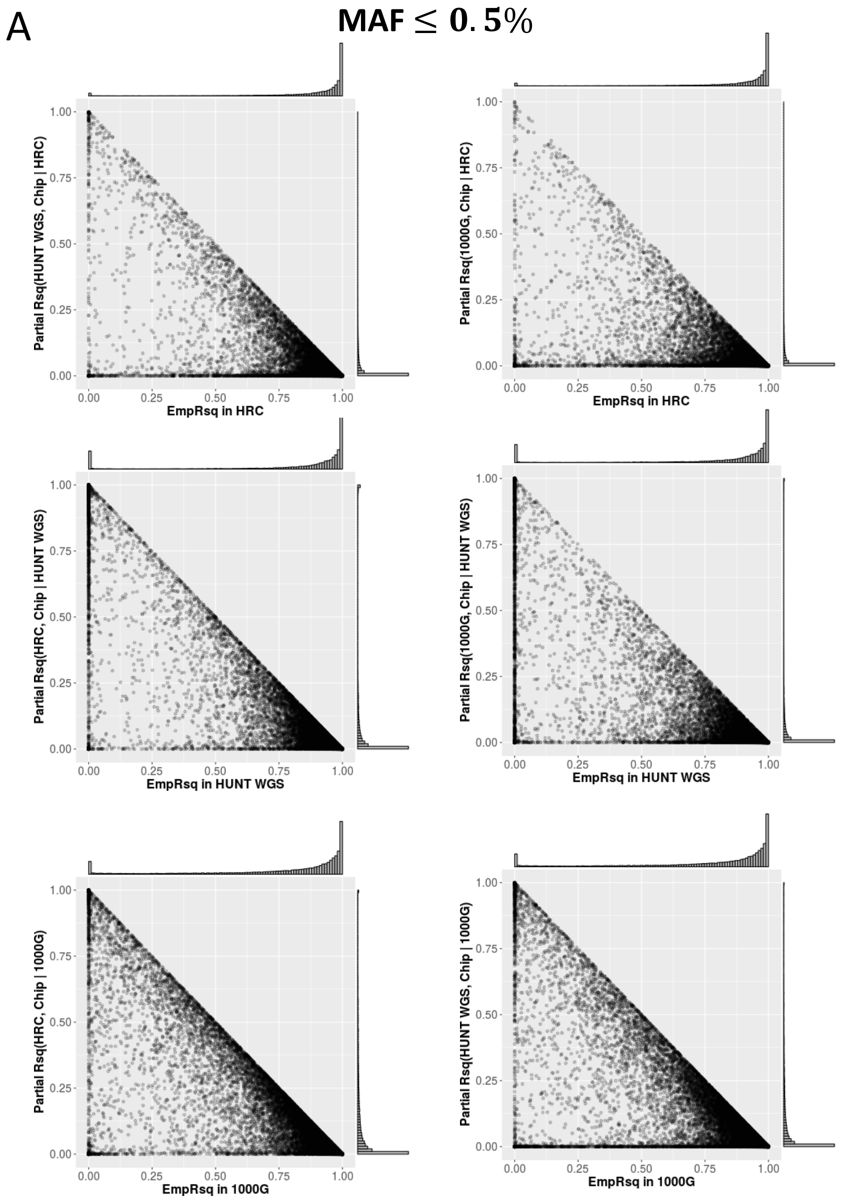
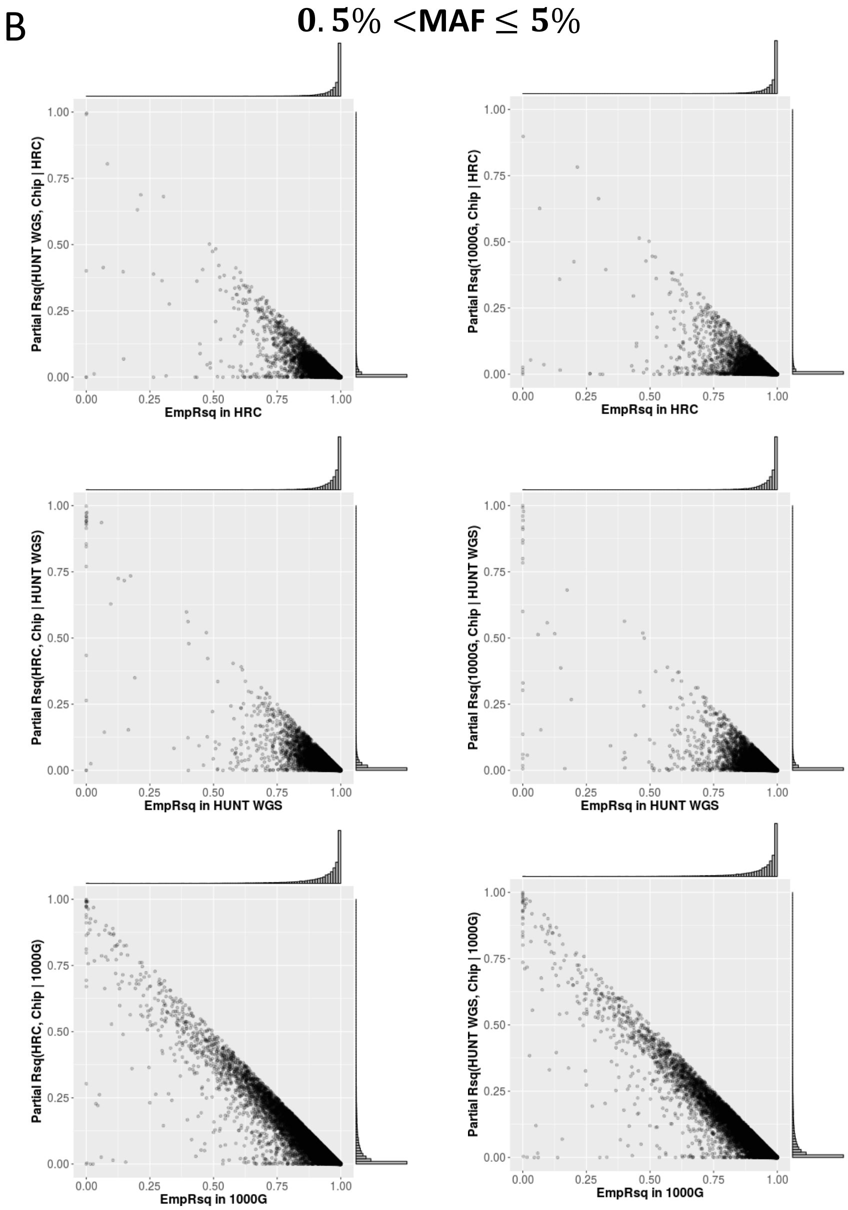


Figure S3-7 Plots of partial correlation between imputed dosages by one reference panel and the chip genotypes given imputed dosages by another reference panel.

The plots only contain the genotypes variants found in all three reference panels: HRC(McCarthy, et al., 2016), 1000G Phase 3(Auton, et al., 2015), and HUNT WGS. MAF was calculated based on the chip genotypes. A. 19,337 variants with MAF ≤ 0.5%. B. 26,661 variants with MAF between 0.5% and 5%. C. 243,378 variants with MAF > 5%.




B
$0.5\% <$MAF $\leq 5\%$
Partial Rsq(HUNT WGS, Chip | HRC)
EmpRsq in HRC
Partial Rsq(1000G, Chip | HRC)
EmpRsq in HRC
Partial Rsq(HRC, Chip | HUNT WGS)
EmpRsq in HUNT WGS
Partial Rsq(1000G, Chip | HUNT WGS)
EmpRsq in HUNT WGS
Partial Rsq(HRC, Chip | 1000G)
EmpRsq in 1000G
Partial Rsq(HUNT WGS, Chip | 1000G)
EmpRsq in 1000G
0.00
0.25
0.50
0.75
1.00

C

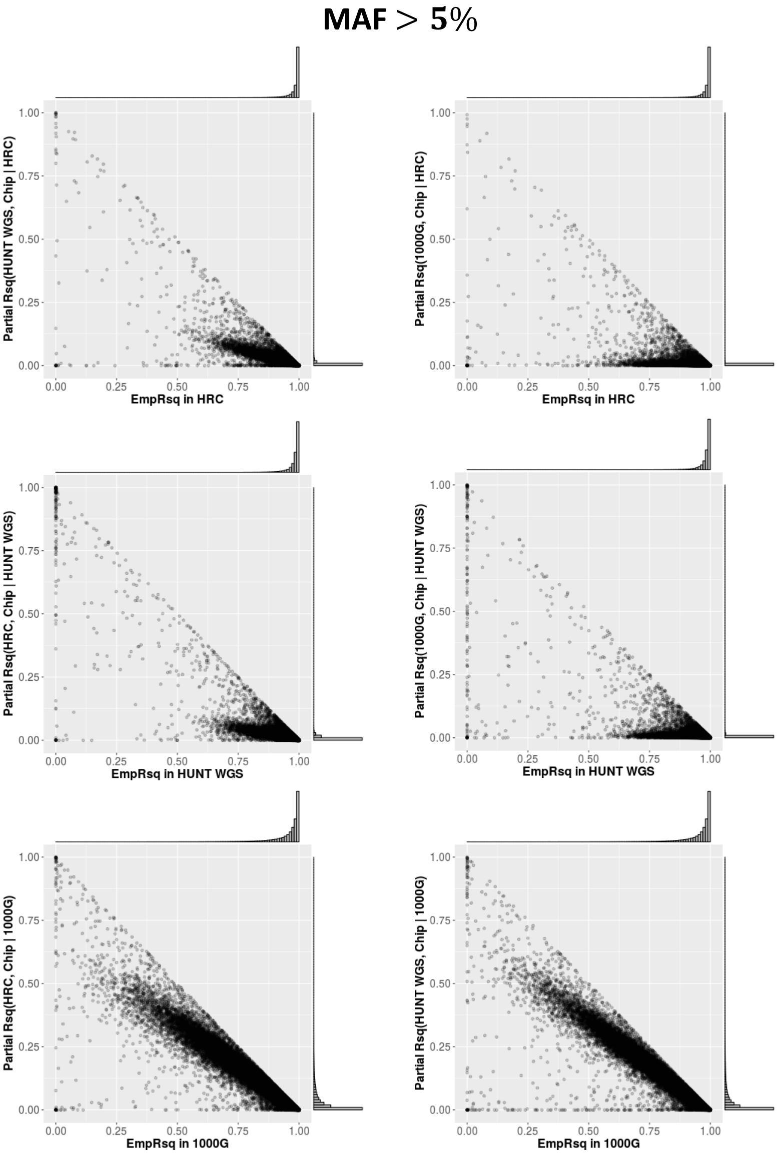

MAF > 5%
Partial Rsq(HUNT WGS, Chip | HRC)
EmpRsq in HRC
Partial Rsq(1000G, Chip | HRC)
EmpRsq in HRC
Partial Rsq(HRC, Chip | HUNT WGS)
EmpRsq in HUNT WGS
Partial Rsq(1000G, Chip | HUNT WGS)
EmpRsq in HUNT WGS
Partial Rsq(HRC, Chip | 1000G)
EmpRsq in 1000G
Partial Rsq(HUNT WGS, Chip | 1000G)
EmpRsq in 1000G
0.00
0.25
0.50
0.75
1.00

Table S3-1 Percentage of variants detected as significant (association p value reaches the significant p threshold) using the best p and the best R^2 methods when the ImpRsq > 0.3 was applied.

True p	0 < MAF <= 0.001				0.001 < MAF <= 0.01				0.01 < MAF <= 0.05				MAF > 0.05			
	Best p with permutation threshold $2.53x10^{-9}$	Best Rsq with permutation threshold $4.69x10^{-9}$	Best p with Bonferroni threshold $1.05x10^{-9}$	Best Rsq with Bonferroni threshold $2.10x10^{-9}$	Best p with permutation threshold $2.53x10^{-9}$	Best Rsq with permutation threshold $4.69x10^{-9}$	Best p with Bonferroni threshold $1.05x10^{-9}$	Best Rsq with Bonferroni threshold $2.10x10^{-9}$	Best p with permutation threshold $2.53x10^{-9}$	Best Rsq with permutation threshold $4.69x10^{-9}$	Best p with Bonferroni threshold $1.05x10^{-9}$	Best Rsq with Bonferroni threshold $2.10x10^{-9}$	Best p with permutation threshold $2.53x10^{-9}$	Best Rsq with permutation threshold $4.69x10^{-9}$	Best p with Bonferroni threshold $1.05x10^{-9}$	Best Rsq with Bonferroni threshold $2.10x10^{-9}$
1.00	85.3%	77.5%	82.8%	75.0%	91.3%	89.1%	88.9%	86.9%	92.9%	92.9%	90.9%	90.9%	92.7%	92.8%	90.5%	90.8%
2.00	83.8%	75.9%	80.9%	73.5%	90.0%	87.7%	87.1%	84.9%	91.8%	91.5%	89.2%	89.4%	91.3%	91.5%	88.8%	89.3%
5.00	81.5%	73.9%	78.2%	70.9%	87.6%	85.3%	84.4%	82.3%	89.7%	89.6%	86.8%	87.0%	89.3%	89.4%	86.3%	86.9%
1.00	79.4%	71.9%	75.9%	68.6%	85.6%	83.1%	82.2%	80.1%	87.9%	87.7%	84.8%	85.0%	87.4%	87.7%	84.1%	84.8%
2.00	77.2%	69.7%	73.4%	66.3%	83.5%	81.0%	79.9%	77.5%	85.8%	85.7%	82.6%	82.9%	85.4%	85.7%	81.6%	82.5%
5.00	74.0%	66.6%	69.8%	63.0%	80.3%	77.7%	76.2%	73.9%	82.9%	83.0%	79.2%	79.6%	82.0%	82.7%	78.1%	79.1%
1.00	71.3%	64.1%	66.9%	60.2%	77.7%	75.0%	73.2%	70.7%	80.4%	80.4%	76.3%	76.7%	79.4%	80.0%	75.1%	76.3%
2.00	68.3%	61.3%	63.4%	57.6%	74.7%	72.0%	69.8%	67.6%	77.6%	77.7%	73.3%	73.6%	76.4%	77.3%	72.0%	72.9%
5.00	63.9%	57.8%	58.9%	53.5%	70.3%	67.7%	65.4%	63.4%	73.5%	73.5%	68.6%	69.0%	72.4%	72.9%	67.4%	68.7%
1.00	60.5%	54.5%	55.5%	50.4%	66.9%	64.4%	61.7%	60.1%	70.0%	70.1%	64.8%	65.4%	68.9%	69.8%	63.7%	65.1%
2.00	57.0%	51.5%	51.9%	47.2%	63.2%	60.9%	57.9%	56.3%	66.3%	66.4%	60.8%	61.6%	65.1%	66.2%	59.9%	61.4%
5.00	52.1%	47.1%	46.7%	42.7%	58.1%	56.1%	52.6%	51.1%	60.9%	61.3%	55.3%	56.2%	60.0%	61.1%	54.6%	55.9%
1.00	47.9%	43.8%	42.5%	39.2%	54.0%	52.1%	48.5%	47.0%	56.7%	57.1%	51.0%	51.9%	55.8%	56.9%	50.3%	51.8%
2.00	44.0%	40.1%	38.6%	35.5%	49.7%	48.1%	44.3%	42.8%	52.4%	52.7%	46.5%	47.3%	51.6%	52.7%	45.9%	47.6%
5.00	38.6%	35.2%	33.8%	31.2%	44.1%	42.3%	38.0%	37.3%	46.2%	46.7%	40.4%	41.3%	45.7%	46.8%	40.3%	41.7%
1.00	35.0%	31.7%	30.4%	28.1%	39.4%	38.0%	33.8%	32.6%	41.5%	42.0%	36.2%	37.0%	41.3%	42.4%	35.7%	37.5%
2.00	31.2%	28.6%	26.8%	24.9%	34.7%	33.3%	29.5%	28.8%	37.3%	37.5%	31.9%	32.8%	36.7%	38.1%	31.7%	33.1%
5.00	26.3%	24.2%	22.0%	20.6%	29.0%	27.9%	24.2%	23.8%	31.4%	31.9%	26.3%	27.4%	31.2%	32.2%	26.3%	27.7%

Chapter 4 Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

4.1 Abstract

In genome-wide association studies (GWAS) for thousands of phenotypes in large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, linear mixed model and the recently proposed logistic mixed model, perform poorly -- producing large type I error rates -- in the analysis of phenotypes with unbalanced case-control ratios. Here we propose a scalable and accurate generalized mixed model association test that uses the saddlepoint approximation (SPA) to calibrate the distribution of score test statistics. This method, SAIGE, provides accurate p-values even when case-control ratios are extremely unbalanced. It utilizes state-of-art optimization strategies to reduce computational time and memory cost of generalized mixed model. The computation cost linearly depends on sample size, and hence can be applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK-Biobank data of 408,961 white British European-ancestry samples, we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.

4.2 Introduction

Decreases in genotyping cost allow for large biobanks to genotype all participants, enabling genome-wide scale phenome-wide association studies (PheWAS) in hundreds of thousands of samples. In a typical genome-wide PheWAS**,** GWAS for tens of million variants are performed for thousands of phenotypes constructed from Electronic Health Records (EHR) and/or survey questionnaires from participants in large cohorts(Bush, et al., 2016; Cronin, et al., 2014; Denny, et al., 2013; Denny, et al., 2011; Dumitrescu, et al., 2015; Hall, et al., 2014; Hebbring, et al., 2015; Hebbring, et al., 2013; Liao, et al., 2013; Millard, et al., 2015; Moore, et al., 2015; Namjou, et al., 2014; Neuraz, et al., 2013; Pendergrass, et al., 2013; Ritchie, et al., 2013; Shameer, et al., 2014; Ye, et al., 2015). For binary traits based on disease/condition status in PheWAS, cases are typically defined as individuals with specific International Classification of Disease (ICD) codes within the EHR. Controls are usually all participants without the same or other related conditions(Bush, et al., 2016; Denny, et al., 2013). Due to the low prevalence of many conditions/diseases, case-control ratios are often unbalanced (case:control=1:10) or extremely unbalanced (case:control<1:100). The scale of data and the unbalanced nature of binary traits pose substantial challenges for genome-wide PheWAS in biobanks.

Population structure and relatedness are major confounders in genetic association studies and also need to be controlled in PheWAS. Linear mixed models (LMM) are widely used to account for these issues in GWAS for both binary and quantitative traits(Aulchenko, et al., 2007; Kang, et al., 2010; Lippert, et al., 2011; Loh, et al., 2015; Yang, et al., 2011; Zhang, et al., 2010; Zhou and Stephens, 2012). However, since LMM is not designed to analyze binary traits, it can have inflated type I error rates, especially in the presence of unbalanced case-control ratios. Recently, Chen, H. *et al*. have proposed to use logistic mixed models and developed a score test called the generalized mixed model association test (GMMAT)(Chen, et al., 2016). GMMAT assumes that score test statistics asymptotically follow a Gaussian distribution to estimate asymptotic p-values. Although GMMAT test statistics are more robust than the LMM based approaches, it can also suffer type I error rate inflation when case-control ratios are unbalanced, because unbalanced case-control ratios invalidate asymptotic assumptions of logistic regression. In addition, since GMMAT requires O(MN^2) computation and O(N^2) memory space, where M is the number of genetic variants to be tested and N is the number of individuals, it cannot handle data with hundreds of thousands of samples.

Here, we propose a novel method to allow for analysis of very large samples, for binary traits with unbalanced case-control ratios, which also infers and accounts for sample relatedness. Our method, Scalable and Accurate Implementation of GEneralized mixed model (SAIGE), uses the saddlepoint approximation (SPA)(Daniels, 1954; Dey, et al., 2017; Kuonen, 1999) to calibrate unbalanced case-control ratios in score tests based on logistic mixed models. Since SPA uses all the cumulants, and hence all the moments, it is more accurate than using the Gaussian distribution, which uses only the first two moments. Similar to BOLT-LMM(Loh, et al., 2015), the large sample size method for linear mixed-models, our method utilizes state-of-art optimization strategies, such as the preconditioned conjugate gradient (PCG) approach(Hestenes, 1952; Kaasschieter, 1988). for solving linear systems for large cohorts without requiring a pre-computed genetic relationship matrix (GRM). The overall computation cost of this proposed method is O(MN), which is substantially lower than the computation cost of GMMAT(Chen, et al., 2016) and many popular LMM methods, such as GEMMA(Zhou and Stephens, 2012). In addition, we reduce the memory use by compactly storing raw genotypes instead of calculating and storing the GRM.

We have demonstrated that SAIGE controls for the inflated type I error rates for binary traits with unbalanced case-control ratios in related samples through simulation and the UK Biobank data of 408,961 white British samples(Bycroft, et al., 2017; Sudlow, et al., 2015). By evaluating its computation performance, we demonstrate the feasibility of SAIGE for large-scale PheWAS.

4.3 Results

4.3.1 Overview of methods

The SAIGE method contains two main steps: 1. Fitting the null logistic mixed model to estimate variance component and other model parameters. 2. Testing for association between each genetic variant and phenotypes by applying SPA to the score test statistics. Step 1 iteratively estimates the model parameters using the computational efficient average information restricted maximum likelihood (AI-REML) algorithm(Gilmour, et al., 1995), which is also used in GMMAT(Chen, et al., 2016). Several optimization strategies have been applied in step 1 to make fitting the null logistic mixed model practical for large data sets, such as the UK Biobank(Bycroft, et al., 2017; Sudlow, et al., 2015). First, the spectral decomposition has been replaced by the PCG to solve linear systems without inversing the $N \times N$ GRM(Hestenes, 1952) (as in BOLT-LMM(Loh, et al., 2015)). The PCG method iteratively finds solutions of the linear system in a computation and memory efficient way. Thus, instead of requiring a pre-computed GRM, which costs a significant amount of time to calculate when sample sizes are large, SAIGE uses the raw genotypes as input. The computation time is about O(M_1N) times the number of iterations for the conjugate gradient to converge, where M_1 is a number of variants to be used for constructing GRM. Second, to further reduce the memory usage during the model fitting, the raw genotypes are stored in a binary vector and elements of GRM are calculated when needed rather than being stored, so the memory usage is $M_1N/4$ bytes (as in BOLT-LMM(Loh, et al., 2015) and GenABEL(Aulchenko, et al., 2007). For example, for the UK Biobank data with M_1 = 93,511 and N = 408,961 (white British participants), the memory usage drops from 669 Gigabytes(Gb) for storing the GRM with float numbers to 9.56 Gb for the raw genotypes in a binary vector.

After fitting the null logistic mixed model, the estimate of the random effects for each individual is obtained. The ratio of the variances of the score statistics with and without incorporating the variance components for the random effects is calculated using a subset of randomly selected genetic variants, similar to BOLT-LMM(Loh, et al., 2015) and GRAMMAR-Gamma(Svishcheva, et al., 2012). This ratio has been previously suggested to be constant for score tests based on LMMs(Svishcheva, et al., 2012). We have shown that the ratio is also approximately constant for all genetic variants with MAC ≥ 20 in the scenario of the logistic mixed models through analytic derivation and simulations (Supplementary Notes and Figure S4-1**).**

In step 2, for each variant, the variance ratio is used to calibrate the score statistic variance that does not incorporate variance components for random effects. Since GRM is no longer needed for this step, the computation time to obtain the score statistic for each variant is O(N). SAIGE next approximates the score test statistics using the SPA to obtain more accurate p-values than the normal distribution. A faster version of the SPA test, similar to the fastSPA method in the SPAtest R package that we recently developed(Dey, et al., 2017), , is used to further improve the computation time, which exploits the sparsity in low frequency or rare variants to reduce the computation cost.

4.3.2 Computation and memory cost

The key features of SAIGE compared to other existing methods are presented in Table 4-1, showing that SAIGE is the only mixed-model association method that is able to account for the unbalanced case-control ratios while remaining computationally practical for large data sets. To further evaluate the computational performance of SAIGE, we randomly sampled subsets from the 408,458 white British UK Biobank participants who are defined as either coronary artery disease (CAD) cases (31,355) or controls (377,103) based on the PheWAS Code 411(Bycroft, et al., 2017; Denny, et al., 2013; Sudlow, et al., 2015) followed by benchmarking association tests using SAIGE and other existing methods on 200,000 genetic markers randomly selected out of the 71 million with imputation info ≥ 0.3. The non-genetic covariates sex, birth year, and principal components 1 to 4 were adjusted in all tests. The log10 of the memory usage and projected computation time for testing the full set of 71 million genetic variants are plotted against the sample size as shown in **Error! Reference source not found.** and Table S4-1. Although SAIGE and BOLT-LMM have the same order of computational complexity (Table 4-1), SAIGE was slower than BOLT-LMM across all sample sizes (ex. 517 vs 360 CPU hours when *N*=408,458). This is due to the fact that fitting logistic mixed model requires more iterative steps than linear mixed model, and applying SPA requires additional computation. SAIGE requires slightly less memory than BOLT-LMM (10 to 11 Gb when *N*=408,458) and the low memory usage makes both methods feasible for the large data set. In contrast, GMMAT and GEMMA requires substantially more computation time and memory usage. For example, when *N*=400,000, projected memory usages of both GMMAT and GEMMA are more than 600 Gb. The actual computation time and memory usage of association tests for the full UK Biobank data for CAD are given in Table 4-1. SAIGE required 517 CPU hours and 10.3 Gb memory to analyze 71 million variants that have imputation info ≥ 0.3 for 408,458 samples, which indicates that the analysis will be done in ~26 hours with 20 CPU cores.

4.3.3 Association analysis of binary traits in UK Biobank data

We applied SAIGE to several randomly selected binary traits defined by the PheWAS Codes (PheCode) of UK Biobank(Bycroft, et al., 2017; Denny, et al., 2013; Sudlow, et al., 2015) and compared the association results with those obtained from the method based on linear mixed models, BOLT-LMM(Loh, et al., 2015) , and SAIGE without the saddlepoint approximation (SAIGE-NoSPA), which is asymptotically equivalent to GMMAT(Chen, et al., 2016). Due to computation and memory cost, the current GMMAT method cannot analyze the UK Biobank data. We restrict our analysis to markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC)(McCarthy, et al., 2016) panel due to quality control issues of non-HRC markers reported by the UK BioBank. Approximately 28 million markers with minor allele counts (MAC) ≥ 20 and imputation info score > 0.3 were used in the analysis. Among 408,961 white British participants in the UK Biobank, 132,179 have at least one up to the third degree relative among the genotyped individuals(Bycroft, et al., 2017; Sudlow, et al., 2015) . We used 93,511 high quality genotyped variants to construct the GRM. In the UK Biobank data, most binary phenotypes based on PheCodes (1,431 out of 1,688; 84.8%) have case-control ratio lower than 1:100 (Figure S4-3) and would likely demonstrate problematic inflation of association test statistics without SPA.

