Distribution Agreement

Signature:

S. Taylor Head Date

Innovative methods for investigating the genetic architecture of complex human traits

By

S. Taylor Head

Michael P. Epstein, Ph.D.

David J. Cutler, Ph.D.

Zhaohui Qin, Ph.D.

Joellen Schildkraut, Ph.D.

Jingjing Yang, Ph.D.

Accepted:

Kimberly J. Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Innovative methods for investigating the genetic architecture of complex human traits

By

S. Taylor Head
BS, Georgia Institute of Technology, 2013
MSPH, Emory University, 2017

Advisor: Michael P. Epstein, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2024

Abstract

Innovative methods for investigating the genetic architecture of complex human traits

Over the past two decades, there has been a rapid increase in the amount of publicly-available genetic datasets necessary to advance gene mapping of complex human traits and diseases. As the amount of genome-wide association (GWAS) data has grown, so has the need for novel statistical methods that aim to not only locate risk regions across the genome, but also shed light on the mechanisms by which these risk loci exert their effect on traits of interest. In this dissertation, we develop and apply innovative statistical methods to help fill these important gaps in such knowledge.

In the first project, we develop a population-based test for parent-of-origin effects (POEs) leveraging GWAS data on multiple phenotypes. A POE exists when maternally- and paternally-transmitted alleles exhibit differential effects on phenotype expression. We show that the presence of a POE at a given locus induces a difference in the covariance structure among multiple phenotypes between homozygotes and heterozygotes. Based on a robust omnibus test for homogeneity of covariance matrices, our method can be applied to normal and non-normal phenotypes and can easily adjust for population stratification and other non-genetic confounders. We evaluate our method through simulation studies and apply it to GWAS data of BMI and two cholesterol phenotypes from the UK Biobank, identifying 338 genome-wide significant variants.

In the second project, we apply a recently proposed transcriptome-wide association study (TWAS) method to publicly available summary statistic GWAS data for breast and ovarian cancer. This Bayesian genome-wide method (BGW-TWAS) incorporates both cis- and trans-expression quantitative trait loci (eQTLs). We first train gene expression imputation models using GTEx V8 transcriptomic data separately in breast and ovarian tissue. We then identify genes significantly associated with risk of both cancers and 10 common subtypes of these cancers and investigate the eQTL architecture of these top genes. We show that several novel loci are identified driven primarily by trans-eQTL effects. We replicate several associations using independent GWAS data and expression data in tumor and tumor-adjacent breast tissue from the Cancer Genome Atlas.

In the third project, we expand upon a recent method for TWAS that circumvents the need for individual-level genotype and transcriptomic data. This method leverages summary-level eQTL data and polygenic risk score (PRS) models to impute gene expression in individuals of a given ancestral group. In contrast to ancestrally homogenous populations, recently admixed populations have genomes that are a mosaic of distinct local ancestral (LA) segments, and it is well-known that PRS methods port very poorly across ancestral groups. Motivated by this, we propose a method to perform TWAS with summary-level eQTL data in recently admixed subjects. We compare the imputation accuracy, power, and type I error rate of this LA-aware approach to LA-unaware PRS methods. We apply our method to 29 blood biochemistry phenotypes in two-way African/European admixed individuals in the UK Biobank.

Innovative methods for investigating the genetic architecture of complex human traits

By

S. Taylor Head
BS, Georgia Institute of Technology, 2013
MSPH, Emory University, 2017

Advisor: Michael P. Epstein, Ph.D.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The desire to understand the role that genetic factors play in the manifestation of human traits has spurred great advancements in recent years. These advancements lie not only in the development of technology to accurately and affordably sequence large numbers of individuals, but also in the development of statistical methodologies that are appropriate to analyze this unique class