Association results of three exemplary binary traits that have various case-control ratios are plotted in Manhattan plots shown in Figure 4-1 and in the quantile-quantile (QQ) plots stratified by minor allele frequency (MAF) shown in Figure 4-2. The four binary traits are coronary artery disease (PheCode 411) with 31,355 cases and 377,103 controls (1:12), colorectal cancer (PheCode 153) with 4,562 cases and 382,756 controls (1:84), glaucoma (PheCode 365) with 4,462 cases and 397,761 controls (1:89), and thyroid cancer (PheCode 193) with 358 cases and 407,399 controls

(1:1138). In the Manhattan plots in Figure 4-1, each locus that contains any variant with p-value < $5x10^{-8}$ is highlighted as blue or green to indicate whether this locus has been reported by previous studies or not. Table S4-2 presents the number of all significant loci and those that have not been previously reported by each method for each trait and Table S4-3 lists all significant loci identified by SAIGE.

Both Manhattan and QQ plots show BOLT-LMM and SAIGE-NoSPA have greatly inflated type I error rates. The inflation problem is more severe as case-control ratios become more unbalanced and the MAF of the tested variants decreases. The genomic inflation factors (λ) at the 0.001, 0.01 p-value percentiles are shown for several MAF categories in Table S4-4. For the colorectal cancer GWAS which has case-control ratio 1:84, λ at the 0.001 p-value percentile is 1.68 and 1.71 for variants with MAF< 0.01 by SAIGE-NoSPA and BOLT-LMM, while λ is 0.99 by SAIGE. The inflation is even more severe for the test results by SAIGE-NoSPA and BOLT-LMM for the thyroid cancer, which has case-control ratio 1:1138, with the λ at the 0.001 p-value percentile around 4 to 5 for variants with MAF< 0.01 and all variants, respectively. With the unbalanced case-control ratio accounted for in SAIGE, the λ is again very close to 1.

We have generated summary statistics for all 1,403 PheCode-derived binary traits in 408,961 UK Biobank white British European-ancestry samples using SAIGE software and made them available in a public repository (see below for URL).

4.3.4 Simulation studies

We investigated the type I error control and power of two logistic mixed model approaches, SAIGE and GMMAT, and the linear mixed model method BOLT-LMM that computes mixed model association statistics under the infinitesimal and non-infinitesimal models through simulation studies. We followed the steps described on the Methods section to simulate genotypes for 1,000 families, each with 10 family members (N=10,000), based on the pedigree shown in Figure S4-4.

4.3.4.1 Type I error rates

The type I error rates for SAIGE, SAIGE-NoSPA, GMMAT, and BOLT-LMM have been evaluated based on the association tests performed on 10^9 simulated genetic variants. The variants were simulated using the same MAF spectrum of the UK Biobank HRC imputation data with case-control ratio 1:99, 1:9, and 1:1. Two different values of variance component parameter τ=1 and 2 were considered, which correspond to the liability scale heritability 0.23 and 0.38, respectively. The empirical type I error rates at the $\alpha = 5x10^{-4}$ and $\alpha = 5x10^{-8}$ are shown in the Table S4-5. Both SAIGE-NoSPA, GMMAT, and BOLT-LMM have greatly inflated type I error rates when the case-control ratios are moderately or extremely unbalanced and slightly deflated type I error rates when the case-control ratios are balanced. This is expected as previous studies have suggested inflation of the score tests in the presence of the unbalanced case-control ratios and deflation in balanced studies(Dey, et al., 2017; Ma, et al., 2013). We also observed that GMMAT score test statistics do not follow the normal distribution when MAF is low and case-control is unbalanced (Figure S4-5). Unlike GMMAT and BOLT-LMM, SAIGE has no inflation when case-control ratios are unbalanced. SAIGE also has no deflation when the case-control ratios are balanced.

To further investigate the type I error rates by MAF and case-control ratios, we carried out additional simulations.

Figure S4-6 shows QQ plots of 1,000,000 rare variants (MAF = 0.005) with various case-control ratios (1:1, 1:9, and 1:99) and Figure S4-7 shows QQ plots of 1,000,000 variants with different

MAF (0.005, 0.01, 0.05, 0.1 and 0.3) when case-control ratio was 1:99. Consistent to what has been observed in the real data study, GMMAT and SAIGE-NoSPA is more inflated for less frequent variants with more unbalanced case-control ratios. In contrast, SAIGE has successfully corrected this problem.

To evaluate whether SAIGE can control type I error rates in the presence of population stratification, we have simulated two subpopulations with Fst 0.013, which corresponds to the average Fst between Finnish and non-Finnish Europeans[20]. We assumed that subpopulations have different disease prevalences (0.01 for subpopulation 1 and 0.02 for subpopulation 2, 0.1 for subpopulation 1 and 0.2 for subpopulation 2, and 0.5 for subpopulation 1 and 0.4 for subpopulation 2). Both subpopulations have 1,000 families, each with 10 family members based on the pedigree shown in Figure S4-4. Association tests were performed on 10 million simulated markers and the first four principle components were included as covariates (Figure S4-8). QQ plots (Figure S4-9) show that the test statistics were well calibrated regardless of the variance component parameter τ and prevalence. This simulation result demonstrates that SAIGE produces well-calibrated p-values in the presence of population stratification.

4.3.4.2 Power

Next we evaluated empirical power. Since power simulation requires re-estimating a variance component parameter for each variant to test, to reduce computational burden, we used SAIGE-NoSPA instead of the original GMMAT software. Due to the inflated type I error rates of BOLT-LMM and GMMAT (and SAIGE-NoSPA), for a fair comparison, we estimated power at the test-specific empirical α levels that yield type I error rate $\alpha = 5x10^{-8}$ (Table S4-6). Figure S4-10 shows the power curve by odds ratios for variants with MAF 0.05, 0.1 and 0.2. When the case-control ratio is balanced, the power of SAIGE, SAIGE-NoSPA and BOLT-LMM were nearly identical. For studies with moderately unbalanced case-control ratio (case:control=1:9), SAIGE has higher power than SAIGE-NoSPA and BOLT-LMM, which is due to very small empirical α for SAIGE-NoSPA and BOLT-LMM resulted from type I error inflation. The power gap is much larger when the case-control ratios are extremely unbalanced. Power results for $\tau=2$ yielded the same conclusion regarding the methods comparison (data not shown).

Overall simulation studies show that SAIGE can control type I error rates even when case-control ratios are extremely unbalanced and can be more powerful than GMMAT and BOLT-LMM. In contrast, GMMAT and BOLT-LMM suffer type I error inflation, and the inflation is especially severe with low MAF and unbalanced case-control ratios.

4.3.4.3 Code and data availability

SAIGE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/. The GWAS results for 1,403 binary phenotypes with the PheCodes(Denny, et al., 2013) constructed based on ICD codes in UK Biobank using SAIGE are currently available for public download at https://www.dropbox.com/sh/wuj4y8wsqjz78om/AAACfAJK54KtvnzSTAoaZTLma?dl=0

We also display the results for 397 binary phenotypes in the Michigan PheWeb http://pheweb.sph.umich.edu/UKBiobank, which consists of Manhattan plots, Q-Q plots, and regional association plots for each phenotype as well as the PheWAS plots for every genetic marker. We will populate the pheweb with results for all UK biobank phenotypes (> 1,400).

4.4 Discussion

In this paper, we have presented a method to perform the association tests for binary traits in large cohorts in the presence of sample relatedness, which provides accurate p-value estimates for even extremely unbalanced case-control settings (with a prevalence < 0.1%). The dramatic decrease of the genotyping cost over the last decade allows more and more large biobanks to genotype all of their participants followed by genome-wide PheWAS, in which GWASs are performed for all thousands of diseases/conditions characterized based on EHR and/or survey questionnaires to identify genetic risk factors across different phenotypes(Bush, et al., 2016; Cronin, et al., 2014; Denny, et al., 2013; Denny, et al., 2011; Dumitrescu, et al., 2015; Hall, et al., 2014; Hebbring, et al., 2015; Hebbring, et al., 2013; Liao, et al., 2013; Millard, et al., 2015; Moore, et al., 2015; Namjou, et al., 2014; Neuraz, et al., 2013; Pendergrass, et al., 2013; Ritchie, et al., 2013; Shameer, et al., 2014; Ye, et al., 2015). Several challenges exist for PheWAS studies by large cohorts. Statistically, inflated type I error rates caused by unbalanced case-control ratios and sample relatedness need to be corrected. Computationally, most of existing mixed model association methods are not feasible for large sample sizes. Our method, SAIGE, uses logistic mixed model to account for the sample relatedness and applies the saddle point approximation (SPA) to correct the inflation caused by the unbalanced case-control ratio in the score tests based on logistic mixed models.

SAIGE successfully corrects the inflation of type I error rates of low-frequency variants with binary traits that have unbalanced case-control ratios while also accounting for the relatedness among samples. Furthermore, our method uses several optimization strategies that are similar to those used by BOLT-LMM to improve its computational feasibility for large cohorts. For example, the preconditioned conjugate gradient algorithm is used to solve linear systems instead of the Cholesky decomposition method so that the time complexity for fitting the null logistic model is decreased from O(N^3) to approximately O($M_1N^{1.5}$), where M_1 is the number of pruned markers used for estimating the genetic relationship matrix and the N is the sample size. Compared to large N, M_1 is usually small. For instance, in the UK Biobank(Bycroft, et al., 2017; Sudlow, et al., 2015), M_1 = 93,511 and N = 408,961 (white British participants).

There are several limitations in SAIGE. First, the time for algorithm convergence may vary among phenotypes and study samples given different heritability levels and sample relatedness. Second, SAIGE has been observed to be slightly conservative when case-control ratios are extremely unbalanced (Table S4-5). Third, the accurate odds ratio estimation requires fitting the model under the alternative and is not computational efficient. Similar to several other mixed model methods(Kang, et al., 2010; Loh, et al., 2015; Svishcheva, et al., 2012) , SAIGE estimates odds ratios for genetic markers using the parameter estimates from the null model. Fourth, SAIGE estimates the genetic relationship matrix using genome-wide genetic markers instead of using the leave-one-chromosome-out (LOCO) scheme, which can avoid proximal contamination(Lippert, et al., 2011; Listgarten, et al., 2012; Loh, et al., 2015; Yang, et al., 2014) . Last, SAIGE assumes that the effect sizes of genetic markers are normally distributed, which follows an infinitesimal architecture. With this assumption, SAIGE may sacrifice power to detect genetic signals whose genetic architecture is non-infinitesimal. In future direction, we will incorporate the LOCO scheme, which is straightforward based on the current model and method, and model non-infinitesimal architecture as needed to improve power. In addition, we will extend the current single variant test to gene- or region-based multiple variant test to improve power for identifying disease susceptibility rare variants.

With the emergence of large-scale biobank, PheWAS will be an important tool to identify genetic components of complex traits. Here we describe a scalable and accurate method, SAIGE, for the analysis of binary phenotypes in genome-wide PheWAS. Currently, SAIGE is the only available approach to adjust for both case-control imbalance and family relatedness, which are commonly observed in PheWAS datasets. In addition, the optimization approaches used in SAIGE make it scalable for the current largest (UK Biobank) and future much larger datasets. Through simulation and real data analysis, we have demonstrated that our method can efficiently analyze a dataset with 400,000 samples and adjust for type I error rates even when the case-control ratios are extremely unbalanced. Our method will provide an accurate and scalable solution for large scale biobank data analysis and ultimately contribute to identify genetic mechanism of complex diseases.

4.5 Methods

4.5.1 Generalized linear mixed model for binary traits

In a case-control study with sample size N, we denote the status of the *ith* individual using $y_i = 1$ or 0 for being a case or a control. Let the $1\times (1+p)$ vector X_i represent p covariates including the intercept and G_i represent the allele counts $(0, 1$ or $2)$ for the variant to test. The logistic mixed model can be written as

$$logit(\mu_i) = X_i\alpha + G_i\beta + b_i$$

where $\mu_i = P(y_i = 1 \mid X_i, G_i, b_i)$ is the probability for the *ith* individual being a case given the covariates and genotypes as well as the random effect, which is denoted as b_i. The random effect b_i is assumed to be distributed as $N(0,\ \tau\ \psi)$, where ψ is an $N \times N$ genetic relationship matrix

(GRM) and τ is the additive genetic variance. The α is a $(1+p) \times 1$ coefficient vector of fixed effects and β is a coefficient of the genetic effect.

4.5.2 Estimate variance component and other model parameters (Step 1)

To fit the null model, $logit(\mu_{i0}) = X_i\alpha + b_i$, penalized quasi-likelihood (PQL) method(Breslow and Clayton, 1993) and the AI-REML algorithm(Gilmour, et al., 1995) are used to iteratively estimate $(\hat{\tau}, \hat{\alpha}, \hat{b})$. At iteration k, let $(\hat{\tau}^{(k)}, \hat{\alpha}^{(k)}, \hat{b}^{(k)})$ be estimated $(\hat{\tau}, \hat{\alpha}, \hat{b})$, $\hat{\mu}_i^{(k)}$ be the estimated mean of y_i, $\widehat{W}^{(k)} = diag[\hat{\mu}_i^{(k)}\left(1-\hat{\mu}_i^{(k)}\right)]$, and $\hat{\Sigma}^{(k)}$= $(\widehat{W}^{(k)})^{-1} + \hat{\tau}^{(k)}\psi$ be an $n \times n$ matrix of the variance of working vector $\tilde{y}_i = X_i\alpha^{(k)} + b_i^{(k)} + (y_i - \hat{\mu}_i^{(k)})/\left\{\hat{\mu}_i^{(k)}\left(1-\hat{\mu}_i^{(k)}\right)\right\}$. To obtain log quasi-likelihood and average information at each iteration, the current GMMAT approach calculates the inverse of $\hat{\Sigma}^{(k)}$. Since it is computationally too expensive for large N, we use the preconditioned conjugate gradient (PCG)(Hestenes, 1952; Kaasschieter, 1988) , which allows calculating quasi-likelihood and average information without calculating $(\hat{\Sigma}^{(k)})^{-1}$ (See Supplementary for details). PCG is a numerical method to find solutions of linear system. It is particularly useful when the system is very large. BOLT-LMM(Loh, et al., 2015) successfully used it to estimate variance component in linear mixed model.

A score test statistics for H_o: $\beta = 0$ is $T = G^T(Y-\hat{\mu}) = \tilde{G}^T(Y-\hat{\mu})$ where G and Y are $N \times 1$ genotype and phenotype vectors, respectively, and $\hat{\mu}$ is the estimated mean of Y under the null hypothesis, and $\tilde{G} = G - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G$ is the covariate adjusted genotype vector. The variance of T, Var(T) = $\tilde{G}^T P \tilde{G}$, where $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$. For each variant, given $\hat{P}$, calculation of Var(T) requires O(N^2) computation. In addition, since our approach does not calculate $\hat{\Sigma}^{-1}$, and hence $\hat{P}$, obtaining Var(T) requires applying PCG for each variant, which can be computationally very expensive. To reduce computation cost, we use the same approximation approach used in BOLT-LMM and GRAMMAR-GAMMAR(Svishcheva, et al., 2012), , in which we estimate a variance of *T* with assuming that true random effect b is given, and then calculate ratio between these two variance. Suppose Var(*T*)* = $\tilde{G}^T\widehat{W}\tilde{G}$, which is a variance estimate of T assuming $\hat{b}$ is given. Let *r* = Var(*T*)/Var(*T*)* ratio of these two different types of variance estimates. In Supplementary materials, we have shown that the ratio is approximately constant for all variants. Using this fact, we can estimate *r* using a relatively small number of variants. In all the numerical studies in this paper, we used 30 variants to estimate *r*.

4.5.3 Score test with SPA (Step 2)

Suppose $\hat{r}$ is the estimated ratio (i.e. r) in Step 1. Now the variance adjusted test statistics is

$$T_{adj} = \frac{\tilde{G}^T(Y-\hat{\mu})}{\sqrt{\hat{r}\tilde{G}^T\widehat{W}\tilde{G}}},$$

which has mean zero and variance 1 under the null hypothesis. The computation of T_{adj} requires O(*N*) computation. The traditional score tests assume that *T* (and hence T_{adj}) asymptotically follows a Gaussian distribution under H_o: $\beta = 0$, which is using only the first two moments (mean and variance). When the case-control ratios are unbalanced and variants have low MAC, the underlying distribution of T_{adj} can be substantially different from Gaussian distribution. To obtain accurate p-values, we use Saddlepoint approximation method (SPA)(Dey, et al., 2017; Imhof, 1961; Kuonen, 1999) , which approximates distribution using the entire cumulant generating function (CGF). A fast version of SPA (fastSPA)(Dey, et al., 2017) has recently been developed and applied to PheWAS, and provides accurate p-values even when case-control ratios are extremely unbalanced (ex. case:control=1:600).

To apply fastSPA to *T_{adj}* we need to obtain CGF of $\mathrm{T}_{\mathrm{adj}}$ first. To do this, we use the fact that given $\hat{b}$, *T_{adj}* is a weighted sum of independent Bernoulli random variables. The approximated cumulant generating function is

$$K(t;\ \hat{\pi}, c) = \sum_{i=1}^{N} \log\left(1 - \hat{\pi}_i + \hat{\pi}_i e^{ct\tilde{G}_i}\right) - ct\sum_{i=1}^{N} \tilde{G}_i\,\hat{\pi}_i$$

where the constant c=Var*(T)$^{-1/2}$. Let $K'(t)$ and $K''(t)$ are first and second derivatives of K with respect to t. To calculate the probability that $T_{adj} < q$, where q is an observed test statistic, we use the following formula(Imhof, 1961; Johnson & Kotz, 1970; Kuonen, 1999)

$$pr\left(T_{adj} <\ q\right) \simeq F(q) =\ \Phi\left\{w + \frac{1}{w}\log\left(\frac{v}{w}\right)\right\},$$

where $w = sign(\hat{\zeta})\left[2\left\{\hat{\zeta}q - K(\hat{\zeta})\right\}\right]^{\frac{1}{2}}$, $v = \hat{\zeta}\left\{K''\left(\hat{\zeta}\right)\right\}^{\frac{1}{2}}$ and $\hat{\zeta} =\ \hat{\zeta}(q)$ is the solution of the equation $K'\left(\hat{\zeta}\right) = q$. As fastSPA(Dey, et al., 2017), we exploit the sparsity of genotype vector when MAF of variants are low. In addition, since normal approximation works well when the test statistic is close to the mean, we use the normal distribution when the test statistic is within two standard deviation of the mean.

4.5.4 Data simulation

We carried out a series of simulations to evaluate and compare the performance of SAIGE to GMMAT. We randomly simulated a set of 1,000,000 base-pair “pseudo” sequences, in which variants are independent to each other. Alleles for each variant were randomly drawn from Binomial(n = 2, p = MAF). Then we performed the gene-dropping simulation(Abecasis, et al., 2001) using these sequences as founder haplotypes that were propagated through the pedigree of

10 family members shown in Figure S4-4. Binary phenotypes were generated from the following logistic mixed model

$$logit(\pi_{i0}) = \alpha_0 + b_i + \ X_1 + X_2 + G_i\beta$$

where G_i is the genotype value, β is the genetic log odds ratio, b_i is the random effect simulated from $N(0, \tau\ \psi)$ with $\tau = 1$. Two covariates, X_1 and X_2, were simulated from Bernoulli(0.5) and N(0,1), respectively. The intercept α_0 was determined by given prevalence (i.e. case-control ratios).

To evaluate the type I error rates at genome-wide $\alpha=5\times10^{-8}$, 10 million markers along with 100 sets of phenotypes with different random seeds for case-control ratios 1:99, 1:9, and 1:1 were simulated with $\beta = 0$. Given $\tau = 1$, the estimated heritability is 0.015, 0.092, and 0.17 for phenotypes with case-control ratios 1:99, 1:9, and 1:1, respectively(de Villemereuil, et al., 2016) . Association tests were performed on the 10 million genetic markers for each of the 100 sets of phenotypes using SAIGE, GMMAT, and BOLT-LMM, therefore in total 10^9 tests were performed. To have a realistic MAF spectrum, MAFs were randomly sampled from the MAF spectrum in UK Biobank data (Figure S4-11). Additional type I error simulations were carried out for five different MAFs (0.005, 0.01, 0.05, 0.1 and 0.3) to evaluate type I error rates by MAFs.

For the power simulation, phenotypes were generated under the alternative hypothesis $\beta \neq 0$. For each of the MAF 0.05 and 0.2, we simulated 1,000 datasets, and power was evaluated at test-specific empirical α, which yields nominal $\alpha=5\times10^{-8}$. The empirical α was estimated from the previous type I error simulations. As the same as type I error simulations, three different case-control ratios (1:99, 1:9, and 1:1) were considered.

Note that since we evaluated the empirical type I error rates and power based on genetic variants that were simulated independently, the LD Score regression(Bulik-Sullivan, et al., 2015) calibration and the leave-one-chromosome-out (LOCO) scheme were not applied in BOLT-LMM.

4.5.5 Phenotype definition in UK Biobank

We used a previously published scheme to defined disease-specific binary phenotypes by combining hospital ICD-9 codes into hierarchical PheCodes, each representing a more or less specific disease group(Denny, et al., 2013)

ICD-10 codes were mapped to PheCodes using a combination of available maps through the Unified Medical Language System(https://www.nlm.nih.gov/research/umls/) and other sources, string matching, and manual review. Study participants were labeled a PheCode if they had one or more of the PheCode-specific ICD codes. Cases were all study participants with the PheCode of interest and controls were all study participants without the PheCode of interest or any related PheCodes. Gender checks were performed, so PheCodes specific for one gender could not mistakenly be assigned to the other gender.

4.6 Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 24460. SL and RD were supported by NIH R01 HG008773. CJW was supported by NIH R35 HL135824. WZ was supported by the University of Michigan Rackham Predoctoral Fellowship. JBN was supported by the Danish Heart Foundation and the Lundbeck Foundation. JCD, AG, LB, and WQW was supported by NIH R01 LM010685 and U2C OD023196.

4.7 Supplementary Notes

4.7.1 Algorithm Details: Step 1. Fitting the logistic mixed model under the null hypothesis

4.7.1.1 Generalized linear mixed model and penalized quasi-likelihood

Details of fitting the null logistic mixed model and estimating the parameters for fixed effects and variance components are provided in this section. Note that although we use the same restricted log likelihood and average information matrix as in GMMAT(Chen, et al., 2016), we use a different approach to estimate parameters to make our method feasible for very large datasets. In particular, we use the preconditioned conjugate gradient method(Kaasschieter, 1988)to solve linear systems instead of obtaining an inverse of the covariance matrix of the phenotypes. For the derivation of the likelihood and information matrix, please refer the GMMAT paper (Chen, et al., 2016).

Logistic mixed model is a part of the larger generalized linear mixed model (GLMM) with the logistic link function for binary outcome. The model can be written as

$$logit(\mu_i) = X_i\alpha + G_i\beta + b_{i,}$$

where $\mu_i = P(y_i = 1 \mid X_i, G_i, b_i)$ is the probability for the *ith* individual being a case given the covariates X_i and genotypes G_i as well as the random effect b_i, assumed to be distributed as $N(0,$ $\tau\,\psi)$, where ψ is an $N \times N$ genetic relationship matrix (GRM)(Kang, et al., 2010) and τ is an additive genetic variance. The phenotype y_i is assumed to be conditionally independent given (X_i, G_i, b_i) and follows the binomial distribution with mean $E(y_i|\, b_i) = \pi_i$ and variance $Var(y_i|\, b) = \ \phi v(\pi_i)$, where $v(\pi_i) = \mu_i(1-\mu_i)$ is the variance function, and the dispersion parameter $\phi = 1$.

Under the null hypothesis that $H_0\colon \beta = 0$, to estimate (α, ϕ, τ), the log integrated quasi-likelihood function can be written as

$ql(\alpha, \beta = 0, \phi, \tau) = \log \int exp\{\sum_{i=1}^{n} ql_i(\alpha, \beta = 0 \mid b)\} \times (2\pi)^{-\frac{n}{2}} |\tau\, \psi|^{-\frac{1}{2}} \times$

$exp\left\{-\frac{1}{2} b^T (\tau\, \psi)^{-1} b\right\} db, \qquad (1)$

where $ql_i(\alpha, \beta = 0 \mid b) = \int_{y_i}^{\mu_i} \frac{a_i(y_i - \mu)}{\phi v(\mu)} d\mu$ is the quasi-likelihood for the *ith* individual given the random effect b . Let $\kappa(b) = \sum_{i=1}^{n} ql_i(\alpha, \beta = 0 \mid b) - \frac{1}{2} b^T (\tau\, \psi)^{-1}$. Approximation for the integral $\int exp\{\kappa(b)\}db$ can be obtained using Laplace's method with the first and second derivatives. Let $\tilde{b}$ denote the solution of $\kappa'(b) = 0$, which maximizes $\kappa(b)$, and Wdenote the weight matrix, which is a diagonal matrix with diagonal terms $\frac{1}{\phi v(\mu_i)[g'(\mu_i)]^2}$. Note that since logistic is a canonical link function, the diagonal element of W can be simplified as $v(\mu_i)$. Equation (1) can be written as

$ql(\alpha, \beta = 0, \phi, \tau) = \kappa(\tilde{b}) - \frac{1}{2} log|\tau\, \psi\, W + I| \qquad (2)$

4.7.1.2 Estimate parameters using AI-REML

Here we describe iterative steps to estimate (α, b, ϕ, τ). To obtain the estimates of the fixed effect coefficients and the random effects given (ϕ, τ), $(\hat{\alpha}(\phi, \tau), \hat{b}(\phi, \tau))$, that jointly maximize the $ql(\alpha, \beta = 0, \phi, \tau)$, we take the derivative of equation (2) with respect to α and b and get the solution for the derivatives to be zero. Assuming the weight matrix W varies slowly as a function of the conditional mean, the last term in the expression of $ql(\alpha, \beta = 0, \phi, \tau)$ in equation (2) can be ignored. Let $\Sigma = W^{-1} + \tau\psi$, $P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$ and $\tilde{Y}$ be a working vector with the ith element being $X_i \alpha + b_i + g'(\mu_i)(y_i - \mu_i)$, and then

$\hat{\alpha} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \tilde{Y} \qquad (3)$

$\hat{b} = \tau\, \psi \Sigma^{-1} (\tilde{Y} - X\hat{\alpha}) \quad (4)$

Given $\hat{\alpha}$ and $\hat{b}$ estimated,

$ql(\hat{\alpha}(\phi,\tau)\,,\beta=0,\phi,\tau) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\tilde{Y}^T P\tilde{Y}$ (5)

The restricted maximum likelihood (REML) version:

$$ql_R(\hat{\alpha}(\phi,\tau)\,,\beta=0,\phi,\tau) = c_R - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}\tilde{Y}^T P\tilde{Y} \quad (6)$$

To obtain the estimates of the variance components, (ϕ,τ) , that jointly maximize the $ql(\hat{\alpha}(\phi,\tau)\,,\beta=0,\phi,\tau)$, we take the derivative of equation (6) with respect to ϕ and τ:

$\frac{\partial ql_R(\hat{\alpha}(\phi,\tau)\,,\beta=0,\phi,\tau)}{\partial\phi} = \frac{1}{2\phi}\tilde{Y}^T PW^{-1}P\tilde{Y} \; - \frac{1}{2\phi}tr(PW^{-1})$ (7)

$\frac{\partial ql_R(\hat{\alpha}(\phi,\tau)\,,\beta=0,\phi,\tau)}{\partial\tau} = \frac{1}{2}(\tilde{Y}^T P\psi P\tilde{Y} \; - tr(P\psi))$ (8)

$\hat{\phi}$ and $\hat{\tau}$ are estimated by obtaining the solutions to make equations (7) and (8) equal to zero. Let θ represents the vector of variance component parameters. In this case, θ is a vector containing ϕ and τ. In the REML iterative process, the estimates for θ in the (i+1)th iteration is updated by $\theta^{(i+1)} = \theta^{(i)} + J(\theta^{(i)})^{-1}S(\theta^{(i)})$, where $S(\theta) = \frac{\partial ql_R(\theta)}{\partial\theta}$ as the equation (7) and (8) and $J(\theta) = -\frac{\partial S(\theta)}{\partial\theta} = -\frac{\partial^2 ql_R(\theta)}{\partial\theta^2}$.

The elements of the observed information matrix J(θ)(Gilmour, et al., 1995)are

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi^2} = -\frac{1}{2}tr(P\psi_0P\psi_0) + \tilde{Y}^T P\psi_0P\psi_0P\tilde{Y}$$

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi\,\partial\tau} = -\frac{1}{2}tr(P\psi_0P\psi) + \tilde{Y}^T P\psi_0P\psi P\tilde{Y}$$

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\tau^2} = -\frac{1}{2}tr(P\psi P\psi) + \tilde{Y}^T P\psi P\psi P\tilde{Y} \qquad (9)$$

The elements of the expected information matrix(Gilmour, et al., 1995) are

$$\mathrm{E}\left(-\frac{\partial \mathrm{ql}_{\mathrm{R}}(\widehat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi^2}\right)=\frac{1}{2}\mathrm{tr}(\mathrm{P}\psi_0\mathrm{P}\psi_0)$$

$$\mathrm{E}\left(-\frac{\partial \mathrm{ql}_{\mathrm{R}}(\widehat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi\,\partial\tau}\right)=\frac{1}{2}\mathrm{tr}(\mathrm{P}\psi_0\mathrm{P}\psi)$$

$$\mathrm{E}\left(-\frac{\partial \mathrm{ql}_{\mathrm{R}}(\widehat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\tau^2}\right)=\frac{1}{2}\mathrm{tr}(\mathrm{P}\psi\mathrm{P}\psi) \qquad (10)$$

To avoid the trace evaluation in (9), which has high computational cost, an average information matrix AI is then defined as the average of the observed information in (9) and the expected information in (10) in place of the $\mathrm{J}(\theta)$ matrix to estimate $\widehat{\phi}$ and $\hat{\tau}$ iteratively(Chen, et al., 2016; Gilmour, et al., 1995; Yang, et al., 2011).

$$\mathrm{AI}_{\phi\phi}=\frac{1}{2}\,\widetilde{\mathrm{Y}}^{\mathrm{T}}\mathrm{P}\psi_0\mathrm{P}\psi_0\mathrm{P}\widetilde{\mathrm{Y}}$$

$$\mathrm{AI}_{\phi\tau}=\mathrm{AI}_{\tau\phi}=\frac{1}{2}\,\widetilde{\mathrm{Y}}^{\mathrm{T}}\mathrm{P}\psi_0\mathrm{P}\psi\mathrm{P}\widetilde{\mathrm{Y}}$$

$$\mathrm{AI}_{\tau\tau}=\frac{1}{2}\,\widetilde{\mathrm{Y}}^{\mathrm{T}}\mathrm{P}\psi\mathrm{P}\psi\mathrm{P}\widetilde{\mathrm{Y}} \qquad (11)$$

Note that for the logistic mixed model, $\phi=1$, so we do not need to obtain (7) and the first two equations in (9-11) that contain derivatives with ϕ .

4.7.1.3 Approaches to reduce computation and memory cost.

Preconditioned Conjugate Gradient (PCG): To obtain equations (3)-(8), we need to compute expression forms containing a product of Σ^{-1} and a vector or a matrix, such as $\Sigma^{-1}X$, which is very challenging for large cohorts. Computing the $N\times N$ empirical genetic relationship matrix (GRM) $\psi = \frac{G_c^T G_c}{M_1}$ costs O(M_1N^2), where G_c is an $M_1\times N$ matrix with genotypes for M_1 genetic markers of N individuals that are normalized with the means and standard deviations of raw genotypes. Moreover, the Cholesky decomposition used by GMMAT[1] to invert Σ takes $O(N^3)$ computation and very large memory space, which are not practical for studies with large sample sizes ($N > 20{,}000$).

Similar to BOLT-LMM (Loh, et al., 2015), we use two strategies to reduce the computation and memory cost. First, instead of requiring the pre-computed GRM ψ as an input, we store genotypes for computing GRM in a binary vector and calculate elements of Σ as needed, which reduces the memory usage from $4N(N+1)$ bytes, given double precision floating number is used to store ψ, to $\frac{NM_1}{4}$ bytes. For instance, with N =408,961 white British participants and M_1 = 93,511 markers, the memory usage drops from 669 Gb to 9.56 Gb with this strategy. Second, the conjugate gradient method is used to calculate the product of Σ^{-1} and a vector by iteratively solving the linear system $Ax = u$, where $A = \Sigma$ and u is a known vector, such as any column vector in X matrix. The number of iterations required for convergence of the conjugate gradient algorithm is proportional to $\sqrt{\kappa(A)}$, where $\kappa(A)$ is the condition number for A. To make the convergence faster, a preconditioner matrix Q is used so that $\hat{A}$=$Q^{-1}A$ and $\kappa(\hat{A}) < \kappa(A)$. Here, Q is an $N \times N$ diagonal matrix with the diagonal elements of Σ and the calculation of Q requires O(NM_1).

The numerical accuracy of the PCG method has been evaluated based on the Euclidean distance for the vector Σ^{-1}y computed by PCG and by calculating Σ^{-1} for the simulated data sets as described in the Data Simulation section. With the tolerance 1x10^{-5} for PCG to converge, the average Euclidean distances for 100 simulated data sets with case-control ratio 1:99, 1:9 and 1:1 are 2.46x10^{-11}, 7.70x10^{-10}, and 1.53x10^{-9}, respectively, suggesting the PCG method is highly accurate. The average numbers of PCG interactions to convergence are 4, 6 and 7 for case-control ratio of 1:99, 1:9, and 1:1, respectively. The average iterations for PCG to converge for the 1,283 non-sex specific binary phenotypes in the UK Biobank have been plotted in Figure S4-12. There was no phenotype with an average number of iterations larger than 10, indicating PCG converges reasonably fast in UK Biobank data analysis.

Randomized trace estimator for $tr(PW^{-1})$ and $tr(P\psi)$: The computation of (7) and (8) requires the traces of matrices PW^{-1} and $P\psi$. For this, we use Hutchinson's randomized trace estimator[6,7]. The trace of a matrix B, such as PW^{-1} and $P\psi$, is estimated by $\frac{1}{R}\sum_{i=1}^{R} z_i^T B z_i$, where z_i's are R independent random vectors whose entries are i.i.d Rademacher random variables ($\mathrm{P}(\mathrm{z_i} = \pm 1) =$ 0.5). A vector z_i with size N is randomly drawn from the Rademacher distribution, followed by the calculation for $z_i^T B z_i$. This procedure is repeated for R times and the average of the results for $z_i^T B z_i$ is the estimate for the trace of the B matrix. The by default value for R is set to be 30.

The numerical stability and convergence of the randomized trace estimator has been evaluated using data sets that were simulated as described in the Data Simulation section. During the process of fitting the null generalized logistic model iteratively, the trace of the matrix $P\psi$ was estimated using different numbers of independent random vectors (R = 10, 20, 30, 40 and 50). The estimated traces were plotted against the true traces that were computed as the sum of the elements on the main diagonal of matrix $P\psi$ in the Figure S4-13. As the number of random vectors that were used for trace estimation increases, the estimator is more stable and more consistent to the true value. Given that the trace is estimated as the average of $z_i^T B z_i$, i=1, …., R, the coefficient of variation (CV), which is defined as the ratio of standard error to the mean (i.e. SE/Mean) and measures relative variability, is used to determine whether R independent random vectors provide stable trace estimation. When R=30, in most simulated datasets, CV < 0.0025, which indicates that trace can be accurately estimated using 30 independent random vectors for the simulated data sets. Therefore, the default number of random vectors to use (R) in SAIGE is set to be 30. But it is possible that R=30 is not enough to stably calculate the trace in some datasets. In this case, R should be increased. A function to adaptively increase R when the CV is larger than a certain threshold has been implemented in the SAIGE R package.

Parallel computation for the vector multiplication: The most time-consuming step of the proposed algorithm is performing PCG, which involves computing a product of the GRM ψ and a vector x, i.e. $\psi x = G_c^T G_c$x. We use parallel computing techniques to speed up this procedure. In particular, we use Intel Threading Building Block (TBB) implemented in RcppParallel package[8] for the multi-threading computation. Our approach utilized nearly all CPU cores allocated. For example, the CPU usages on average were 14.6 when 16 CPU cores were allocated.

A low-rank GRM to correct for sample relatedness: Since the computation and memory cost of step 1 in SAIGE is linear to the number of markers (M_1) used to construct the Kinship matrix, the computation and memory cost can be reduced using a subset of markers, instead of using all available markers. In the UK Biobank data analysis, for example, 93,511 independent, high quality genotyped variants were used for the step 1 (M_1 = 93,511), which is the same set of markers used by the UK Biobank data group to estimate the kinship coefficients between samples(Bycroft, et al., 2017). This low-rank GRM approach was first proposed by Lippert C, *et al.*(Lippert, et al., 2011) and has been shown to provide similar p-values to using the more complete set of genetic markers to construct GRM(Lippert, et al., 2011). Later, Yang *et al.* suggested that using a few thousand genetic markers to construct GRM would reduce the ability to correct for sample relatedness(Yang, et al., 2014). Therefore the marker selection for step 1 should be based on careful consideration for the trade-off between computation cost and performance of adjusting for sample relatedness. In Supplementary Section 2, a sensitivity analysis has been reported when increasing M_1 to be 340,447. Using more markers for the step 1 produced generally similar p-values but with lambdas closer to 1.

4.7.2 Algorithm Details: Step 2. Single variant score tests with SPA

4.7.2.1 Score tests based on logistic mixed model

Given the estimates from step 1 for fixed effect coefficients $\hat{\alpha}$, random effects $\hat{b}$, and the variance component parameters $\hat{\phi}$ and $\hat{\tau}$ under the null hypothesis H_o: $\beta = 0$, the score test can be constructed for each genetic marker to be tested. Suppose G is the $\mathrm{N}\times 1$ genotype vector, $\hat{\mu}$ is estimate for $P(Y = 1 \mid X, \hat{b})$, are the probabilities for study individuals being a case given the covariates X and the estimated random effect $\hat{b}$ from step 1, $\widehat{W}$ is a diagonal vector with diagonal elements $\hat{\mu}(1 - \hat{\mu})$, and $\tilde{G} = G - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G$ is the covariate adjusted genotype vector with covariate effects projected out from the raw genotypes[9]. Suppose $\hat{\Sigma} = \hat{W}^{-1} + \hat{\tau}\psi$ and $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$, and then $\hat{P}G = \hat{P}\tilde{G}$. The score test statistics can be written as

$$T = G^T(Y - \hat{\mu}) = G^T\hat{P}\tilde{Y} = \tilde{G}^T\hat{P}\tilde{Y} = \tilde{G}^T(Y - \hat{\mu}),$$

where $\tilde{Y}$ is the working vector previously defined. The variance of T, Var(T) = $G^T\hat{P}G = \tilde{G}^T\hat{P}\tilde{G}$.

4.7.2.2 Estimation of Var(T)

Calculating $\hat{P}\tilde{G}$ is required for the estimation of Var(T), which is computationally expensive. To avoid to calculate $\hat{P}\tilde{G}$ to all the variants, we use similar approximation approaches used in BOTL-LMM(Loh, et al., 2015) and GRAMMAR-GAMMAR(Svishcheva, et al., 2012) in which we obtain the ratio between $\text{Var}(T)$ and $\text{Var}(T)^* = \tilde{G}^TW\tilde{G}$ using a small number of variants, and estimate variant as $r\text{Var}(T)^*$, where $r = \text{Var}(T)\ /\ \text{Var}(T)^*$. Note that $\text{Var}(T)^*$ is a variance estimator without accounting the fact that the random effect *b* is estimated from data, and the calculation of $\text{Var}(T)^*$ only requires O(N) computation.

Here we show that the ratio *r* is approximately constant across all variants. For this, we assume that $\frac{w_i}{\sum_{j=1}^{N} w_j} = o(1)$, for all i=1, …, N, where w_i is the i^{th} element of W. Note that this assumption can only be violated when the covariates are extremely sparse, which rarely happens in real data. First, $Var(T)$ can be written as

$$Var(T) = \tilde{G}^TP\tilde{G} = \tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} - \tilde{G}^T\hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} \qquad (3)$$

Suppose $\tilde{G}_i$ is the i*th* element of $\tilde{G}$. Since $\tilde{G}$ is adjusted by covariates including the intercept, $\tilde{G}_i$ can be treated as a mean zero random variable uncorrelated with the covariates, and hence $N^{-1/2}X^T\hat{\Sigma}^{-1}\tilde{G}$ asymptotically have mean zero and variance $N^{-1}X^T\hat{\Sigma}^{-1}Var\left(\tilde{G}\right)\hat{\Sigma}^{-1}X = O(1)$. By

Chebyshev's inequality $N^{-1/2}X^T\hat{\Sigma}^{-1}\tilde{G} = O_P(1)$. Since $\left(X^T\hat{\Sigma}^{-1}X\right)^{-1} = O(N^{-1})$, the second term in (3) is $\tilde{G}^T\hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} = O_p(1)$. The first term in (3) is $\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} = O_p(N)$, so (3) can be approximated by $\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G}$. Let $\overline{w}$ be the mean of the diagonal element of $\widehat{W}^{-1}$ and $\xi = \widehat{W}^{-1} - \overline{w}I$. And then

$$\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} \approx \tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} - G^T(\overline{w}I + \hat{\tau}\psi)^{-1}\xi(\overline{w}I + \hat{\tau}\psi)^{-1}G \quad (4)$$

With the assumption $\frac{w_i}{\sum_{j=1}^{N} w_j} = o(1)$, $\tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\xi(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = \sum\xi_i d_i$, where ξ_i is the i*th* diagonal element of ξ and d_i is the square of the i*th* element of $(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G}$. Since the mean of ξ_i is zero, and ξ_i and d_i are uncorrelated, $\sum\xi_i d_i = o_p(N)$. Combining a fact that $\tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = O_P(N)$, $\frac{G^T(\overline{w}I+\hat{\tau}\psi)^{-1}\xi(\overline{w}I+\hat{\tau}\psi)^{-1}G}{\tilde{G}^T(\overline{w}I+\hat{\tau}\psi)^{-1}\tilde{G}} = o_p(1)$, therefore (4) can be approximated by the first term, in which

$$(4) \approx \tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = \tilde{G}^T\psi^{-\frac{1}{2}}\psi^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\psi)^{-1}\psi^{\frac{1}{2}}\psi^{-\frac{1}{2}}\tilde{G} = a^T U\Lambda^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\Lambda)^{-1}\Lambda^{\frac{1}{2}}Ua \quad (5)$$

where U and Λ are eigenvector and eigenvalue matrices of ψ, and $a = \psi^{-\frac{1}{2}}\tilde{G}$. Since correlation matrix of a is an identity matrix, asymptotically, (4) is closely approximated by the trace of $cU\Lambda^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\Lambda)^{-1}\Lambda^{\frac{1}{2}}U$, which is $c\sum_{i=1}^{n}\lambda_i/(\overline{w} + \hat{\tau}\lambda_i)$, where c=*MAF*(1-*MAF*). As the same way, $\mathrm{Var}(T)^* = \tilde{G}^T\widehat{W}\tilde{G} \approx c\sum_{i=1}^{n}\lambda_i/\overline{w}$. And hence the ratio is

$$r = \frac{\mathrm{Var}(T)}{\mathrm{Var}(T)^*} \approx \frac{\sum_{i=1}^{n}\frac{\lambda_i}{\overline{w} + \hat{\tau}\lambda_i}}{\sum_{i=1}^{n}\frac{\lambda_i}{\overline{w}}}$$

which is constant across all variants. The variance adjusted score test statistic is

$$T_{adj} = \left(\hat{r}\tilde{G}^T\widehat{W}\tilde{G}\right)^{-1/2}\tilde{G}^T(Y - \hat{\mu})$$

where $\hat{r}$ is the estimated *r*, which is estimated from 30 randomly selected genetic markers. Under the null hypothesis of no association, T_{adj} has mean zero and variance one. Figure S4-1 shows the ratio *r* by minor allele counts (MAC) from 1000 simulated markers. The ratio was nearly identical for markers with MAC > 20 and then variation was increased for extremely rare variants. This figure provides empirical evidence that the equal ratio assumption holds.

In analysis of simulated and real data, 30 randomly selected genetic markers with MAC > 20 were used to estimate $\hat{r}$. To evaluate the numerical stability of the $\hat{r}$ estimation, the coefficient of variance (CV) of $\hat{r}$ using simulated datasets was used. In most simulated datasets, the CV for $\hat{r}$ was smaller than 0.001 (Figure S4-14) with 30 randomly selected markers, indicating that $\hat{r}$ can be accurately estimated using 30 markers. As a sensitivity analysis, $\hat{r}$ has been calculated based on 500 randomly selected markers, and the estimated $\hat{r}$ were nearly identical (Figure S4-14). But it is also possible that using 30 markers is not enough to stably calculate $\hat{r}$ in some datasets. In this case, the number of markers for $\hat{r}$ should be increased. As the same as the random trace estimation (Section 1.1.3), a function is included in the SAIGE package, in which the number of markers for $\hat{r}$ is automatically increased if CV is larger than a given threshold (current default=0.001).

4.7.2.3 P-value calculation using SPA

The traditional score test, such as GMMAT, used the fact that the score test statistic asymptotically follows a normal distribution under the null hypothesis of no association. When the case-control ratios are unbalanced and MAC is small, this asymptotic result does not hold and type I error rates can be inflated. To obtain more accurate p-value, we use a fast-version of SPA (fastSPA)(Dey, et al., 2017), which we have previously developed for logistic regression model. For this, we utilize the fact that phenotype Y_i independently follows Bernoulli distribution given π_i, and T_{adj} is a weighted sum of independent Bernoulli random variable. The approximated cumulant generating function (CGF) of T_{adj} is

$$K(t;\ \hat{\pi}, c) = \sum_{i=1}^{N} \log\left(1 - \hat{\pi}_i + \hat{\pi}_i e^{ct\tilde{G}_i}\right) - ct \sum_{i=1}^{N} \tilde{G}_i \hat{\pi}_i$$

where the constant c=Var*(T)$^{-1/2}$, which provide K'(0)=0 and K''(0) = 1, where K' and K'' are first and second derivate of K with respect to t. Note that since K uses $\hat{\pi}$, which is estimated from data, it is an approximation of the true CGF. Now we use the saddle point method to estimate the p-value. To calculate the probability that $T_{adj} < q$, where q is an observed test statistic, we use the following formula(Kuonen, 1999) (Imhof, 1961) (Johnson & Kotz, 1970).

$$pr\left(T_{adj} < q\right) \simeq F(q) = \Phi\left\{w + \frac{1}{w}\log\left(\frac{v}{w}\right)\right\}$$

,where $w = sign(\hat{\zeta})\left[2\{\hat{\zeta}q - K(\hat{\zeta})\}\right]^{\frac{1}{2}}$, $v = \hat{\zeta}\{K''(\hat{\zeta})\}^{\frac{1}{2}}$ and $\hat{\zeta} = \hat{\zeta}(q)$ is the solution of the equation $K'(\hat{\zeta}) = q$. As the fastSPA(Dey, et al., 2017), we exploit the sparsity of genotype vector when MAF of variants are low. In addition, since normal approximation performs well when the test statistic is close to the mean, we use normal distribution when the test statistic is within two standard deviations of the mean.

4.7.2.4 Effect size estimation

To rapidly estimate the effect size $\hat{\beta}$, which equals to the natural logarithm of the odds ratio, we use the variance component estimate under the null hypothesis. Note that a similar approach has been used in EMMAX(Kang, et al., 2008) and GRAMMAR-Gamma(Svishcheva, et al., 2012). Our $\hat{\beta}$ estimate is

$$\hat{\beta} = \left(\tilde{G}^T \hat{P} \tilde{G}\right)^{-1} \tilde{G}^T \hat{P} \tilde{Y}$$

Since $T = \tilde{G}^T \hat{P} \tilde{Y}$ and $\text{Var}(T) = \tilde{G}^T \hat{P} \tilde{G}$, $\hat{\beta}$ can be written as $T/\text{Var}(T)$. In the section 1.2.2, we have shown that $\text{Var}(T) = \hat{r}\text{Var}(T)^* = \hat{r}\tilde{G}^T \widehat{W} \tilde{G}$. Therefore, $\hat{\beta}$ can estimated using $T, \text{Var}(T)^*$, and $\hat{r}$, which have already been calculated for association p-value estimation. To estimate the standard error and confidence interval, we use p-values. The standard error of $\hat{\beta}$, $SE(\hat{\beta}) = |\hat{\beta}/z|$, where z-score corresponds to the association p-value/2.

4.7.2.5 Leave-one-chromosome-out

To avoid contamination from correlated markers(Lippert, et al., 2011), we implemented an option to apply the leave-one-chromosome-out (LOCO) scheme in SAIGE. In step 1, given the variance component parameter $\hat{\boldsymbol{\tau}}$ that was estimated using GRM constructed with genome-wide markers, the random and fixed effects were estimated for each chromosome using a GRM constructed with genetic markers excluding that chromosome. In the following step 2 for association tests, the estimates from step 1 using all other chromosomes are then used for testing genetic markers on that chromosome. We evaluated this approach by comparing p-values with and without the LOCO scheme.

Figure S4-15 shows the scatter plots for the p-values of the 28 million genotyped and imputed markers for the four randomly selected phenotypes in the UK Biobank data. We found that the p-values estimated with and without LOCO are highly correlated.

4.7.3 Additional simulation and real data analysis results

4.7.3.1 Simulation studies with different τ values and heritability estimation

In SAIGE, penalized quasi-likelihood (PQL), which provides easy implementation and fast computation, is used to estimate the variance component parameter $\hat{\tau}$. Although PQL is the mostly widely used method in Generalized Linear Mixed Model and also used by the recently developed GMMAT method(Chen, et al., 2016), it is known to produce biased estimate of the variance component ($\hat{\tau}$)(Breslow, 2004; Breslow and Clayton, 1993; Capanu, et al., 2013), and therefore, the heritability estimates. This may be due to the fact that PQL approximates true-likelihood using Laplace method, and hence after the approximation, $\boldsymbol{\tau}$ in true likelihood is no longer the same as $\boldsymbol{\tau}$ in the approximated model.

Table S4-7 shows $\hat{\boldsymbol{\tau}}$ estimated by PQL (as in SAIGE) for simulated data with four different τ values, 0.5, 1, 2, and 3, corresponding to h_{latent}^2 =0.13, 0.23, 0.38, and 0.48, respectively, where h_{latent}^2 is a liability scale heritability. The h_{latent}^2 was obtained using the fact that the logistic regression can be described as a liability threshold model with standard logistic distribution, which has variance=$\pi^2/3$ = 3.23. Therefore the variance component parameter τ can be converted to the heritability on latent scale as

$$h_{latent}^2 = \frac{\tau}{\pi^2/3 + \tau}$$

Using the relationship between τ and h_{latent}^2, τ can be estimated from the liability scale heritability estimates from other methods, such as phenotype correlation–genotype correlation (PCGC) regression method(Golan, et al., 2014). PCGC is a moment-based method and known to produce unbiased heritability estimation. We estimated $\hat{\boldsymbol{\tau}}$ as $\frac{\pi^2 h_{pcgc}^2}{3(1-h_{pcgc}^2)}$, where h_{pcgc}^2 is the latent scale heritability estimated by PCGC. Table S4-7 clearly suggests that $\hat{\boldsymbol{\tau}}$ from SAIGE is substantially biased. Therefore， $\hat{\boldsymbol{\tau}}$ estimated by SAIGE should not be used to interpret the heritability. $\hat{\boldsymbol{\tau}}$ from PCGC was more accurate than that from SAIGE; however, it was still biased especially when true τ was large.

To evaluate whether SAIGE can control type I errors in wide ranges of heritability, additional type I error simulations with four τ values (0.5, 1, 2, and 3) have been performed and the results are similar for different τ values (Figure S4-16). The results with $\tau = 1$ and 2 are shown in Table S4-7. To evaluate whether using more accurate τ estimate can have impact on type I error control, we also included approaches assuming 1) true τ is known (true- τ), and 2) estimating τ using PCGC regression (PCGC- τ). For both approaches, fixed and random effect terms were calculated from Equations (3) and (4) given τ. Note that since true τ is unknown in real data, the first approach (i.e true- τ) can be used in simulation study only. Figure S4-16 shows QQ plots when the variant MAF=0.01 and case control ratio=1:99. In all τ -values, the proposed PQL-based approach has very well calibrated QQ plots. Interestingly, both true- τ and PCGC- τ have deflated QQ plots, indicating that these approaches produce conservative results. As aforementioned, this may due to the fact that our score test statistics were derived from PQL not from original likelihood. We note that type I error simulations with different MAFs (0.3 and 0.05) and case control ratios (1:9 and 1:1) yielded nearly identical results (data not shown).

Overall these simulation studies clearly show that although PQL is biased for the heritability estimation, it works well for adjusting for sample-relatedness.

4.7.3.2 Simulation studies with Population stratification

To evaluate whether SAIGE can control type I error rates in the presence of population stratification, we have simulated two subpopulations with Fst 0.013, which corresponds the Fst between Finnish and non-Finnish Europeans(Nelis, et al., 2009), assuming that subpopulations have different disease prevalence. Each subpopulation has 1000 families, each with 10 family members based on the pedigree shown in Figure S4-4**.** 93,511 genetic markers were simulated with the overall minor allele frequency (MAF) following the MAF spectrum of the genotyped markers that were used for constructing the GRM in the UK Biobank data. Three different disease prevalences were considered for subpopulations 1 and 2 (0.01 and 0.02; 0.1 and 0.2; 0.5 and 0.4). Four different τ values are used to simulate the phenotypes: 0.5, 1, 2, and 3. Association tests were performed on 10 million markers including the first four principle components as covariates. The overall MAF of 10 million markers follows the same MAF spectrum of the imputed genetic markers in the UK Biobank data. The plots for the PCs were presented in the Figure S4-8, which shows that PC1 well separated two populations. QQ plots (Figure S4-9) were well calibrated regardless of τ and prevalence. This simulation results clearly demonstrate that our approach can produce well calibrated p-values in the presence of population stratification.

4.7.3.3 UK Biobank data analysis with different M_1

As a sensitivity analysis, we used 340,447 genotyped markers for step 1, which were obtained by using the following pruning parameters on directly genotyped markers: using windows of 500,000 base pairs (bp), a step-size of 50 markers, and pairwise $r^2 < 0.2$. We compared association p-values for four randomly selected phenotypes in the UK Biobank data when GRM was constructed using the 93,511 genotyped markers and the 340,447 genotyped markers, respectively. Scatter plots comparing p-values of the 28 million tested genetic markers are presented in Figure S4-17, suggesting highly correlated association p-values. We also note that when 340,447 markers were used for GRM , -log10 p-values were slightly lower than those of using 93,511 markers, especially for coronary artery disease (PheCode 411) (Figure S4-17) and the genomic inflation factors (λ) at the 0.001, 0.01 p-value percentiles slightly decreased (Table S4-8). Manhattan plots of these two approaches were largely similar (Figure S4-18), in which colorectal cancer (PheCode 153), glaucoma (PheCode 365), and thyroid cancer (PheCode 193) have the exactly same number of GWAS hits.

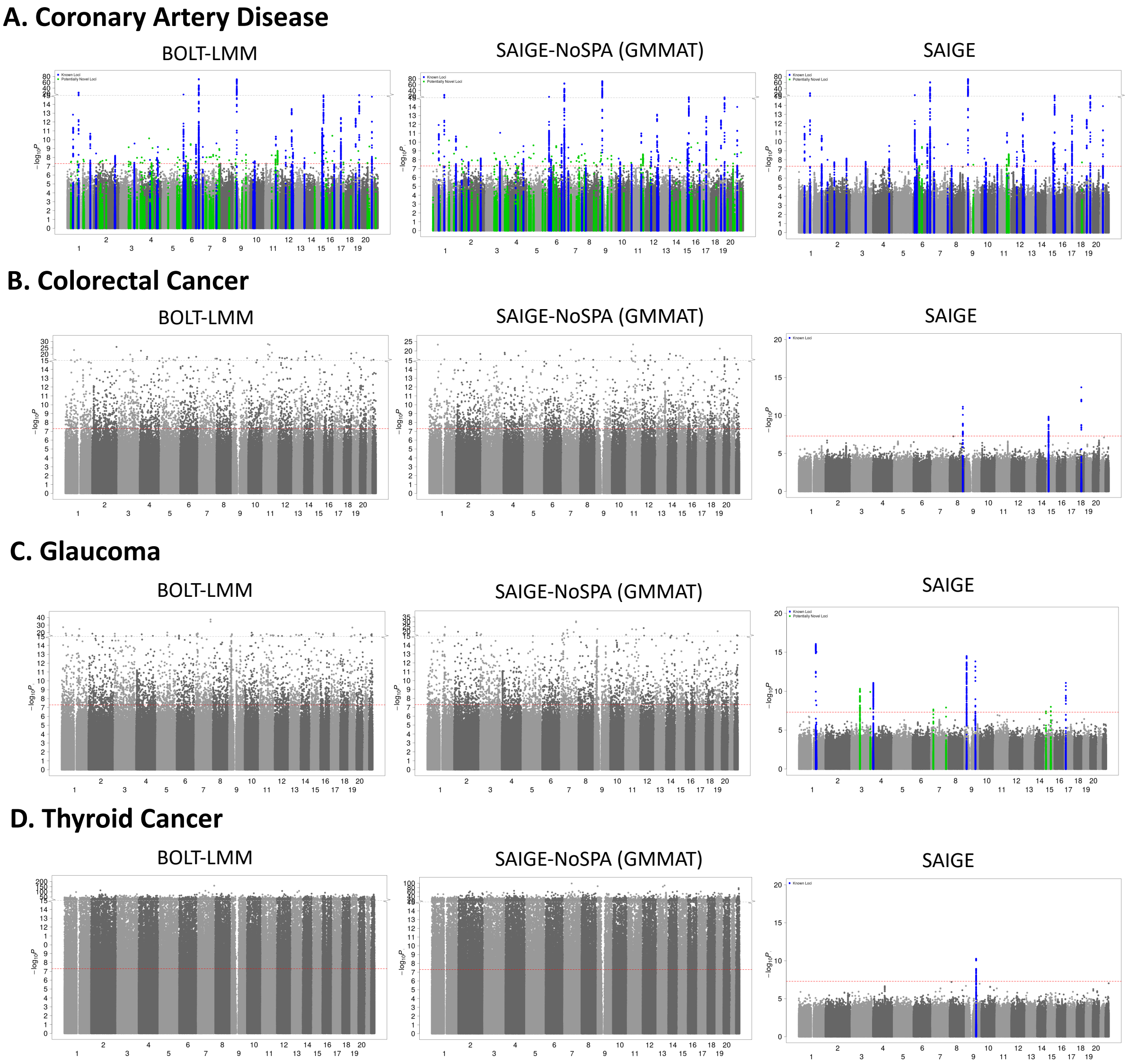
4.7.3.4 Additional rare variant associations in UK Biobank

Among SAIGE results for 1,283 non sex-specific binary phenotypes constructed based on the PheCodes in the UK Biobank data, there are total 1,609 genetic variants, including variants in the same locus, with minor allele frequency < 0.5% with SAIGE p-values < 5x10^{-8}. Examples include the *HBB* locus (rs11549407, MAF=0.027%, p-value=2.4x10^{-12}) associated with hereditary hemolytic anemias (http://pheweb.sph.umich.edu:5003/pheno/282), and two different rare variants associated with breast cancer: the *ZNRF3* locus (rs6223688, MAF=0.26%, p-value=1.8x10^{-23}) and the *TTC28* locus (rs62237617, MAF=0.3%, p-value=3.5x10^{-22}) (http://pheweb.sph.umich.edu:5003/pheno/174).

As shown in Table S4-3, a well-known stop-gain variant rs74315329 in the gene *MYOC* for glaucoma was identified for glaucoma (PheCode 365 with 4,462 cases and 397,761 controls). This rare variant has MAF 0.14%. If rare variants were excluded from the analysis due to difficulties appropriately analyzing them, these associations would not be identified.

Figure 4-1 Manhattan plots of association p values resulting from SAIGE, SAIGE-NoSPA(asymptotically equivalent to GMMAT) and BOLT-LMM

A. coronary artery disease (PheCode 411, case:control = 1:12), B. colorectal cancer (PheCode 153, case:control = 1:84), C. glaucoma (PheCode 365, case: control = 1:89), and D. thyroid cancer (PheCode 193, case:control=1:1138). Blue: loci with association p-value < 5x10-8, which have been previously reported, Green: loci that have association p-value < 5x10-8 and have not been reported before. Since results from SAIGE-noSPA and BOLT-LMM contain many false positive signals for colorectal cancer, glaucoma, and thyroid cancer, the significant loci are not highlighted.



A. Coronary Artery Disease

BOLT-LMM

SAIGE-NoSPA (GMMAT)

SAIGE

B. Colorectal Cancer

BOLT-LMM

SAIGE-NoSPA (GMMAT)

SAIGE

C. Glaucoma

BOLT-LMM

SAIGE-NoSPA (GMMAT)

SAIGE

D. Thyroid Cancer

BOLT-LMM

SAIGE-NoSPA (GMMAT)

SAIGE

Figure 4-2 Quantile-quantile plots of association p-values resulting from SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT) and BOLT-LMM (non-infinitesimal mixed model association test p-value)

A. coronary artery disease (PheCode 411, case: control = 1:12), B. colorectal cancer (PheCode 153, case: control = 1:84), C. glaucoma (PheCode 365, case: control = 1:89), and D. thyroid cancer (PheCode 193, case: control=1:1138).

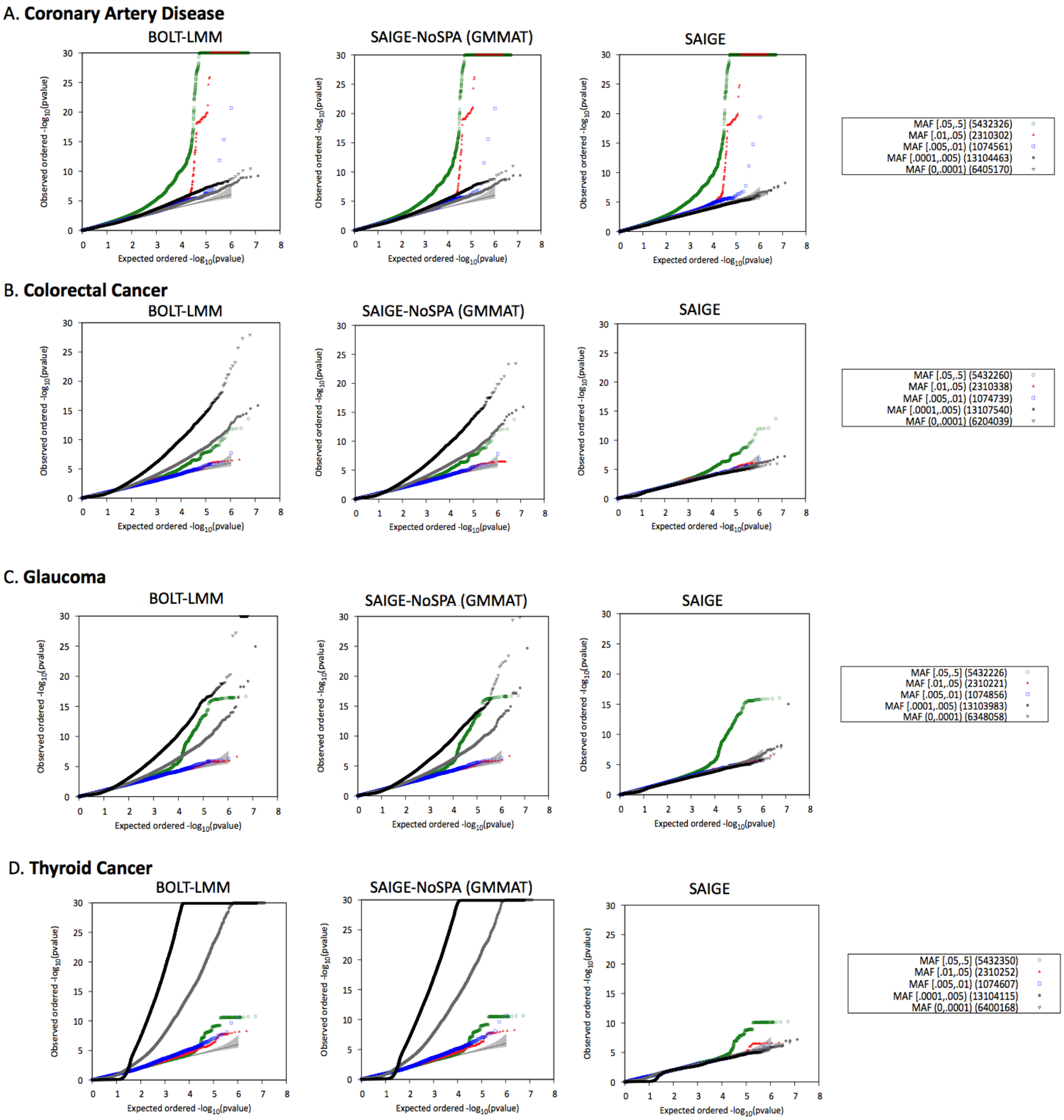


Table 4-1 Comparison of different methods for GWAS with mixed effect models

		Method Features					**Algorithm Complexity**				**Benchmarks for UK Biobank Data Coronary Artery Disease (PheCode 411)**	
		Does not require pre-computed genetic relationship matrix	**Feasible for large sample sizes**	**Developed for binary traits**	**Accounts for unbalanced case-control ratio**	**Tests quantitative traits**	**Time complexity**		**Memory usage (Gbyte)**		**Time CPU hrs**	**Memory**
							Step 1	**Step 2**	**Step 1**	**Step 2**		
Logistic mixed model	**SAIGE**	✓	✓	✓	✓	✓	$O(PM_1N^{1.5})$ *	$O(MN)$	$M_1N/4$	N	517	10.3G
	GMMAT			✓		✓	$O(PN^3)$	$O(MN^2)$	FN^2	FN^2	NA	NA
Linear mixed model	**BOLT-LMM**	✓	✓			✓	$O(PM_1N^{1.5})$ *	$O(MN)$	$M_1N/4$	N	360	10.9G
	GEMMA					✓	$O(N^3)$	$O(MN^2)$	FN^2	FN^2	NA	NA

N: number of samples

P: number of iterations required to reach convergence

M_1: number of markers used to construct the kinship matrix

M: total number of markers to be tested

F: Byte for floating number

* Number of iterations in PCG is assumed as $O(N^{0.5})$ (Loh, et al., 2015)

Figure S4-1 Plot of the ratio of the variances of the score statistics with and without incorporating the variance components for the random effects

for A. 1,000 simulated markers with MAF spectrum shown in Figure S4-10 and B. 669 out of 1,000 markers that have MAC < 200. 1,000 families were simulated based on the pedigree structure shown in Figure S4-4 with case control ratio 1:9.

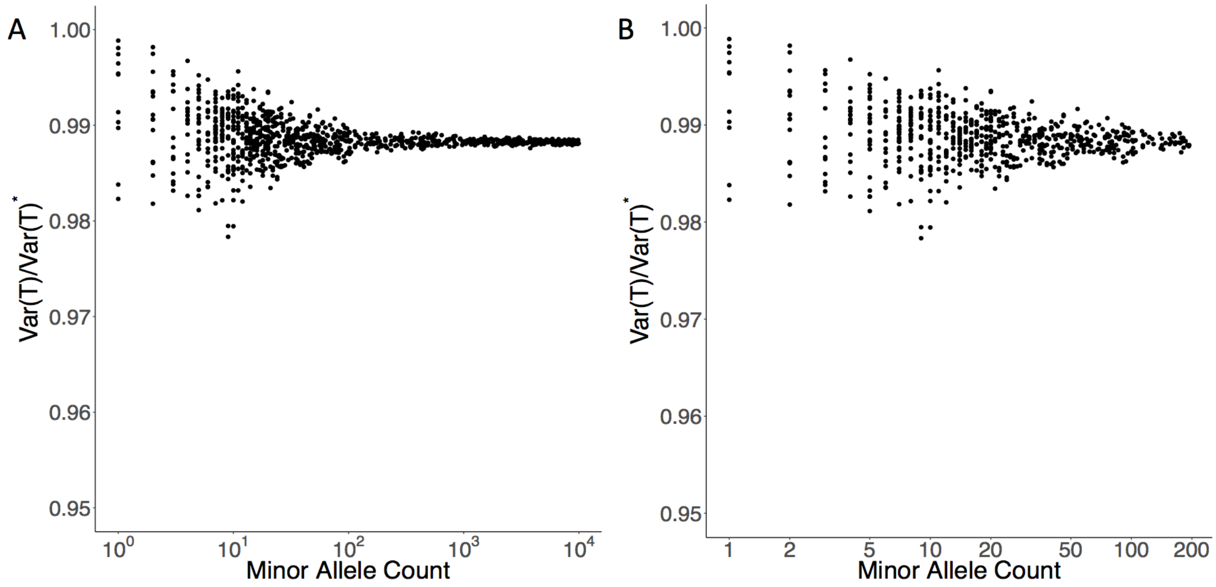


Figure S4-2 Log-log plots of the estimated run time (A) and memory use (B) as a function of sample size (N). Numerical data are provided in Table S4-1.

Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,458 white British participants and 200,000 markers for the cardiovascular diseases (PheCode = 411). The plotted run time is the projected computation time for testing 71 million markers with info ≥ 0.3. The reported run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. Software versions: BOLT-LMM, v2.3; GEMMA, v0.96. BOLT-LMM: compute association statistics under the non-infinitesimal model; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

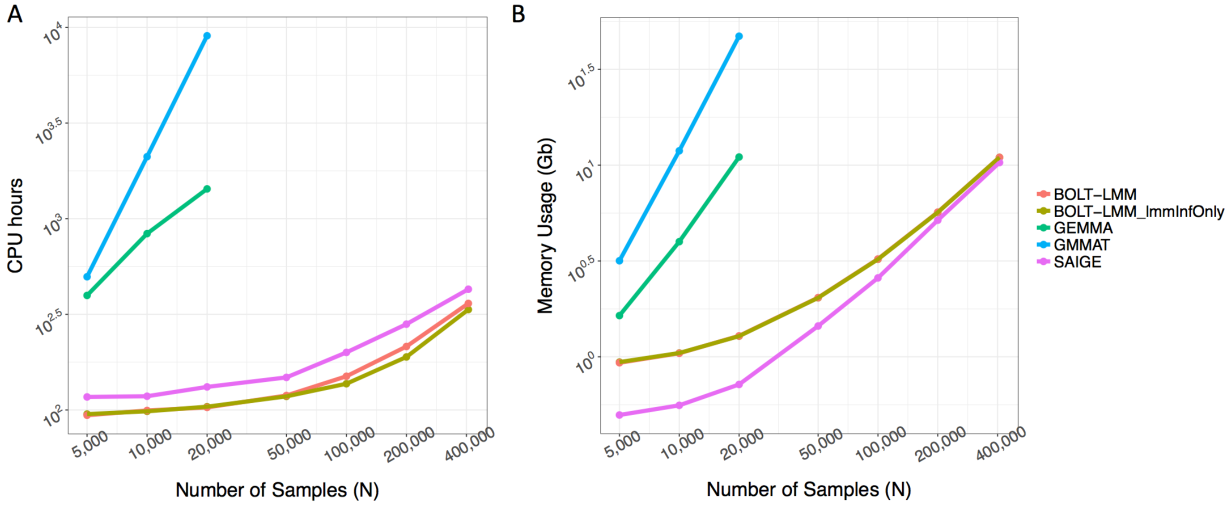


Figure S4-3 Histogram of case-control ratios of 1,688 disease-specific binary phenotypes in the UK Biobank data.

Phenotypes were constructed based on ICD-9 and ICD-10 codes using a previously described scheme(Denny, et al., 2013).

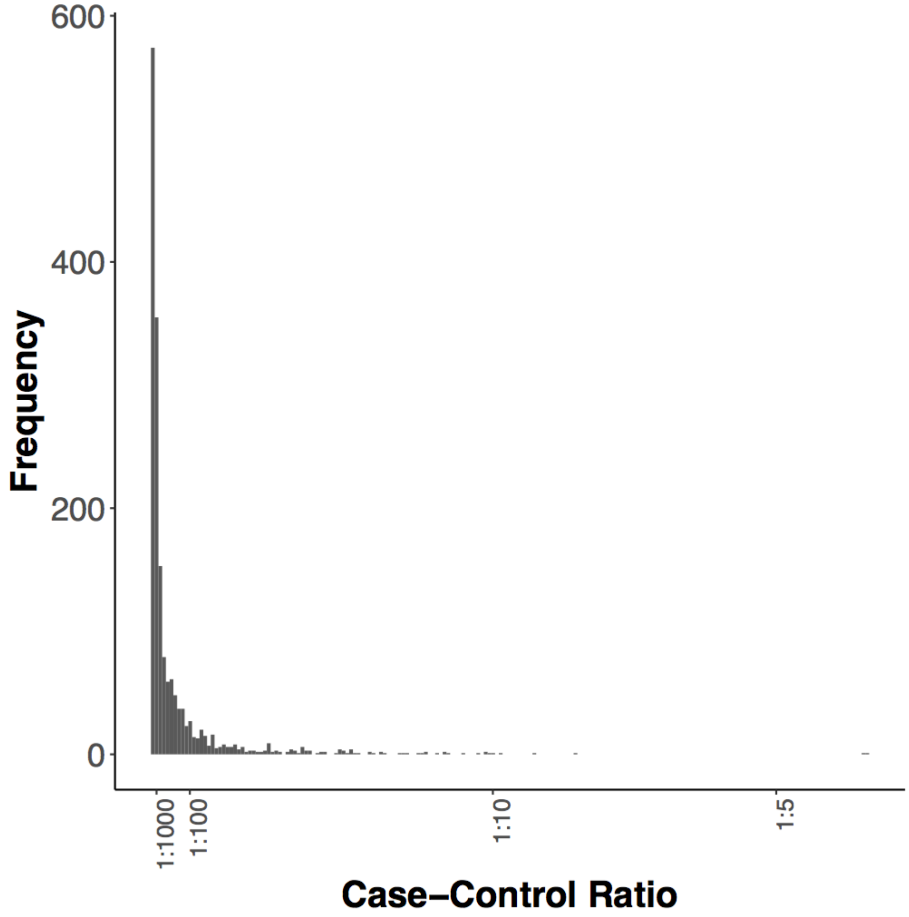


Figure S4-4 Pedigree of families, each with 10 members, in the simulation study.

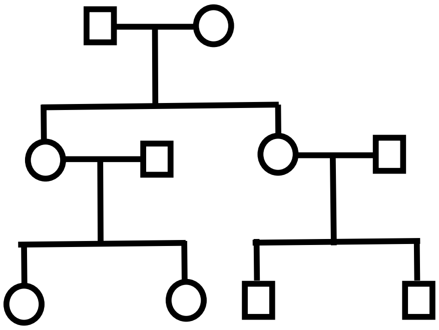


Figure S4-5 Histogram of the GMMAT test statistics (solid black line) overlaid with the standard normal density curve (red dotted line) for 1,000,000 simulated genetic markers for different case-control ratios

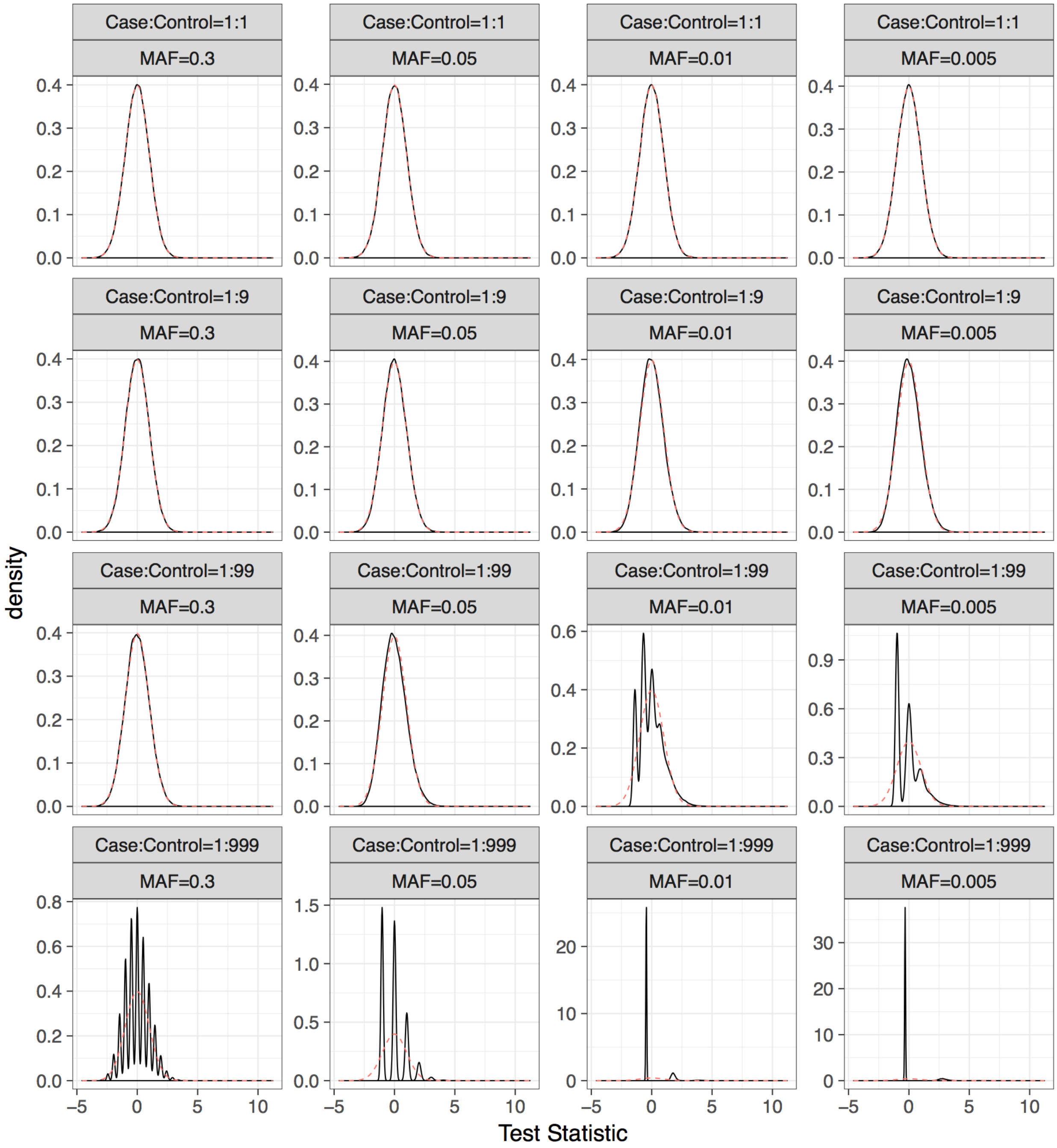


Figure S4-6 Quantile-quantile plots of association p-values for 1,000,000 variants having MAF = 0.005 from the simulation study.

The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is for p -values from the GMMAT (Chen, et al., 2016) program. The fourth column is comparing the p-values from SAIGE and from GMMAT (Chen, et al., 2016). The fifth column is comparing the p-values from SAIGE-NoSPA and from GMMAT (Chen, et al., 2016). The black lines indicate x = y. τ: variance component parameter.

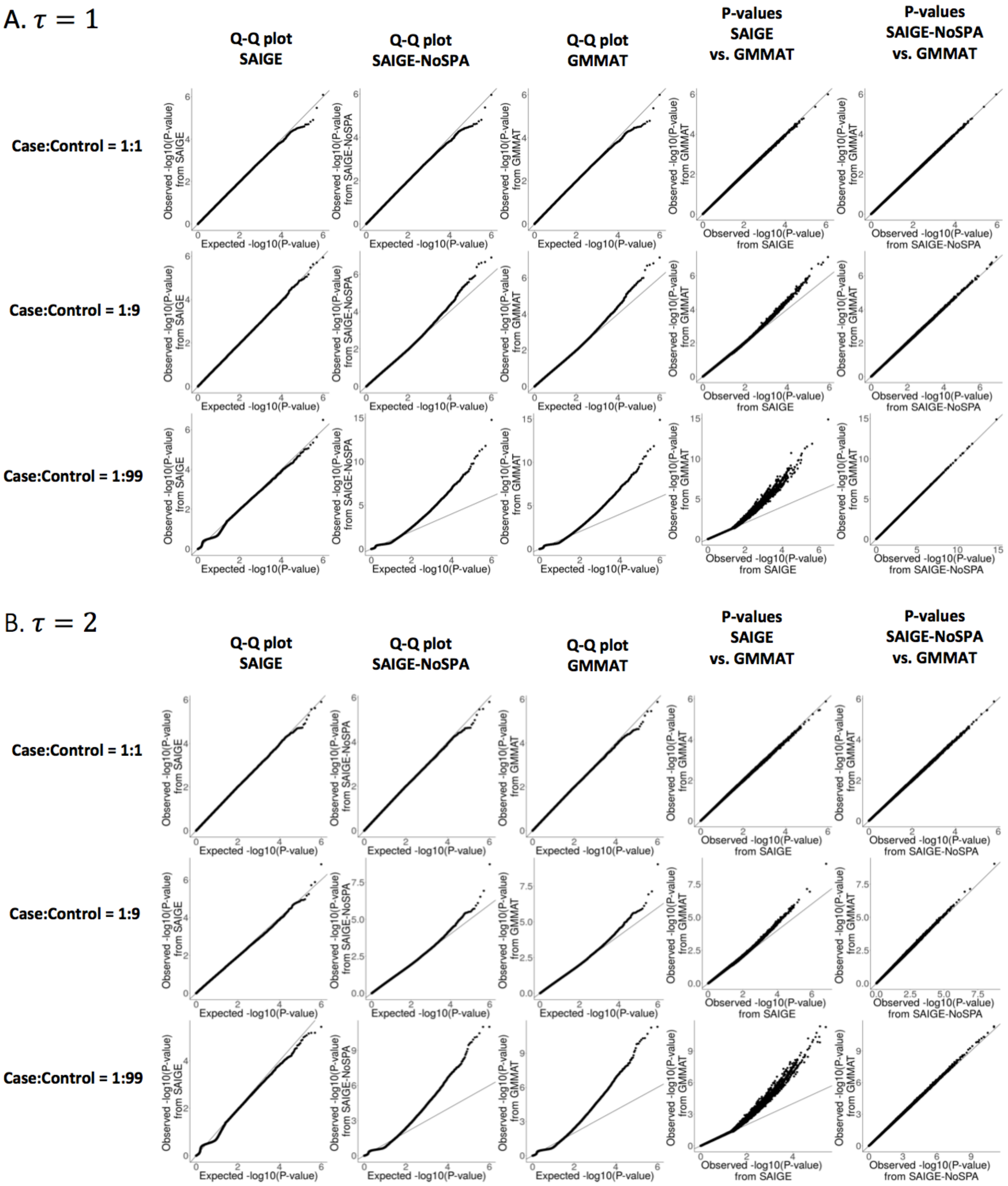

A. $\tau = 1$
Q-Q plot SAIGE
Q-Q plot SAIGE-NoSPA
Q-Q plot GMMAT
P-values SAIGE vs. GMMAT
P-values SAIGE-NoSPA vs. GMMAT
Case:Control = 1:1
Case:Control = 1:9
Case:Control = 1:99
Observed -log10(P-value) from SAIGE
Observed -log10(P-value) from SAIGE-NoSPA
Observed -log10(P-value) from GMMAT
Expected -log10(P-value)
Observed -log10(P-value) from SAIGE
Observed -log10(P-value) from SAIGE-NoSPA
B. $\tau = 2$
Q-Q plot SAIGE
Q-Q plot SAIGE-NoSPA
Q-Q plot GMMAT
P-values SAIGE vs. GMMAT
P-values SAIGE-NoSPA vs. GMMAT
Case:Control = 1:1
Case:Control = 1:9
Case:Control = 1:99
Observed -log10(P-value) from SAIGE
Observed -log10(P-value) from SAIGE-NoSPA
Observed -log10(P-value) from GMMAT
Expected -log10(P-value)
Observed -log10(P-value) from SAIGE
Observed -log10(P-value) from SAIGE-NoSPA

Figure S4-7 Quantile-quantile plots of association p-values for 1000,000 variants with 10,000 samples with very unbalanced case-control ratio 1:99 from the simulation study.

The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is for p -values from the GMMAT (Chen, et al., 2016) program. The fourth column is comparing the p-values from SAIGE and from GMMAT (Chen, et al., 2016). The fifth column is comparing the p-values from SAIGE-NoSPA and from GMMAT (Chen, et al., 2016). The black lines indicate x = y. τ: variance component parameter.

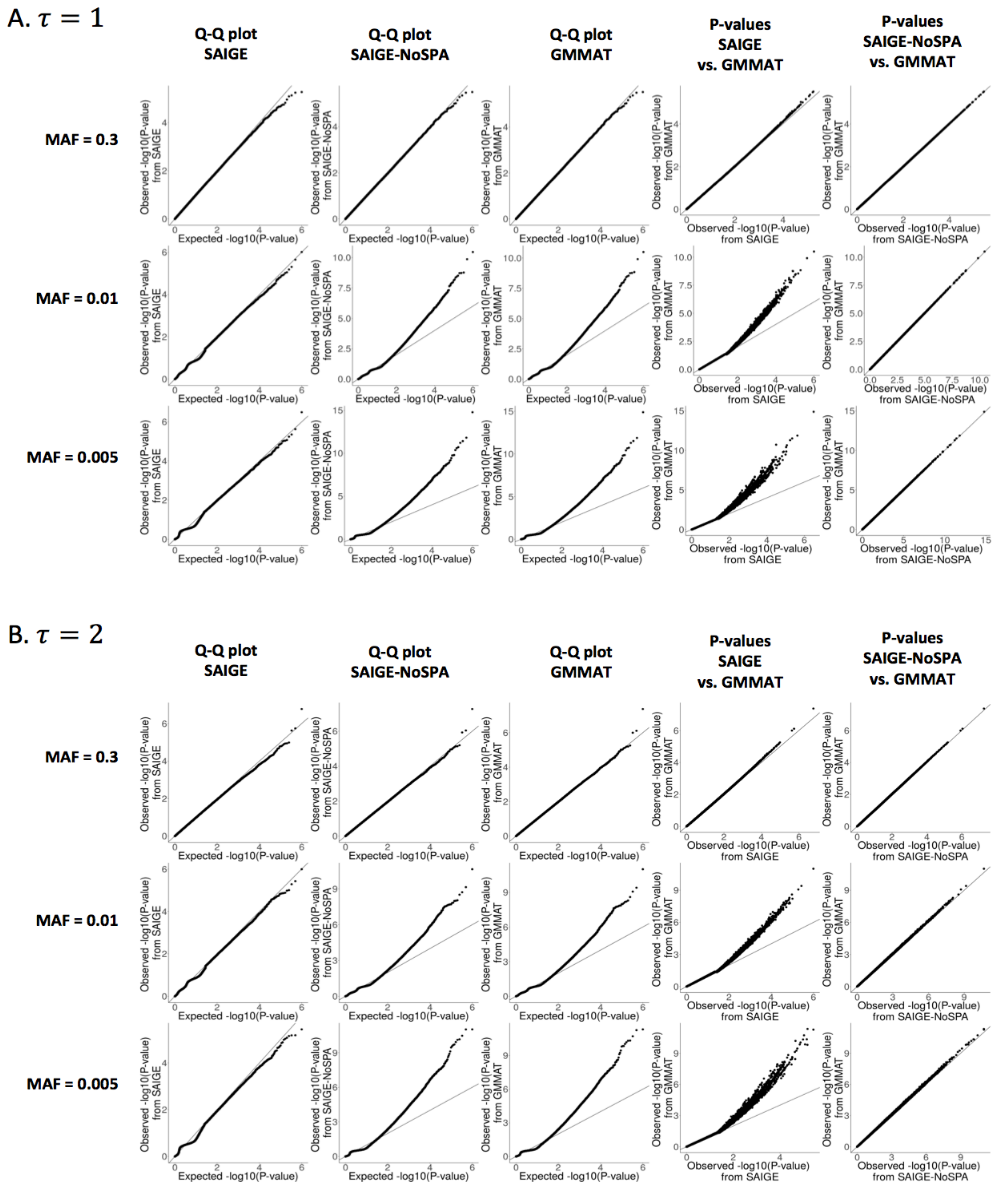


Figure S4-8 Plots for the first four PCs based on the 93,511 simulated markers for samples from the two subpopulations.

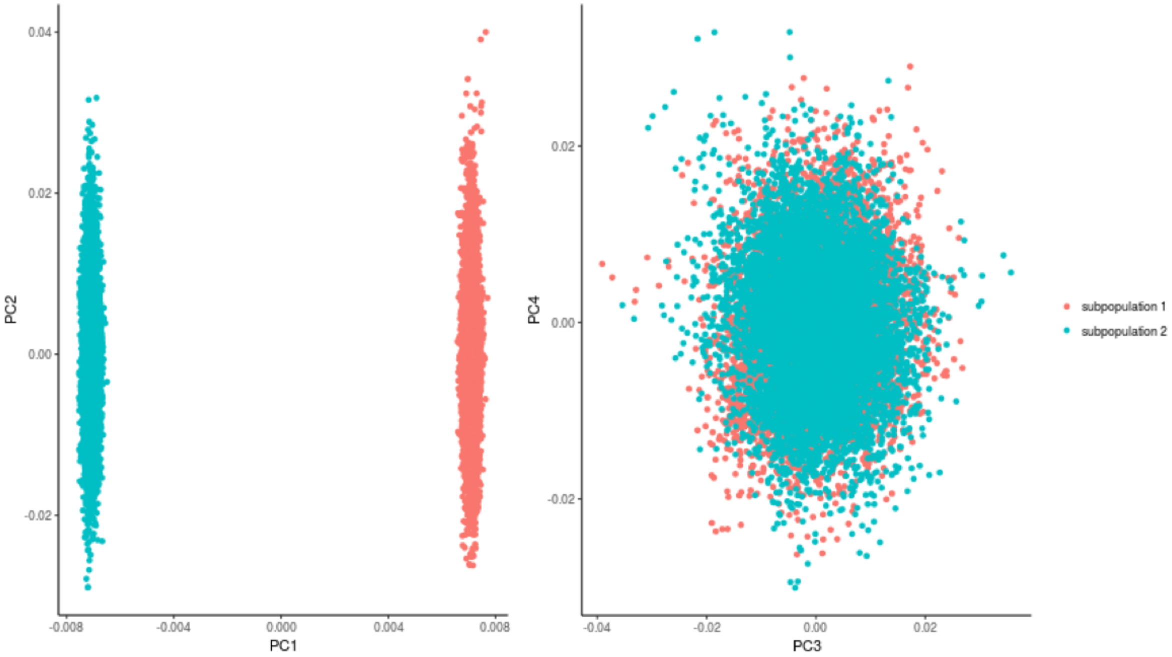


Figure S4-9 Quantile-Quantile plots for the association p-values for ~10 million simulated genetic markers with MAC > 20 in presence of two subpopulations, each having 10,000 samples, with Fst = 0.013. τ: variance component parameter.

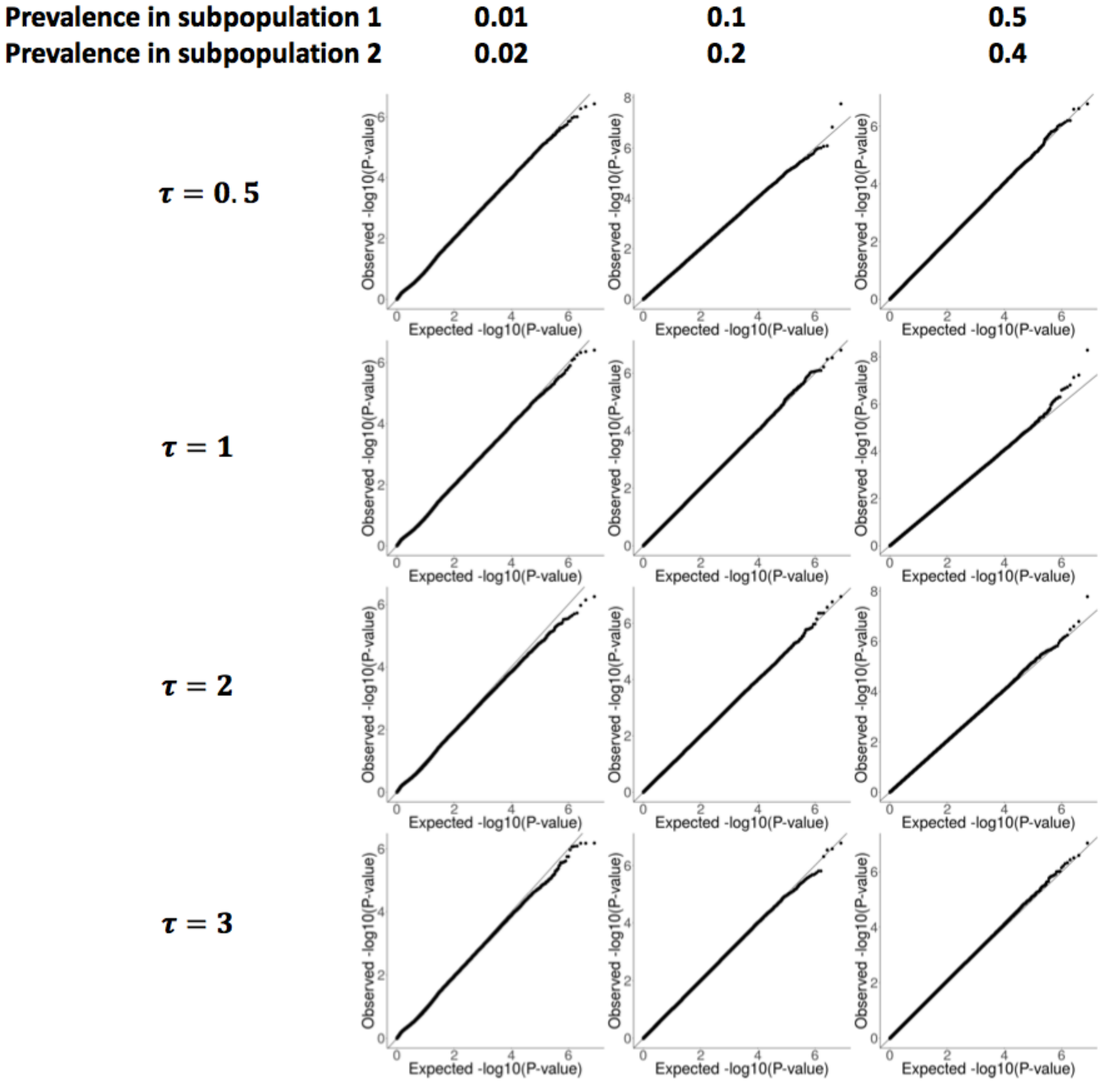


Figure S4-10 Empirical power of SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT), BOLT-LMM_lmmInfOnly (compute mixed model association statistics under the infinitesimal model), and BOLT-LMM (compute mixed model association statistics under the non-infinitesimal model) at the test-specific empirical α levels that yield type I error rate α = 5x10-8, when the variance component parameter τ=1.

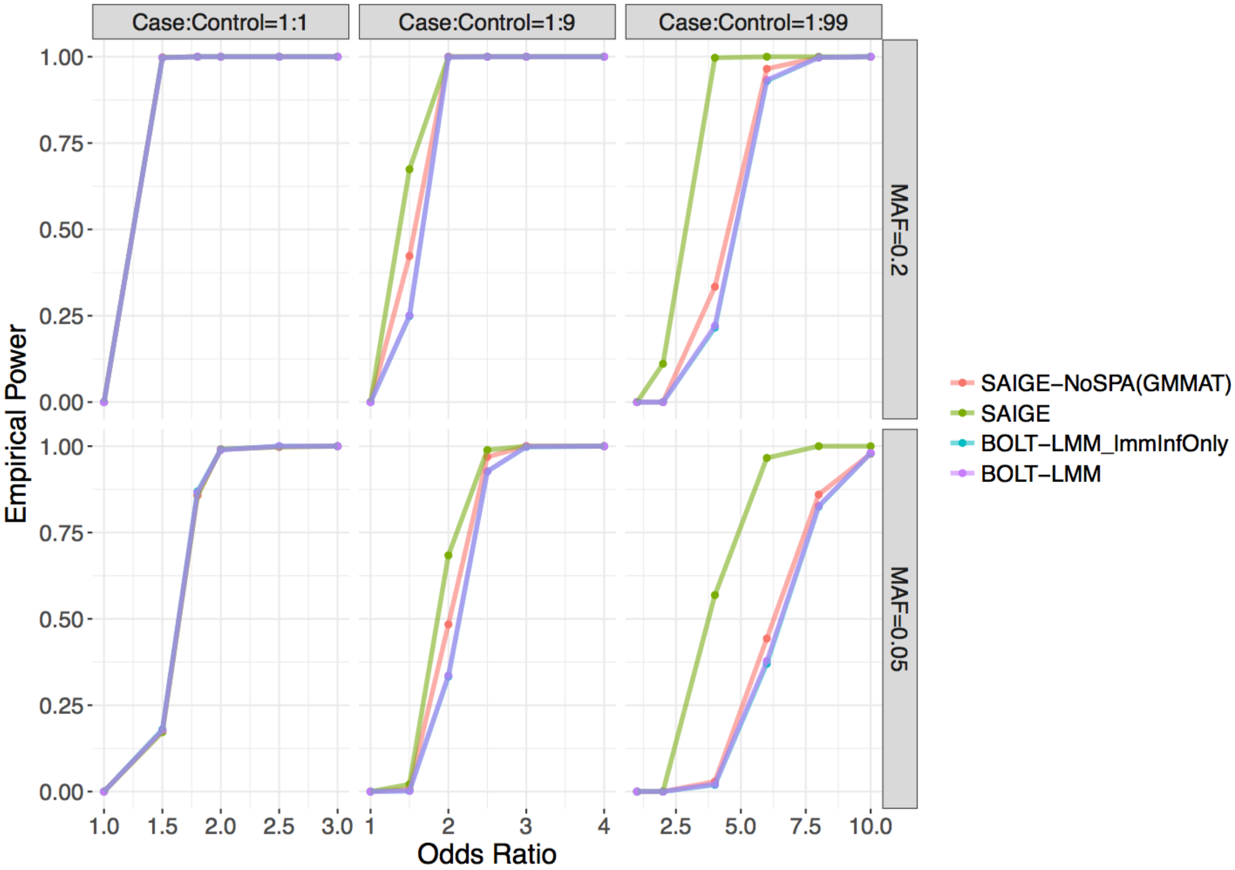


Figure S4-11 Distribution of the minor allele frequency spectrum for randomly selected 1,000,000 markers in the simulation study.

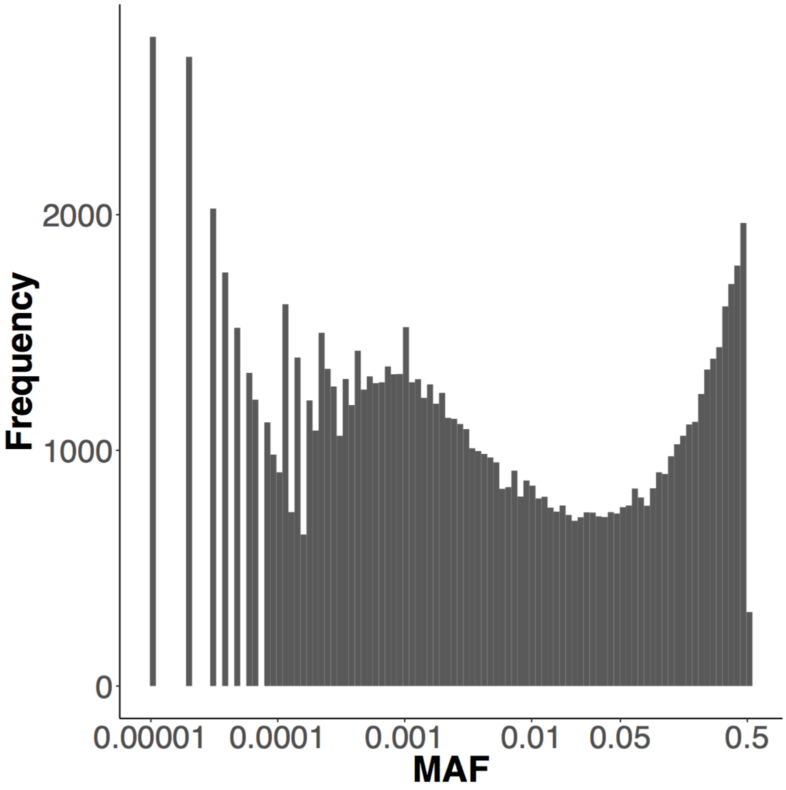


Figure S4-12 Histogram of the average numbers of iterations for PCG to converge in the process of fitting the null logistic mixed model for the 1,283 non-sex specific binary phenotypes that have at least 50 cases in the UK Biobank. The PCG convergence tolerance was 1x10^{-5}.

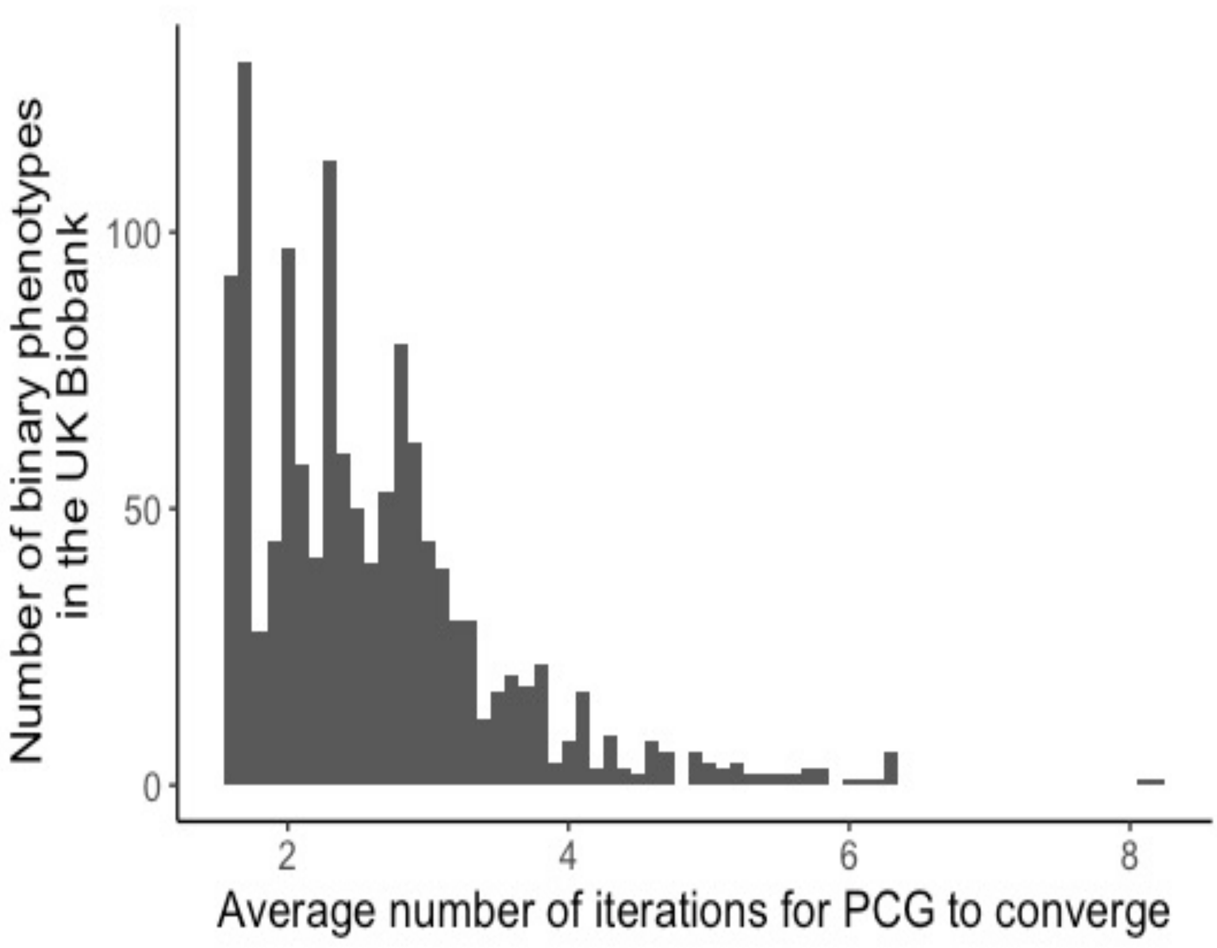


Figure S4-13 A. The trace of the matrix Pψ (estimated using the random trace estimator based on 10, 20, 30, 40 and 50 random vectors) is plotted against the true traces that were computed as the sum of the elements on the main diagonal of matrix Pψ; B. The coefficient of variation (CV) for the trace estimator is plotted

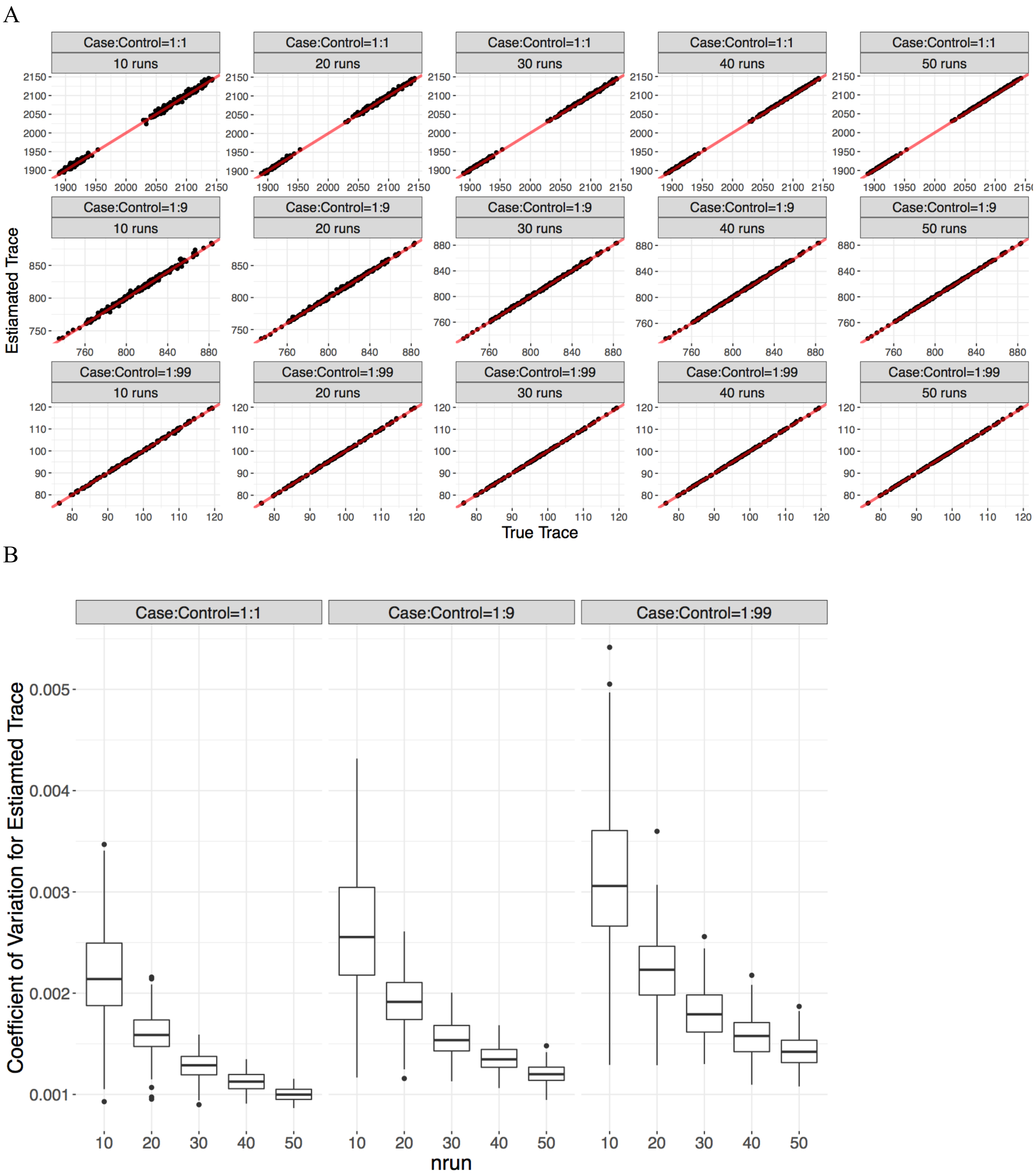
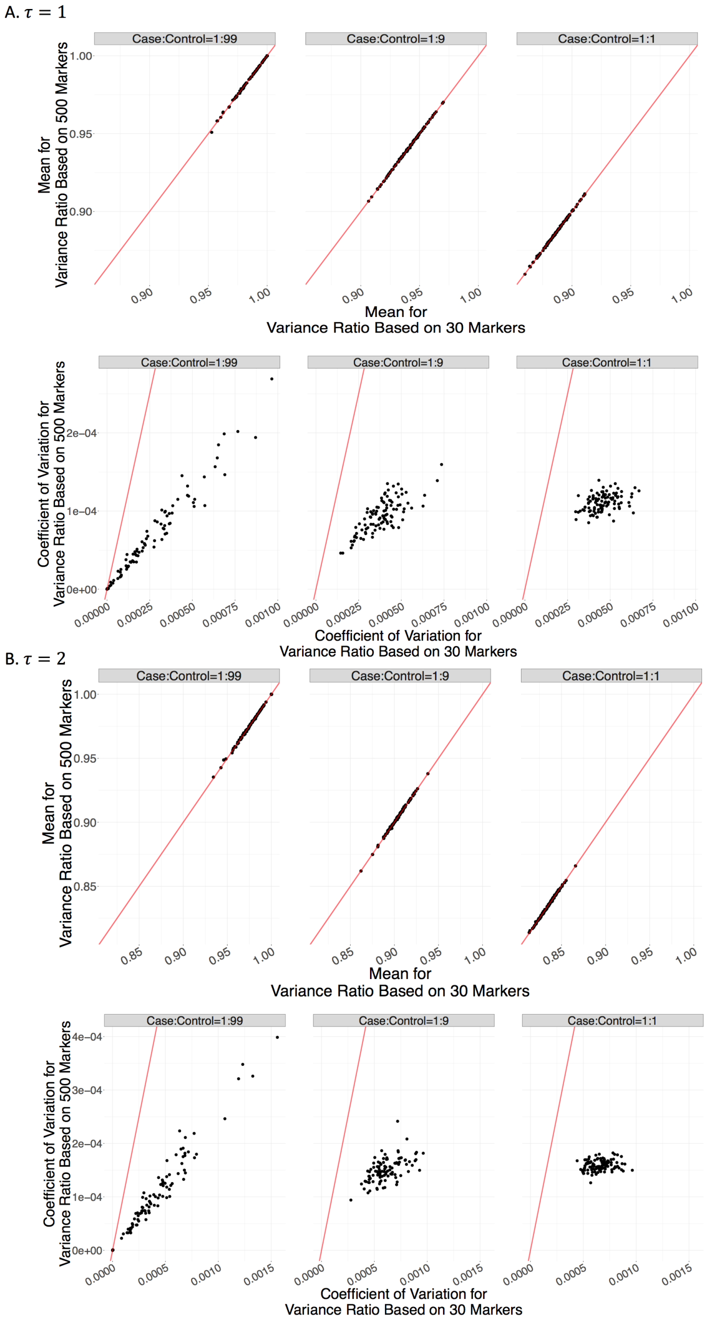


Figure S4-14 The variance ratios and the coefficient of variation (CV) estimated based on 30 random genetic markers were plotted against those based on 500 markers, respectively. The thin red lines indicate x = y. τ: variance component parameter.



A. $\tau = 1$

B. $\tau = 2$

Figure S4-15 Comparing association p-values for ~ 28 million genotyped or HRC imputed genetic markers with and without the leave-one-chromosome-out (LOCO) approach.

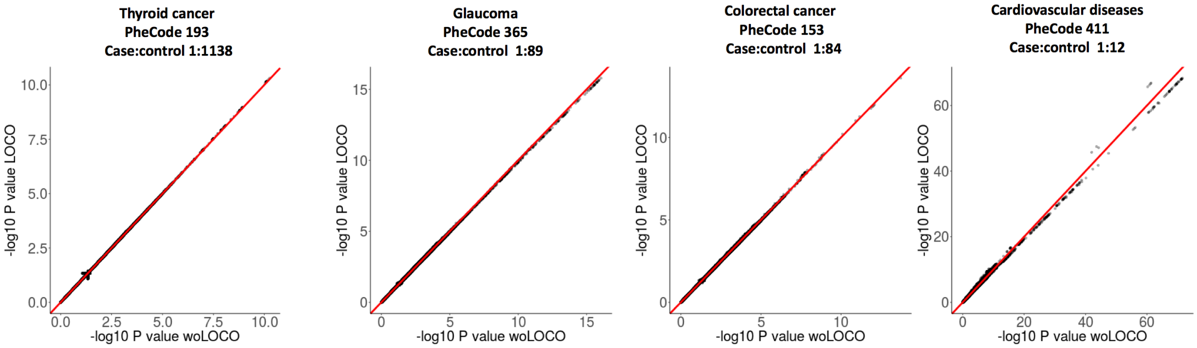


Figure S4-16 Quantile-quantile plots of association p-values for 1,000,000 variants having case-control ratio 1:99 from the simulation study with MAF = 0.005. The first column is p-values using τ̂ estimated by SAIGE. The second column is for p-values using true τ. The third column is for p-values using τ̂ estimated by PCGC. The fourth column is comparing the p-values using τ̂ estimated by SAIGE and using true τ. The fifth column is comparing the p-values using τ̂ estimated by PCGC and using true τ. The thin black lines indicate x = y. τ: variance component parameter.

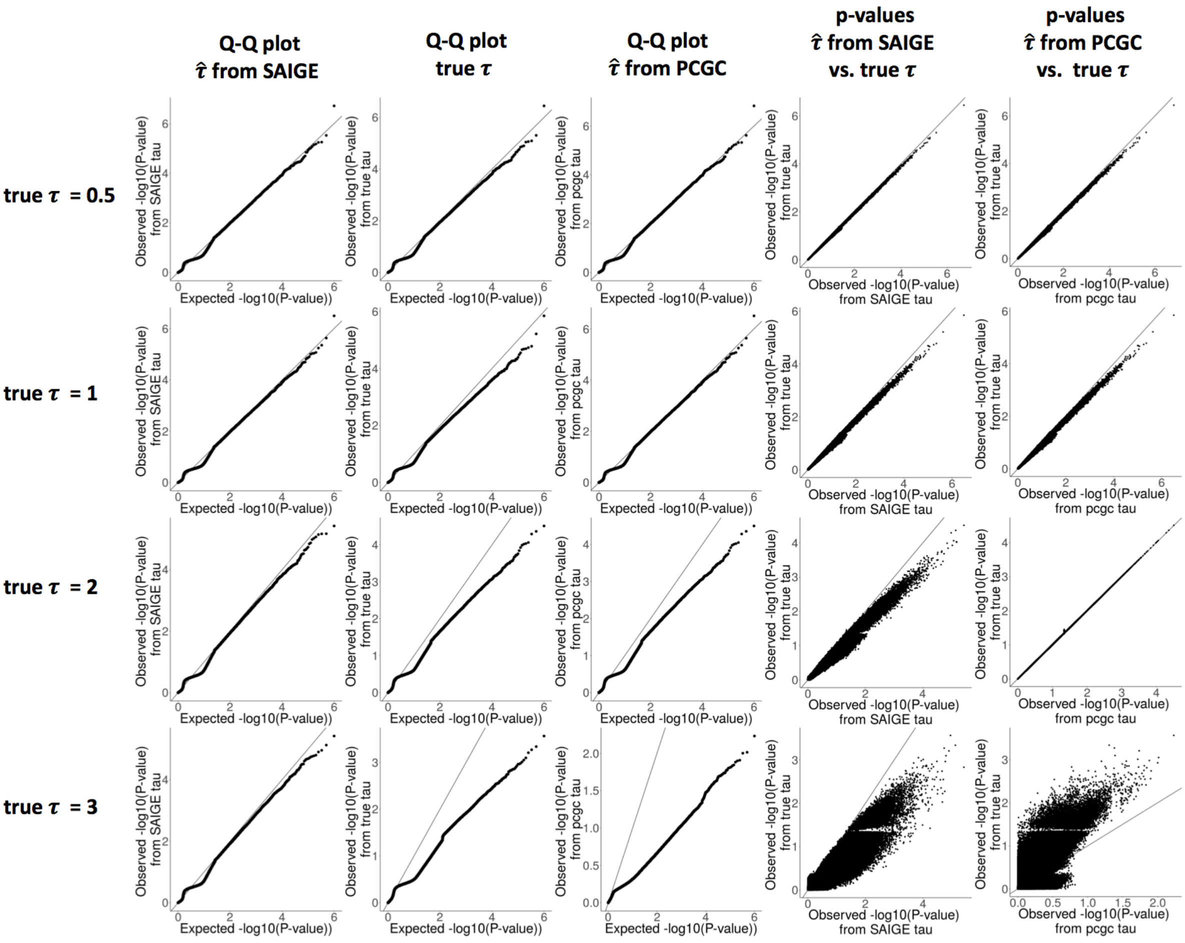
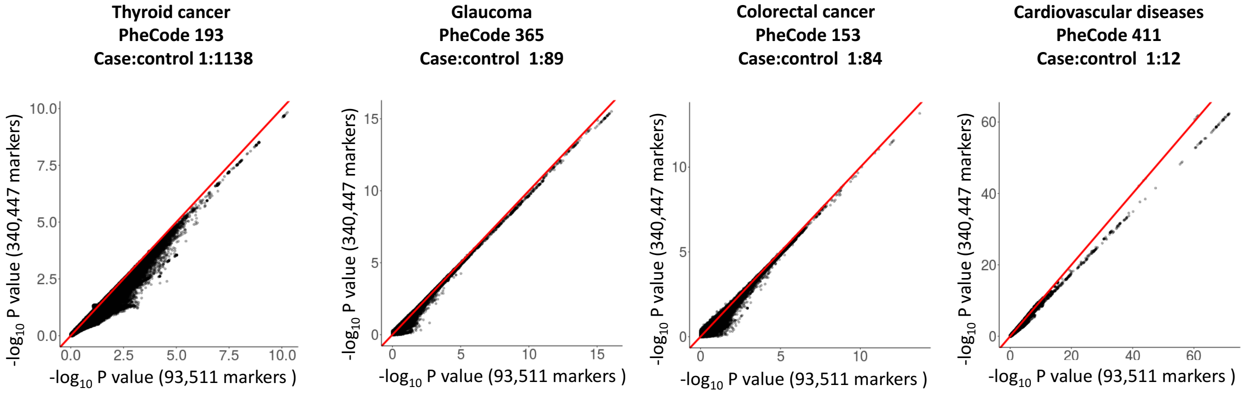


Figure S4-17 Comparing association p-values for ~ 28 million genotyped or HRC imputed genetic markers for all four randomly select exemplary binary phenotypes in the UK Biobank data with a low-rank genetic relationship matrix (GRM) constructed using 93,511 genotyped markers and a GRM constructed using 340,447 directly genotyped markers



Thyroid cancer
PheCode 193
Case:control 1:1138

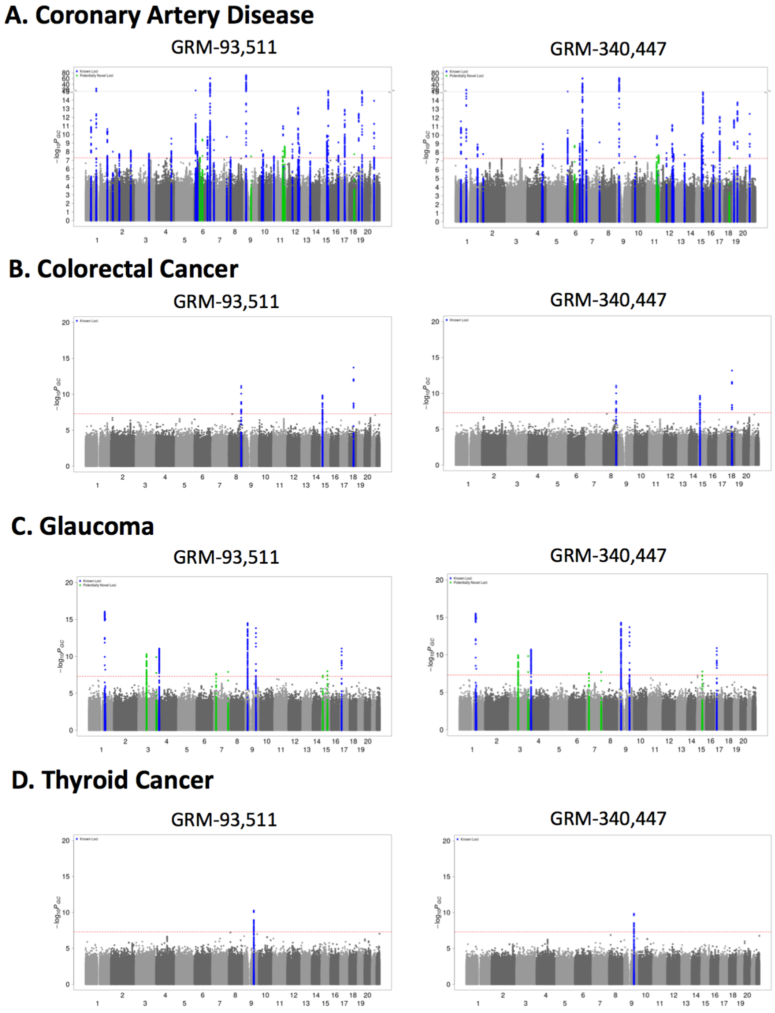
Glaucoma
PheCode 365
Case:control 1:89

Colorectal cancer
PheCode 153
Case:control 1:84

Cardiovascular diseases
PheCode 411
Case:control 1:12

Figure S4-18 Manhattan plots of association p values resulting from SAIGE with a genetic relationship matrix (GRM) constructed using 93,511 genotyped markers and a GRM constructed using 340,447 genotyped markers

A. coronary artery disease (PheCode 411, case:control = 1:12), B. colorectal cancer (PheCode 153, case:control = 1:84),), C. glaucoma (PheCode 365, case: control = 1:89), and D. thyroid cancer (PheCode 193, case:control=1:1138). Blue: loci that have association p-value < 5x10^{-8}, where the top hits are previously reported, Green: loci that have association p-value < 5x10^{-8} and have not been reported before.



A. Coronary Artery Disease

GRM-93,511

GRM-340,447

B. Colorectal Cancer

GRM-93,511

GRM-340,447

C. Glaucoma

GRM-93,511

GRM-340,447

D. Thyroid Cancer

GRM-93,511

GRM-340,447

Table S4-1 The estimated run time (A) and memory use (B) across different sample sizes.

Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,458 white British participants and 200,000 markers for the cardiovascular diseases (PheCode = 411). The run time is the projected computation time for testing 71 million markers with info ≥ 0.3. The reported run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. Software versions: BOLT-LMM, v2.3; GEMMA, v0.96. BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Sample Size(N)	Time (CPU hours)	Memory(Gb)	Tests
5,000	497.19	3.17	GMMAT
10,000	2109.83	11.88	GMMAT
20,000	9046.04	47.09	GMMAT
5,000	95.18	0.94	BOLT-LMM_lmmInfOnly
10,000	98.40	1.05	BOLT-LMM_lmmInfOnly
20,000	104.05	1.28	BOLT-LMM_lmmInfOnly
50,000	117.80	2.03	BOLT-LMM_lmmInfOnly
100,000	137.16	3.23	BOLT-LMM_lmmInfOnly
200,000	189.28	5.67	BOLT-LMM_lmmInfOnly
408,458	335.00	10.98	BOLT-LMM_lmmInfOnly
5,000	93.89	0.93	BOLT-LMM
10,000	99.39	1.04	BOLT-LMM
20,000	103.15	1.29	BOLT-LMM
50,000	119.03	2.04	BOLT-LMM
100,000	150.02	3.24	BOLT-LMM
200,000	214.71	5.69	BOLT-LMM
408,458	360.63	10.98	BOLT-LMM
5,000	397.00	1.64	GEMMA
10,000	835.59	3.99	GEMMA
20,000	1431.69	11.03	GEMMA
5,000	117.50	0.50	SAIGE
10,000	118.83	0.56	SAIGE
20,000	133.32	0.72	SAIGE
50,000	153.60	1.45	SAIGE
100,000	211.21	2.58	SAIGE
200,000	312.81	5.16	SAIGE
408,458	517.38	10.32	SAIGE

Table S4-2 Number of genetic variants and loci that passed the genome-wide significant threshold ($P < 5x10^{-8}$) for the three 'real data' phenotypes identified by SAIGE, SAIGE-NoSPA(asymptotically equivalent to GMMAT), and BOLT-LMM in the UK Biobank data.

Since results from SAIGE-NoSPA and BOLT-LMM contain many false positive signals for colorectal cancer, Glaucoma, and thyroid cancer, the numbers of loci are not provided.

Phenotype	**Tests**	**Number of variants with p-value < $5x10^{-8}$**	**Number of all loci with top p-value < $5x10^{-8}$**	**Number of all loci with top p-value < $5x10^{-8}$ and have not been previously reported**
Coronary artery disease PheCode 411 case:control 1:12	SAIGE	1,733	40	6
	SAIGE-NoSPA	1,820	101	68
	BOLT-LMM	1,886	89	58
Colorectal cancer PheCode 153 case:control 1:84	SAIGE	77	3	3
	SAIGE-NoSPA	2,950	NA	NA
	BOLT-LMM	3,349	NA	NA
Glaucoma PheCode 365 case:control 1:89	SAIGE	362	12	6
	SAIGE-NoSPA	3,278	NA	NA
	BOLT-LMM	4,228	NA	NA
Thyroid cancer PheCode 193 case:control=1:1138	SAIGE	125	1	1
	SAIGE-NoSPA	73,382	NA	NA
	BOLT-LMM	79,269	NA	NA

Table S4-3 Loci that passed the genome-wide significant threshold ($P < 5x10^{-8}$) for the three phenotypes identified by the SAIGE in the UK Biobank data.

Phenotype	Location	Chr:Pos	rsID	Ref	Alt	Function	Gene	MAF	Sample Size	P value	Known for CAD	Previous Findings
Cardiovascular diseases PheCode 411 case:control =1:12	1p32.3	1:55505647	rs11591147	G	T	Exonic	PCSK9	0.018	408,458	2.30E-12	known	(Kathiresan, 2008)
	1p32.2	1:56966350	rs17114046	A	G	Intronic	PLPP3	0.092	408,458	1.36E-11	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	1p13.3	1:109817590	rs12740374	G	T	UTR3	CELSR2	0.222	408,458	1.68E-25	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	1q41	1:222814442	rs2133189	C	T	Intronic	MIA3	0.286	408,458	2.35E-11	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	2p24.1	2:19942473	rs16986953	G	A	Intergenic	OSR1; LINC00954	0.068	408,458	9.96E-09	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	2p11.2	2:85767735	rs2028900	C	T	Intronic	MAT2A	0.450	408,458	1.82E-08	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	2q33.2	2:203968973	rs72934535	T	C	Intronic	NBEAL1	0.108	408,458	7.14E-09	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	3q22.3	3:136294757	rs13065626	C	G	Intronic	STAG1	0.137	408,458	1.63E-08	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	4q32.1	4:156645513	rs13139571	C	A	Intronic	GUCY1A3	0.233	408,458	2.94E-10	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	6p24.1	6:12903957	rs9349379	A	G	Intronic	PHACTR1	0.405	408,458	6.30E-19	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	6p21.33	6:31881731	rs685031	G	A	Intronic	C2	0.389	408,458	9.26E-09	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	6p11.2	6:57113816	rs430918	C	T	Intergenic	RAB23; LOC100506188	0.066	408,458	4.79E-08	potential novel	
	6q14.1	6:82459034	rs78707197	T	C	UTR3	FAM46A	0.022	408,458	3.75E-10	potential novel	
	6q23.2	6:134204247	rs12194592	A	G	ncRNA_intronic	TARID	0.307	408,458	1.95E-10	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	6q26	6:161005610	rs55730499	C	T	Intronic	LPA	0.081	408,458	4.48E-62	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)

	7p21.1	7:19049388	rs2107595	G	A	Intergenic	HDAC9; TWIST1	0.152	408,458	4.23E-10	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	7q36.1	7:150690176	rs3918226	C	T	Intronic	NOS3	0.081	408,458	1.92E-10	known	(Nikpay, et al., 2015)
	8p21.3	8:19870271	rs35237252	C	A	Intergenic	LPL;SLC18A1	0.251	408,458	4.68E-08	known	{CARDIoGRAMplusC4D Consortium, 2013 #212}
	9p21.3	9:22103813	rs1333042	A	G	ncRNA_intronic	CDKN2B-AS1	0.496	408,458	2.29E-72	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	9q21.12	9:73553245	rs150282530	C	T	Intronic	TRPM3	0.001	408,458	3.45E-08	potential novel	
	10p11.23	10:30317073	rs9337951	G	A	Exonic	JCAD	0.345	408,458	7.32E-09	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	10q11.21	10:44687780	rs11238907	T	G	Intergenic	LINC00841; C10orf142	0.115	408,458	1.88E-08	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	11p15.4	11:9766932	rs378825	A	G	Intronic	SWAP70	0.427	408,458	3.43E-08	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	11q22.1	11:100593538	rs633185	G	C	Intronic	ARHGAP42	0.285	408,458	8.81E-09	potential novel	
	11q22.3	11:103673294	rs2839812	T	A	Intergenic	DYNC2H1; MIR4693	0.279	408,458	1.10E-11	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	11q23.3	11:120233626	rs7924772	A	G	intronic	ARHGEF12	0.387	408,458	2.42E-09	potential novel	
	12q13.13	12:54513915	rs11170820	C	G	ncRNA_exonic	FLJ12825	0.058	408,458	1.33E-09	known	(Verweij, et al., 2017)
	12q24.12	12:111904371	rs4766578	T	A	intronic	ATXN2	0.495	408,458	7.97E-14	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	12q24.13	12:112486818	rs17696736	A	G	intronic	NAA25	0.428	408,458	7.93E-11	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	12q24.31	12:121416650	rs1169288	A	C	exonic	HNF1A	0.313	408,458	1.37E-09	known	(Reiner, et al., 2009)
	13q34	13:110837553	rs638634	C	T	intronic	COL4A1	0.302	408,458	1.41E-08	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	15q25.1	15:79132330	rs11072811	A	C	intergenic	ADAMTS7; MORF4L1	0.492	408,458	1.28E-10	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	15q26.1	15:91429287	rs4932373	A	C	intronic	FES	0.326	408,458	1.84E-17	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)

	16q23.3	16:83045790	rs7500448	A	G	Intronic	CDH13	0.254	408,458	8.32E-10	known	(Verweij, et al., 2017)
	17q21.32	17:47340297	rs2011767	C	T	Intergenic	FLJ40194; MIR6129	0.459	408,458	1.33E-13	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	17q21.33	17:47450057	rs7209400	C	T	ncRNA_intronic	LOC102724596	0.453	408,458	2.25E-12	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	18q21.2	18:52723198	rs550780826	A	G	Intergenic	CCDC68; LINC01929	0.004	408,458	1.91E-08	potential novel	
	19p13.2	19:11188164	rs56125973	T	C	Intergenic	SMARCA4; LDLR	0.118	408,458	3.99E-13	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	19q13.32	19:45412079	rs7412	C	T	Exonic	APOE	0.081	408,458	6.98E-17	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
	21q22.11	21:35593827	rs28451064	G	A	Intergenic	LINC00310; KCNE2	0.132	408,458	1.24E-14	known	(CARDIoGRAMplusC4D Consortium, et al., 2013)
Colorectal cancer PheCode 153 case:control = 1:84	8q24.21	8:128413305	rs6983267	G	T	ncRNA_exonic	CCAT2	0.481	387,318	7.03E-12	known	(Haiman, et al., 2007)
	15q13.3	15:33001734	rs58658771	T	A	Intergenic	SCG5; GREM1	0.179	387,318	1.41E-10	known	(Jaeger, et al., 2007)
	18q21.1	18:46448805	rs6507874	T	C	Intronic	SMAD7	0.473	387,318	1.93E-14	known	(Broderick, et al., 2007)
Glaucoma PheCode 365 Case:control 1:89	1q24.1	1:165743523	rs2790049	A	G	ncRNA_exonic	LOC100147773	0.124	402,223	8.71E-17	known	(Burdon, et al., 2011)
	1q24.3	1:171605478	rs74315329	G	A	exonic	MYOC	0.00137	402,223	9.13E-16	known	(Stone, et al., 1997)
	3p12.1	3:85134557	rs9309969	T	G	intronic	CADM2	0.406	402,223	4.94E-11	Potential novel	
	3q27.3	3:186128816	rs56233426	A	G	intergenic	DGKG; LINC02052	0.463	402,223	1.25E-10	Potential novel	
	4p16.1	4:7889096	rs7663205	C	T	intronic	AFAP1	0.400390	402,223	8.82E-12	known	(Gharahkhani, et al., 2014)
	7p15.3	7:22293117	rs113432289	A	C	intronic	RAPGEF5	0.00012	402,223	2.26E-08	Potential novel	
	7q35	7:146348027	rs540694424	G	C	intronic	CNTNAP2	0.00004	402,223	1.27E-08	Potential novel	
	9p21.3	9:22052734	rs6475604	T	C	ncRNA_intronic	CDKN2B-AS1	0.43059	402,223	3.12E-15	known	(Burdon, et al., 2011)
	9q31.1	9:107693201	rs2437812	A	C	intergenic	ABCA1; SLC44A1	0.42358	402,223	1.49E-14	known	(Chen, et al., 2014)
	15q13.1	15:28365618	rs12913832	A	G	intronic	HERC2	0.21416	402,223	4.05E-08	Potential novel	
	15q24.2	15:76049154	rs187112398	C	T	intergenic	DNM1P35; MIR4313	0.00079	402,223	1.03E-08	Potential novel	
	17p13.1	17:10031090	rs12150284	C	T	intronic	GAS7	0.37337	402,223	8.70E-12	known	(Bailey, et al., 2016)

Thyroid cancer PheCode 193 case:control =1:1138	9q22.33	9:100546600	rs925489	C	T	ncRNA_intronic	PTCSC2	0.332	407,757	5.43E-11	known	(Pereira, et al., 2015)

Table S4-4 Estimated inflation factors of the genomic controls at different p-value quantiles and different MAF cutoffs for SAIGE, SAIGE-NoSPA, and BOLT-LMM test applied on three different phenotypes for 28 million successfully imputed genetic markers (imputation info ≥ 0.3 and MAC ≥ 20) from the UK Biobank data.

			Genomic Control at q^{th} p-value quantile			
Phenotype	**Test**	**MAF cutoffs**	**Including previously reported loci**		**Excluding previously reported loci**	
			q=0.01	**q=0.001**	**q=0.01**	**q=0.001**
Coronary artery disease PheCode 411 case:control 1:12	All variants	SAIGE	1.132	1.244	1.112	1.166
		SAIGE-noSPA	1.155	1.329	1.133	1.249
		BOLT-LMM	1.129	1.306	1.108	1.225
	> 0.01	SAIGE	1.363	1.72	1.284	1.445
		SAIGE-noSPA	1.363	1.721	1.284	1.445
		BOLT-LMM	1.356	1.709	1.277	1.433
	< 0.01	SAIGE	1.046	1.041	1.045	1.04
		SAIGE-noSPA	1.069	1.162	1.069	1.16
		BOLT-LMM	1.031	1.13	1.028	1.13
Colorectal cancer PheCode 153 case:control 1:84	All variants	SAIGE	1.014	1.026	1.01	1.014
		SAIGE-noSPA	1.186	1.555	1.181	1.545
		BOLT-LMM	1.188	1.577	1.182	1.567
	> 0.01	SAIGE	1.051	1.116	1.039	1.073
		SAIGE-noSPA	1.052	1.121	1.04	1.077
		BOLT-LMM	1.057	1.126	1.044	1.085
	< 0.01	SAIGE	0.999	0.993	0.998	0.992
		SAIGE-noSPA	1.253	1.683	1.251	1.681
		BOLT-LMM	1.255	1.709	1.255	1.709
Glaucoma PheCode 365 case:control=1:89	All variants	SAIGE	1.024	1.039	1.021	1.033
		SAIGE-noSPA	1.204	1.576	1.2	1.567
		BOLT-LMM	1.222	1.634	1.216	1.621
	> 0.01	SAIGE	1.077	1.141	1.069	1.114
		SAIGE-noSPA	1.078	1.144	1.07	1.118
		BOLT-LMM	1.085	1.153	1.078	1.126
	< 0.01	SAIGE	1.004	1.003	1.003	1.003
		SAIGE-noSPA	1.266	1.702	1.265	1.702
		BOLT-LMM	1.285	1.77	1.285	1.77
Thyroid cancer PheCode 193 ase:control=1:1138	All variants	SAIGE	1.012	0.992	1.011	0.989
		SAIGE-noSPA	1.964	4.195	1.963	4.194
		BOLT-LMM	2	4.497	1.989	4.497

	> 0.01	SAIGE	1.01	1.036	1.007	1.026
		SAIGE-noSPA	1.015	1.069	1.012	1.058
		BOLT-LMM	1.02	1.074	1.017	1.064
	< 0.01	SAIGE	1.013	0.977	1.013	0.977
		SAIGE-noSPA	2.432	4.737	2.434	4.737
		BOLT-LMM	2.479	5.096	2.479	5.096

Table S4-5 Empirical type 1 error rates for SAIGE, SAIGE-NoSPA, GMMAT, and BOLT-LMM estimated based on 10^9 simulated data sets.

BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Variance Component Parameter τ	**Case:Control**	**Test**	**Empirical Type 1 Error Rates**	
			$\alpha = 5\times10^{-4}$	$\alpha = 5\times10^{-8}$
1	1:1	SAIGE	$5.11\text{x}10^{-4}$	$5.45\text{x}10^{-8}$
		SAIGE-NoSPA	$4.71\text{x}10^{-4}$	$4.00\text{x}10^{-8}$
		GMMAT	$4.66\text{x}10^{-4}$	$3.81\text{x}10^{-8}$
		BOLT-LMM_ lmmInfOnly	$4.83\text{x}10^{-4}$	$4.83\text{x}10^{-8}$
		BOLT-LMM	$4.95\text{x}10^{-4}$	$4.99\text{x}10^{-8}$
	1:9	SAIGE	$4.43\text{x}10^{-4}$	$4.01\text{x}10^{-8}$
		SAIGE-NoSPA	$6.72\text{x}10^{-4}$	$7.82\text{x}10^{-7}$
		GMMAT	$7.30\text{x}10^{-4}$	$1.00\text{x}10^{-6}$
		BOLT-LMM_ lmmInfOnly	$9.01\text{x}10^{-4}$	$2.73\text{x}10^{-6}$
		BOLT-LMM	$9.03\text{x}10^{-4}$	$2.71\text{x}10^{-6}$
	1:99	SAIGE	$3.82\text{x}10^{-4}$	$1.44\text{x}10^{-8}$
		SAIGE-NoSPA	$2.93\text{x}10^{-3}$	$9.76\text{x}10^{-5}$
		GMMAT	$3.31\text{x}10^{-3}$	$1.26\text{x}10^{-4}$
		BOLT-LMM_ lmmInfOnly	$4.02\text{x}10^{-3}$	$2.10\text{x}10^{-4}$
		BOLT-LMM	$4.02\text{x}10^{-3}$	$2.10\text{x}10^{-4}$
2	1:1	SAIGE	$5.15\text{x}10^{-4}$	$3.53\text{x}10^{-8}$
		SAIGE-NoSPA	$4.75\text{x}10^{-4}$	$2.72\text{x}10^{-8}$
		GMMAT	$4.64\text{x}10^{-4}$	$2.56\text{x}10^{-8}$
		BOLT-LMM_ lmmInfOnly	$5.03\text{x}10^{-4}$	$3.74\text{x}10^{-8}$
		BOLT-LMM	$5.21\text{x}10^{-4}$	$3.59\text{x}10^{-8}$
	1:9	SAIGE	$4.07\text{x}10^{-4}$	$3.20\text{x}10^{-8}$
		SAIGE-NoSPA	$5.96\text{x}10^{-4}$	$4.94\text{x}10^{-7}$
		GMMAT	$7.07\text{x}10^{-4}$	$8.01\text{x}10^{-7}$
		BOLT-LMM_ lmmInfOnly	$9.88\text{x}10^{-4}$	$3.51\text{x}10^{-6}$
		BOLT-LMM	$9.88\text{x}10^{-4}$	$3.52\text{x}10^{-6}$
	1:99	SAIGE	$3.53\text{x}10^{-4}$	$2.08\text{x}10^{-8}$
		SAIGE-NoSPA	$2.66\text{x}10^{-3}$	$7.75\text{x}10^{-5}$
		GMMAT	$3.13\text{x}10^{-3}$	$1.08\text{x}10^{-4}$
		BOLT-LMM_ lmmInfOnly	$4.13\text{x}10^{-3}$	$2.30\text{x}10^{-4}$
		BOLT-LMM	$4.13\text{x}10^{-3}$	$2.30\text{x}10^{-4}$

Table S4-6 Test-specific $\boldsymbol{\alpha}$ levels SAIGE and GMMAT where empirical type I errors were equal to 5x10^{-8}.

BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Variance Component Parameter $\boldsymbol{\tau}$	**Case:Control**	**Test**	**Test-specific $\boldsymbol{\alpha}$ levels**
1	1:1	SAIGE	4.74x10^{-8}
		SAIGE-NoSPA	5.70x10^{-8}
		BOLT-LMM_ lmmInfOnly	5.20x10^{-8}
		BOLT-LMM	4.80x10^{-8}
		GMMAT	6.79x10^{-8}
	1:9	SAIGE	6.08x10^{-8}
		SAIGE-NoSPA	6.98x10^{-10}
		BOLT-LMM_ lmmInfOnly	1.60x10^{-11}
		BOLT-LMM	1.70x10^{-11}
		GMMAT	5.29x10^{-10}
	1:99	SAIGE	1.02x10^{-7}
		SAIGE-NoSPA	1.54x10^{-22}
		BOLT-LMM_ lmmInfOnly	5.80x10^{-26}
		BOLT-LMM	8.40x10^{-26}
		GMMAT	1.50x10^{-23}
2	1:1	SAIGE	6.76x10^{-8}
		SAIGE-NoSPA	8.01x10^{-8}
		BOLT-LMM_ lmmInfOnly	6.40x10^{-8}
		BOLT-LMM	6.30x10^{-8}
		GMMAT	8.42x10^{-8}
	1:9	SAIGE	7.85x10^{-8}
		SAIGE-NoSPA	2.30x10^{-9}
		BOLT-LMM_ lmmInfOnly	1.40x10^{-11}
		BOLT-LMM	1.40x10^{-11}
		GMMAT	8.73x10^{-10}
	1:99	SAIGE	1.59x10^{-7}
		SAIGE-NoSPA	2.10x10^{-21}
		BOLT-LMM_ lmmInfOnly	8.10x10^{-28}
		BOLT-LMM	9.60x10^{-28}
		GMMAT	6.69x10^{-23}

Table S4-7 The variance component estimates $\hat{\tau}$ were estimated using SAIGE and PCGC for 100 simulated data sets for each combination of prevalence and the variance component parameter τ.

		$\hat{\tau}$ from PCGC		$\hat{\tau}$ from SAIGE	
Prevalence	τ	**Mean**	**SD**	**Mean**	**SD**
0.01	0.5	1.181	1.602	0.257	0.29
0.01	1	1.226	1.691	0.372	0.295
0.01	2	4.287	12.057	0.631	0.373
0.01	3	6.771	16.294	0.834	0.428
0.1	0.5	0.348	0.158	0.182	0.069
0.1	1	0.701	0.223	0.322	0.075
0.1	2	1.473	0.351	0.534	0.072
0.1	3	2.329	0.523	0.689	0.073
0.5	0.5	0.362	0.085	0.185	0.035
0.5	1	0.714	0.11	0.312	0.036
0.5	2	1.396	0.164	0.481	0.035
0.5	3	2.022	0.234	0.58	0.037

Table S4-8 Estimated inflation factors of the genomic factor (λ) at different p-value quantiles and different MAF cutoffs when applying SAIGE using 93,511 genetic markers to construct GRM vs. using 340,447 genetic markers to construct GRM on four different phenotypes for 28 million successfully imputed genetic markers (imputation info ≥ 0.3 and MAC ≥ 20) from the UK Biobank data

τ.

			Genomic Control at qth p-value quantile			
Phenotype	**Test**	**MAF cutoffs**	**Including previously reported loci**		**Excluding previously reported loci**	
			q=0.01	**q=0.001**	**q=0.01**	**q=0.001**
Coronary artery disease PheCode 411 case:control 1:12	All variants	GRM-93,511	1.132	1.244	1.112	1.166
		GRM-340,447	1.048	1.137	1.032	1.074
	> 0.01	GRM-93,511	1.363	1.72	1.284	1.445
		GRM-340,447	1.217	1.523	1.157	1.277
	< 0.01	GRM-93,511	1.046	1.041	1.045	1.04
		GRM-340,447	0.985	0.98	0.984	0.979
Colorectal cancer PheCode 153 case:control 1:84	All variants	GRM-93,511	1.014	1.026	1.01	1.014
		GRM-340,447	0.993	1.004	0.99	0.993
	> 0.01	GRM-93,511	1.051	1.116	1.039	1.073
		GRM-340,447	1.026	1.088	1.014	1.047
	< 0.01	GRM-93,511	0.999	0.993	0.998	0.992
		GRM-340,447	0.981	0.973	0.98	0.972
Glaucoma PheCode 365 case:control=1:89	All variants	GRM-93,511	1.024	1.039	1.021	1.033
		GRM-340,447	1.008	1.022	1.006	1.015
	> 0.01	GRM-93,511	1.077	1.141	1.069	1.114
		GRM-340,447	1.057	1.119	1.049	1.094
	< 0.01	GRM-93,511	1.004	1.003	1.003	1.003
		GRM-340,447	0.99	0.988	0.989	0.988
Thyroid cancer PheCode 193 ase:control=1:1138	All variants	GRM-93,511	1.012	0.992	1.011	0.989
		GRM-340,447	0.957	0.933	0.956	0.931
	> 0.01	GRM-93,511	1.01	1.036	1.007	1.026
		GRM-340,447	0.969	0.991	0.966	0.981
	< 0.01	GRM-93,511	1.013	0.977	1.013	0.977
		GRM-340,447	0.953	0.91	0.953	0.91

Chapter 5 Discussion

5.1 Results Summary and Future Directions

Bicuspid aortic valve (BAV) is a birth defect of the heart characterized by fusion of two of the normal three leaflets of the aortic valve. Despite its the prevalence(Hoffman and Kaplan, 2002; Tutar, et al., 2005), importance(Roberts and Ko, 2005), and heritability(Cripe, et al., 2004; Ellison, et al., 2007; Emanuel, et al., 1978; Garg, 2006), its genetic origins remain elusive. Previous genetic studies of BAV have focused primarily on linkage analysis in families (Ellison, et al., 2007; Martin, et al., 2007) or sequencing candidate genes in cases (Foffa, et al., 2013) under a hypothesis of Mendelian inheritance. Although GWAS is a widely used approach to identify genetic risk factors for complex diseases, it is challenging to perform GWAS for BAV because an open heart surgery is required to diagnose this congenital heart defect. In Chapter 2, in collaboration with clinicians at the Frankel Cardiovascular Center at the University of Michigan, we performed the first large-scale GWAS on 466 BAV cases, each matched with 10 unaffected controls on age, gender and ancestry. We identified a low-frequency intergenic variant rs6601627 (P = 1.5×10^{-8}) with a substantially higher frequency in BAV cases (8.3%) than in controls (4.2%). We performed a replication study in six BAV study cohorts with an additional 1,021 cases and 5,357 controls and we found strong evidence for association (P=3×10^{-15}). We also identified an independent association signal at a common protein-altering variant p.Ser377Gly in *GATA4* (rs3729856) that is 151 kilobases(kb) away from the first variant (P = 8.8×10^{-8}). Following identification of this novel associated region, we examined the gene expression and chromatin conformation patterns in the region and identified the *GATA4* gene, which encodes one of three major transcription factors that are critical for heart differentiation, as a biological candidate. Next, GATA4 was interrupted by CRISPR-Cas9 in induced pluripotent stem cells from healthy donors. The disruption of *GATA4* significantly impaired the transition from endothelial cells into mesenchymal cells, a critical step in heart valve development.

Compared to large-scale GWASs on other common diseases (e.g. type 2 diabetes and cardiovascular diseases), our study on BAV has a much smaller sample size, leading to relatively low association power in principle. However, multiple features, which were considered carefully when planning the study, make it a valid and well-designed GWAS so that genetic variants with modest effect sizes typically seen for complex traits have been successfully captured. 1. BAV cases were diagnosed by cardiac surgeons upon visual inspection of the aortic valve during open surgery for aneurysm repair or valve replacement and any controls with possible aortic disease were excluded. This avoided any bias due to case-control misclassification. 2. Each case and its ten controls were matched based on ancestry, age, and gender, which avoided batch effects due to sampling. 3. Cases and controls were genotyped simultaneously using the same version of an Exome+GWAS array at the same genotyping center followed by uniform calling using the same cluster file and quality control was performed on a merged data set with cases and controls. This reduced batch effects of genotyping. 4. Genotypes were imputed from a sequencing-based reference panel, HRC. This allowed rare or low-frequency genetic variants to be detected and tested. 5. The DNA array used for genotyping is enriched with coding variants (Illumina Human CoreExome), enabling genetic association tests for directly genotyped protein-altering variants, even those with low frequency. 6. To correct for inflation caused by the unbalanced case-control ratio (1:10) in score tests in the traditional logistic regression, the saddlepoint approximation (SPA) tests were used to test for genetic association.

To further investigate the underlying genetic and epigenetic architecture for BAV, next steps would include but not limited to 1. meta-analyzing our results with other BAV cohorts to increase power and therefore detect novel BAV-associated genetic variants, 2. conducting whole-exome or target sequencing to identify novel pathogenic genetic variants for BAV that were unable to be imputed, possibly due to low frequency 3. performing reduced representation bisulfite sequencing(RRBS) and whole-transcriptome RNA sequencing (RNA-seq) on aortic tissues of BAV patients and healthy controls with tricuspid aortic valves to uncover epigenetic and transcriptional variations that are associated with BAV.

Genotype imputation, which statistically infers missing/unobserved genotypes in low-density genotyped study samples using a reference panel containing high-density genome markers, is considered "in silico genotyping". As trade-offs exist between imputing from a larger, multi-ethnic, publicly available imputation reference panel and a smaller, population-specific panel, we sought to quantify the differences. In Chapter 3, we systematically evaluated and compared the imputation accuracy and numbers of imputed variants using multiple reference panels. Our results confirmed that both the population-specific reference panels (e.g. HUNT whole-genome sequencing reference panel) and the large-scale publicly accessible reference panels (e.g. HRC and 1000G) are valuable for genotype imputation. Furthermore, since different whole-genome reference panels may generate discordant imputed genotypes for the same variants of the same study samples, we performed a simulation study to investigate the optimal strategy to incorporate multiple versions of imputed genotypes to perform GWAS. Our results indicate that testing for association for all imputed genotypes for all genetic variants and using the most significant association p-value results in higher association power than retaining imputed genotypes only from the panel with highest imputation quality metrics for each variant. This holds true even with a more stringent adjustment for the multiple testing burden of the additional variants adjusted.

The strategy we proposed is feasible even when individual-level haplotypes within the reference panel are not accessible, a common issue due to ethical issues surrounding sharing of individual-level genetic data, as is the case with the HRC(McCarthy, et al., 2016). Genotype imputation tools that allow merging reference panels without sharing the individual-level data in each panel for imputation would be very useful. As the cost for genome sequencing drops quickly, reference panel sizes continuously increase. The future direction of genotype imputation would also lie in the development of imputation tools to handle the increase in computational burden.

In Chapter 4, to tackle challenges in PheWAS, an emerging unbiased approach to explore the genome-phenome associations, we have developed a novel method, Scalable and Accurate Implementation of GEneralized mixed model (SAIGE). This method allows for analysis of very large sample sizes for binary traits with unbalanced case-control ratios and also infers and accounts for sample relatedness. SAIGE applies the saddlepoint approximation to correct for inflation caused by the unbalanced case-control ratio in the score tests in logistic mixed models. We have demonstrated that SAIGE controls for the inflated type I error rates for analysis of binary traits in related samples even when case-control ratios are extremely unbalanced through simulation and GWASs in the UK Biobank data of 408,961 white British samples. Our work provides an accurate and scalable solution for large scale biobank data analysis. We employed SAIGE to analyze 1,403 binary phenotypes constructed based on PheCodes for 408,961 white British samples in the UK Biobank, within one week. The GWAS summary statistics are available for public download.

5.2 Large-scale PheWAS: Promises and Challenges

PheWAS was initially proposed to examine associations between genetic variants and multiple human phenotypes. Since the first PheWAS published in 2010 (Denny, et al., 2011), multiple PheWAS studies have been reported(Bush, et al., 2016; Dumitrescu, et al., 2015; Ehm, et al., 2017; Hall, et al., 2014; Hebbring, et al., 2015; Millard, et al., 2015; Moore, et al., 2015; Namjou, et al., 2014; Namjou, et al., 2015; Pendergrass, et al., 2013; Ye, et al., 2015), most of which focused on a small set of genetic markers that were previously shown to be associated with certain phenotypes or have biological functions of interest.

In recent years, as large biobanks, such as the MGI Biobank(www.michigangenomics.org), the HUNT study (Krokstad, et al., 2013) and the UK Biobank(Bycroft, et al., 2017; Sudlow, et al., 2015), started genotyping all of their participants and subsequently performing GWASs on the human phenome based on EHR and/or epidemiological data, PheWAS expanded the scope of its genetic tests to the whole genome. The release of the UK Biobank's genotyping and imputation data in combination with EHR and epidemiological questionnaire data for ~500,000 participants earlier this year is expected to invigorate the next-generation PheWAS(Bycroft, et al., 2017; Sudlow, et al., 2015).

Such genome-phenome association studies enable comprehensive mapping of the pleiotropy of human genome. This not only provides potential drug targets like traditional GWAS, but also reveals cross-phenotype associations which may lead to drug repositioning and/or treatment repurposing(Rastegar-Mojarad, et al., 2015). For example, Zidovudine, a reverse transcriptase inhibitor, is a drug prescribed to treat HIV/AIDS, which may inhibit TERT(Telomerase Reverse Transcriptase) activity. A PheWAS reported an association between a SNP rs2736100 in the gene *TERT* and diabetes and an increased telomerase activity was also reported to be associated with increased diabetes complications, suggesting that Zidovudine may be repositioned to treat diabetes (Rastegar-Mojarad, et al., 2015; Sun, et al., 2013).

Interpreting GWAS results is not a simple task, given that the majority of phenotype-associated variants are non-coding and therefore have no easily identifiable effect on protein function and disease. Omics data, such as epigenetic features, RNA expression and chromatin conformation, etc., are often integrated together for post-GWAS identification of candidate functional genes and pathways. Interpretation of PheWAS results is even more challenging, because cross-phenotype associations could be due to true pleiotropy, true comorbidity, confounded phenotype relationship or just false phenotype definition(Rastegar-Mojarad, et al., 2015). Further analysis, such as Mendelian randomization, is needed to help interpret PheWAS results.

The "big data" in large biobanks makes PheWAS computationally challenging. For example, SAIGE is the only mixed model method that is practical for large-scale PheWAS while correcting for sample relatedness and case-control imbalance. While SAIGE is relatively fast, performing PheWAS for 1,403 binary traits on the UK Biobank data(Bycroft, et al., 2017; Sudlow, et al., 2015), which contains ~30 million genetic variants for ~400,000 samples, required ~700,000 CPU hours (80 CPU years). Further efforts will be required to continuously improve the computational efficiency of methods for the association tests.

The next-generation PheWAS, in which multiple GWASs are performed, identifies genetic markers, while incorporating real-time health records and personal environmental information (e.g. lifestyle). This makes it a powerful approach to inform the prediction for individual disease risk and treatment response, aiding the development of precision medicine. The work presented in this dissertation highlights the value of GWAS for identifying novel genetic variants for complex diseases, even for diseases whose diagnoses is so difficult that sample size in GWAS is relatively small, proposes an optimal strategy to increase the study power by imputing genotypes using multiple reference panels, and provides a scalable and efficient statistical tool to perform PheWAS in large biobanks.

Bibliography

The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;45(6):580-585.

W. NHLBI GO Exome Sequencing Project (ESP) Seattle. In.; 2013.

Abecasis, G.R.*, et al.* Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 2001;30:97.

Abney, M., Ober, C. and McPeek, M.S. Quantitative-Trait Homozygosity and Association Mapping and Empirical Genomewide Significance in Large, Complex Pedigrees: Fasting Serum-Insulin Level in the Hutterites. *American Journal of Human Genetics* 2002;70(4):920-934.

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen‐2. *Current protocols in human genetics* 2013:7.20. 21-27.20. 41.

Albert, A. and Anderson, J.A. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* 1984;71(1):1-10.

Amberger, J.S.*, et al.* OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research* 2015;43(Database issue):D789-798.

Ang, Y.S.*, et al.* Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell* 2016;167(7):1734-1749.e1722.

Aulchenko, Y.S., de Koning, D.-J. and Haley, C. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. *Genetics* 2007;177(1):577.

Auton, A.*, et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.

Bahcall, O.G. Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals. *Nat Rev Genet* 2015;16(7):375.

Bailey, J.N.*, et al.* Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nat Genet* 2016;48(2):189-194.

Bernstein, B.E.*, et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.

Bonachea, E.M.*, et al.* Rare GATA5 sequence variants identified in individuals with bicuspid aortic valve. *Pediatr Res* 2014;76(2):211-216.

Boyle, A.P.*, et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;22(9):1790-1797.

Bradbury, P.J.*, et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;23(19):2633-2635.

Breslow, N. Whither PQL? In: Lin, D.Y. and Heagerty, P.J., editors, *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*. New York, NY: Springer New York; 2004. p. 1-22.

Breslow, N.E. and Clayton, D.G. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1993;88(421):9-25.

Broderick, P.*, et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature Genetics* 2007;39:1315.

Browning, B.L. and Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* 2009;84(2):210-223.
Browning, B.L. and Browning, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 2013;194(2):459-471.
Browning, Brian L. and Browning, Sharon R. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics* 2016;98(1):116-126.
Browning, S.R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics* 2008;124(5):439-450.
Bulik-Sullivan, B.K.*, et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 2015;47:291.
Burdon, K.P.*, et al.* Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nat Genet* 2011;43(6):574-578.
Burton, P.R.*, et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661-678.
Bush, W.S., Oetjens, M.T. and Crawford, D.C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet* 2016;17(3):129-145.
Bycroft, C.*, et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 2017.
Capanu, M., Gonen, M. and Begg, C.B. An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Stat Med* 2013;32(26):4550-4566.
CARDIoGRAMplusC4D Consortium*, et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013;45(1):25-33.
Cargill, M.*, et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* 1999;22:231.
Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genetics* 1999;21:56.
Chen, H.*, et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* 2016;98(4):653-666.
Chen, W.-M. and Abecasis, Gonçalo R. Family-Based Association Tests for Genomewide Association Scans. *American Journal of Human Genetics* 2007;81(5):913-926.
Chen, Y.*, et al.* Common variants near ABCA1 and in PMM2 are associated with primary open-angle glaucoma. *Nat Genet* 2014;46(10):1115-1119.
Cheng, T.H. and Thompson, D.J. Five endometrial cancer risk loci identified through genome-wide association analysis. 2016;48(6):667-674.
Clayton, D.G.*, et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* 2005;37:1243.
Cohen, J.C.*, et al.* Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol. *Science* 2004;305(5685):869.
Cooper, J.D.*, et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 2008;40(12):1399-1401.
Cordell, H.J.*, et al.* Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Hum Mol Genet* 2013;22(7):1473-1481.
Cripe, L.*, et al.* Bicuspid aortic valve is heritable. *J Am Coll Cardiol* 2004;44(1):138-143.

Cronin, R.M., *et al.* Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Frontiers in genetics* 2014;5:250.

Daniels, H.E. Saddlepoint Approximations in Statistics. *Ann. Math. Statist.* 1954;25(4):631-650.

Das, S., *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* 2016;48:1284.

De Jager, P.L., *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 2009;41(7):776-782.

de Lange, F.J., *et al.* Lineage and morphogenetic analysis of the cardiac valves. *Circ Res* 2004;95(6):645-654.

de Villemereuil, P., *et al.* General Methods for Evolutionary Quantitative Genetic Inference from Generalized Mixed Models. *Genetics* 2016;204(3):1281-1294.

Deelen, P., *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. 2014;22(11):1321-1326.

Delaneau, O., Marchini, J. and Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2012;9(2):179-181.

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

Denny, J.C., *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* 2013;31:1102.

Denny, J.C., *et al.* Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011;89(4):529-542.

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

Dey, R., *et al.* A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *The American Journal of Human Genetics* 2017;101(1):37-49.

Dumitrescu, L., *et al.* Towards a phenome-wide catalog of human clinical traits impacted by genetic ancestry. *BioData Min* 2015;8:35.

Ehm, M.G., *et al.* Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. *PLOS ONE* 2017;12(11):e0186405.

Eisenhart, C. The Assumptions Underlying the Analysis of Variance. *Biometrics* 1947;3(1):1-21.

Ellison, J.W., *et al.* Evidence of genetic locus heterogeneity for familial bicuspid aortic valve. *The Journal of surgical research* 2007;142(1):28-31.

Emanuel, R., *et al.* Congenitally bicuspid aortic valves. Clinicogenetic study of 41 families. *Br Heart J* 1978;40(12):1402-1407.

Ernst, J. and Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9(3):215-216.

Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80(1):27-38.

Fisher, R.A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 1919;52(2):399-433.

Foffa, I., *et al.* Sequencing of NOTCH1, GATA5, TGFBR1 and TGFBR2 genes in familial cases of bicuspid aortic valve. *BMC Med Genet* 2013;14:44.

Frazer, K.A., *et al.* Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 2009;10:241.

Fritsche, L.G.*, et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics* 2015;48:134.

Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* 2015;31(5):782-784.

G., M. Versuche über Pflanzen-Hybriden. Verhandlungen des Naturforschenden Vereines. *Abhandlungern* 1866; Brünn(4):45.

Garg, V. Molecular genetics of aortic valve disease. *Curr Opin Cardiol* 2006;21(3):180-184.

Garg, V.*, et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* 2003;424(6947):443-447.

Ge, Y.*, et al.* Rare variants in BRCA2 and CHEK2 are associated with the risk of urinary tract cancers. *Scientific reports* 2016;6:33542.

Gharahkhani, P.*, et al.* Common variants near ABCA1, AFAP1 and GMDS confer risk of primary open-angle glaucoma. *Nat Genet* 2014;46(10):1120-1125.

Gilmour, A.R., Thompson, R. and Cullis, B.R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 1995;51(4):1440-1450.

Golan, D., Lander, E.S. and Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* 2014;111(49):E5272.

Haiman, C.A.*, et al.* A common genetic risk factor for colorectal and prostate cancer. *Nature Genetics* 2007;39:954.

Hall, M.A.*, et al.* Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS Genet* 2014;10(12):e1004678.

Hebbring, S.J.*, et al.* Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics* 2015;31(12):1981-1987.

Hebbring, S.J.*, et al.* A PheWAS approach in studying HLA-DRB1*1501. *Genes and immunity* 2013;14(3):187-191.

Heinze, G. and Schemper, M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002;21(16):2409-2419.

Henderson, C.R. Applications of linear models in animal breeding/por Charles R Henderson. 1984.

Hestenes, M.R.a.S., Eduard. Methods of conjugate gradients for solving linear systems. NBS; 1952.

Hirayama-Yamada, K.*, et al.* Phenotypes with GATA4 or NKX2.5 mutations in familial atrial septal defect. *Am J Med Genet A* 2005;135(1):47-52.

Hirschhorn, J.N. and Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 2005;6:95.

Hoffman, J.I. and Kaplan, S. The incidence of congenital heart disease. *J Am Coll Cardiol* 2002;39(12):1890-1900.

Holm, H.*, et al.*

Horikoshi, M.*, et al.* Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLoS Genet* 2015;11(7):e1005230.

Houlston, R.S.*, et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008;40(12):1426-1435.

Howie, B.*, et al.* Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 2012;44:955.

Howie, B.N., Donnelly, P. and Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* 2009;5(6):e1000529.

Huang, G.H. and Tseng, Y.C. Genotype imputation accuracy with different reference panels in admixed populations. *BMC proceedings* 2014;8(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo):S64.

Huang, J.*, et al.* 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet* 2012;20(7):801-805.

Huang, J.*, et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. 2015;6:8111.

Huang, L.*, et al.* Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* 2009;84(2):235-250.

Huang, W.Y., Cukerman, E. and Liew, C.C. Identification of a GATA motif in the cardiac alpha-myosin heavy-chain-encoding gene and isolation of a human GATA-4 cDNA. *Gene* 1995;155(2):219-223.

Huang, X.*, et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* 2010;42:961.

Imhof, J.P. Computing the Distribution of Quadratic Forms in Normal Variables. *Biometrika Trust* 1961;48:8.

International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860.

Jaeger, E.*, et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature Genetics* 2007;40:26.

Jiao, J.*, et al.* Promoting reprogramming by FGF2 reveals that the extracellular matrix is a barrier for reprogramming fibroblasts to pluripotency. *Stem cells (Dayton, Ohio)* 2013;31(4):729-740.

Jiao, J.*, et al.* Modeling Dravet syndrome using induced pluripotent stem cells (iPSCs) and directly converted neurons. *Hum Mol Genet* 2013;22(21):4241-4252.

Jin, Y.*, et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. 2016.

Johnson & Kotz, p. Distributions in Statistics: Continuous Univariate Distributions. New York: Wiley; 1970.

Jun, G.*, et al.* An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* 2015;25(6):918-925.

Kaasschieter, E.F. Preconditioned conjugate gradients for solving singular systems. *Journal of Computational and Applied Mathematics* 1988;24(1):265-275.

Kang, H.M.*, et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 2010;42:348.

Kang, H.M.*, et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* 2008;178(3):1709.

Kathiresan, S. A PCSK9 missense variant associated with a reduced risk of early-onset myocardial infarction. *N Engl J Med* 2008;358(21):2299-2300.

Kathiresan, S.*, et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics* 2009;41(3):334-341.

Kennedy, S.J.*, et al.* Inherited duplication, dup (8) (p23.1p23.1) pat, in a father and daughter with congenital heart defects. *American journal of medical genetics* 2001;104(1):79-80.

Kircher, M., *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46(3):310-315.

Krokstad, S., *et al.* Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol* 2013;42(4):968-977.

Kumar, P., Henikoff, S. and Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4(7):1073-1081.

Kuo, C.T., *et al.* GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes & development* 1997;11(8):1048-1060.

Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 1999;4(86):7.

Laforest, B., Andelfinger, G. and Nemer, M. Loss of Gata5 in mice leads to bicuspid aortic valve. *J Clin Invest* 2011;121(7):2876-2887.

LaHaye, S., *et al.* Utilization of Whole Exome Sequencing to Identify Causative Mutations in Familial Congenital Heart Disease. *Circ Cardiovasc Genet* 2016;9(4):320-329.

Lander, E.S. The New Genomics: Global Views of Biology. *Science* 1996;274(5287):536.

Lander, E.S. and Schork, N.J. Genetic dissection of complex traits. *Science* 1994;265(5181):2037.

Lander, E.S. and Schork, N.J. Genetic Dissection of Complex Traits. *FOCUS* 2006;4(3):442-458.

Lane, J.M., Vlasac, I. and Anderson, S.G. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. 2016;7:10889.

Lek, M., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285-291.

Lek, M., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285-291.

Li, N. and Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* 2003;165(4):2213.

Li, Y., *et al.* Genotype Imputation. *Annual Review of Genomics and Human Genetics* 2009;10(1):387-406.

Li, Y., *et al.* MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* 2010;34(8):816-834.

Liao, K.P., *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis and rheumatism* 2013;65(3):571-581.

Lin, C.J., *et al.* Partitioning the heart: mechanisms of cardiac septation and valve development. *Development (Cambridge, England)* 2012;139(18):3277-3299.

Lincoln, J., Alfieri, C.M. and Yutzey, K.E. Development of heart valve leaflets and supporting apparatus in chicken and mouse embryos. *Developmental dynamics : an official publication of the American Association of Anatomists* 2004;230(2):239-250.

Lippert, C., *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* 2011;8:833.

Listgarten, J., *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* 2012;9:525.

Loh, P.-R., *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* 2015;47:284.

Long, T.*, et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature Genetics* 2017;49:568.

Loos, R.J.*, et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 2008;40(6):768-775.

Losenno, K.L., Goodman, R.L. and Chu, M.W. Bicuspid aortic valve disease and ascending aortic aneurysms: gaps in knowledge. *Cardiology research and practice* 2012;2012:145202.

Lourenco, D.*, et al.* Loss-of-function mutation in GATA4 causes anomalies of human testicular development. *Proc Natl Acad Sci U S A* 2011;108(4):1597-1602.

Low-Kam, C.*, et al.* Whole-genome sequencing in French Canadians from Quebec. *Hum Genet* 2016;135(11):1213-1221.

Ma, C.*, et al.* Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genetic Epidemiology* 2013;37(6):539-550.

MacArthur, J.*, et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* 2017;45(Database issue):D896-D901.

Mahajan, A.*, et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 2014;46(3):234-244.

Manolio, T.A.*, et al.* Finding the missing heritability of complex diseases. *Nature* 2009;461:747.

Marchini, J. and Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* 2010;11(7):499-511.

Marchini, J. and Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 2010;11:499.

Marchini, J.*, et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 2007;39:906.

Marchini, J.*, et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39(7):906-913.

Marouli, E.*, et al.* Rare and low-frequency coding variants alter human adult height. *Nature* 2017;542:186.

Martin, L.J.*, et al.* Evidence in favor of linkage to human chromosomal regions 18q, 5q and 13q for bicuspid aortic valve and associated cardiovascular malformations. *Hum Genet* 2007;121(2):275-284.

Mattapally, S.*, et al.* c.620C>T mutation in GATA4 is associated with congenital heart disease in South India. *BMC Med Genet* 2015;16:7.

McBride, K.L.*, et al.* NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling. *Hum Mol Genet* 2008;17(18):2886-2893.

McCarthy, S.*, et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* 2016.

Michelena, H.I.*, et al.* Natural history of asymptomatic patients with normally functioning or minimally dysfunctional bicuspid aortic valve in the community. *Circulation* 2008;117(21):2776-2784.

Michelena, H.I.*, et al.* Incidence of aortic complications in patients with bicuspid aortic valves. *JAMA* 2011;306(10):1104-1112.

Millard, L.A.*, et al.* MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci Rep* 2015;5:16645.

Mitt, M.*, et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 2017;25(7):869-876.

Molkentin, J.D., *et al.* Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes & development* 1997;11(8):1061-1072.
Moore, C.B., *et al.* Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. *Open forum infectious diseases* 2015;2(1):ofu113.
Nalls, M.A., *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* 2014;46(9):989-993.
Namjou, B., *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Frontiers in genetics* 2014;5:401.
Namjou, B., *et al.* A GWAS Study on Liver Function Test Using eMERGE Network Participants. *PLoS One* 2015;10(9):e0138677.
Nelis, M., *et al.* Genetic structure of Europeans: a view from the North-East. *PLoS One* 2009;4(5):e5472.
Neuraz, A., *et al.* Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS computational biology* 2013;9(12):e1003405.
Nikpay, M., *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;47(10):1121-1130.
Okada, Y., *et al.* Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. 2015;47(7):798-802.
Orho-Melander, M., *et al.* Common Missense Variant in the Glucokinase Regulatory Protein Gene Is Associated With Increased Plasma Triglyceride and C-Reactive Protein but Lower Fasting Glucose Concentrations. *Diabetes* 2008;57(11):3112.
Padang, R., *et al.* Rare non-synonymous variations in the transcriptional activation domains of GATA5 in bicuspid aortic valve disease. *Journal of molecular and cellular cardiology* 2012;53(2):277-281.
Pasta, S., *et al.* Difference in hemodynamic and wall stress of ascending thoracic aortic aneurysms with bicuspid and tricuspid aortic valve. *Journal of biomechanics* 2013;46(10):1729-1738.
Patterson, N., Price, A.L. and Reich, D. Population Structure and Eigenanalysis. *PLOS Genetics* 2006;2(12):e190.
Pe'er, I., *et al.* Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology* 2008;32(4):381-385.
Pehlivan, T., *et al.* GATA4 haploinsufficiency in patients with interstitial deletion of chromosome region 8p23.1 and congenital heart disease. *American journal of medical genetics* 1999;83(3):201-206.
Pendergrass, S.A., *et al.* Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013;9(1):e1003087.
Pereira, J.S., *et al.* Identification of a novel germline FOXE1 variant in patients with familial non-medullary thyroid carcinoma (FNMTC). *Endocrine* 2015;49(1):204-214.
Phanstiel, D.H., *et al.* Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* 2015;31(19):3092-3098.
Piper, J., *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research* 2013;41(21):e201.

Pistis, G.*, et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* 2015;23(7):975-983.

Plomin, R., Haworth, C.M.A. and Davis, O.S.P. Common disorders are quantitative traits. *Nature Reviews Genetics* 2009;10:872.

Posch, M.G.*, et al.* Mutations in GATA4, NKX2.5, CRELD1, and BMP4 are infrequently found in patients with congenital cardiac septal defects. *Am J Med Genet A* 2008;146a(2):251-253.

Price, A.L., Spencer, C.C.A. and Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences* 2015;282(1821).

Price, A.L.*, et al.* New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 2010;11:459.

Pu, W.T.*, et al.* GATA4 is a dosage-sensitive regulator of cardiac morphogenesis. *Developmental biology* 2004;275(1):235-244.

Purcell, S.*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81(3):559-575.

Ran, F.A.*, et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 2013;8(11):2281-2308.

Rao, S.S.*, et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;159(7):1665-1680.

Rastegar-Mojarad, M.*, et al.* Opportunities for drug repositioning from phenome-wide association studies. *Nature Biotechnology* 2015;33:342.

Reich, D.E. and Lander, E.S. On the allelic spectrum of human disease. *Trends in Genetics*;17(9):502-510.

Reiner, A.P.*, et al.* &lt;span hwp:id=&quot;article-title-1&quot; class=&quot;article-title&quot;&gt;Common Coding Variants of the &lt;em&gt;HNF1A&lt;/em&gt; Gene Are Associated With Multiple Cardiovascular Risk Phenotypes in Community-Based Samples of Younger and Older European-American Adults&lt;/span&gt;&lt;span hwp:id=&quot;article-title-2&quot; class=&quot;sub-article-title&quot;&gt;CLINICAL PERSPECTIVE&lt;/span&gt;. *Circulation: Cardiovascular Genetics* 2009;2(3):244.

Ritchie, M.D.*, et al.* Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013;127(13):1377-1385.

Rivera-Feliciano, J.*, et al.* Development of heart valves requires Gata4 expression in endothelial-derived cells. *Development (Cambridge, England)* 2006;133(18):3607-3618.

Roberts, W.C. and Ko, J.M. Frequency by decades of unicuspid, bicuspid, and tricuspid aortic valves in adults having isolated aortic valve replacement for aortic stenosis, with or without associated aortic regurgitation. *Circulation* 2005;111(7):920-925.

Roshyara, N.R. and Scholz, M. Impact of genetic similarity on imputation accuracy. *BMC Genet* 2015;16:90.

Ruth, K.S.*, et al.* Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *Eur J Hum Genet* 2015.

Saint Pierre, A. and Génin, E. How important are rare variants in common disease? *Briefings in Functional Genomics* 2014;13(5):353-361.

Sarkozy, A.*, et al.* Spectrum of atrial septal defects associated with mutations of NKX2.5 and GATA4 transcription factors. *J Med Genet* 2005;42(2):e16.

Scheet, P. and Stephens, M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* 2006;78(4):629-644.
Schork, N.J. Genetics of Complex Disease. *American Journal of Respiratory and Critical Care Medicine* 1997;156(4):S103-S109.
Sebastiani, P*., et al.* Four Genome-Wide Association Studies Identify New Extreme Longevity Variants. *The Journals of Gerontology: Series A* 2017;72(11):1453-1464.
Sham, P.C. and Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* 2014;15:335.
Shameer, K*., et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2014;133(1):95-109.
Sherry, S.T*., et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 2001;29(1):308-311.
Siu, S.C. and Silversides, C.K. Bicuspid aortic valve disease. *J Am Coll Cardiol* 2010;55(25):2789-2800.
Spencer, C.C.A*., et al.* Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLOS Genetics* 2009;5(5):e1000477.
Springer-Verlag, J.I.T. Principal component analysis. NEW YORK; 1986.
Stone, E.M*., et al.* Identification of a gene that causes primary open angle glaucoma. *Science* 1997;275(5300):668-670.
Su, R.J*., et al.* Efficient generation of integration-free ips cells from human adult peripheral blood using BCL-XL together with Yamanaka factors. *PLoS One* 2013;8(5):e64496.
Sudlow, C*., et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 2015;12(3):e1001779.
Sun, X*., et al.* The effect of telomerase activity on vascular smooth muscle cell proliferation in type 2 diabetes in vivo and in vitro. *Molecular medicine reports* 2013;7(5):1636-1640.
Sveinbjornsson, G*., et al.* Rare mutations associating with serum creatinine and chronic kidney disease. *Hum Mol Genet* 2014;23(25):6935-6943.
Svishcheva, G.R*., et al.* Rapid variance components–based method for whole-genome association analysis. *Nature Genetics* 2012;44:1166.
Teslovich, T.M*., et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466(7307):707-713.
The Genomes Project, C. A global reference for human genetic variation. *Nature* 2015;526:68.
The International HapMap, C. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851.
Tomita-Mitchell, A*., et al.* GATA4 sequence variants in patients with congenital heart disease. *J Med Genet* 2007;44(12):779-783.
Tutar, E*., et al.* The prevalence of bicuspid aortic valve in newborns by echocardiographic screening. *American heart journal* 2005;150(3):513-515.
Vaisse, C*., et al.* Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbid obesity. *The Journal of Clinical Investigation* 2000;106(2):253-262.
van Leeuwen, E.M*., et al.* Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *J Med Genet* 2016;53(7):441-449.

Verweij, N., *et al.* Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Scientific Reports* 2017;7(1):2761.
Walter, K., *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 2015;526(7571):82-90.
Wang, E., *et al.* Identification of functional mutations in GATA4 in patients with congenital heart disease. *PLoS One* 2013;8(4):e62138.
Wang, J., *et al.* A novel GATA4 mutation responsible for congenital ventricular septal defects. *Int J Mol Med* 2011;28(4):557-564.
Ward, C. Clinical significance of the bicuspid aortic valve. *Heart* 2000;83(1):81-85.
Ward, L.D. and Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* 2012;40(Database issue):D930-934.
Weeks, D.E. and Lathrop, G.M. Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics* 1995;11(12):513-519.
Wirrig, E.E. and Yutzey, K.E. Conserved transcriptional regulatory mechanisms in aortic valve development and disease. *Arterioscler Thromb Vasc Biol* 2014;34(4):737-741.
Wooten, E.C., *et al.* Application of Gene Network Analysis Techniques Identifies AXIN1/PDIA2 and Endoglin Haplotypes Associated with Bicuspid Aortic Valve. *PLoS One* 2010;5(1).
Yang, J., *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88(1):76-82.
Yang, J., *et al.* Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* 2014;46:100.
Yang, Y.Q., *et al.* GATA4 loss-of-function mutations underlie familial tetralogy of fallot. *Hum Mutat* 2013;34(12):1662-1671.
Yang, Y.Q., *et al.* Mutation spectrum of GATA4 associated with congenital atrial septal defects. *Archives of medical science : AMS* 2013;9(6):976-983.
Yang, Y.Q., *et al.* Novel GATA4 mutations in patients with congenital ventricular septal defects. *Medical science monitor : international medical journal of experimental and clinical research* 2012;18(6):CR344-350.
Ye, Z., *et al.* Phenome-wide association studies (PheWASs) for functional variants. *Eur J Hum Genet* 2015;23(4):523-529.
Yu, J., *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 2005;38:203.
Zeggini, E., *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008;40(5):638-645.
Zeggini, E., *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;316(5829):1336-1341.
Zhang, W., *et al.* GATA4 mutations in 486 Chinese patients with congenital heart disease. *European journal of medical genetics* 2008;51(6):527-535.
Zhang, Z., *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 2010;42:355.
Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 2012;44:821.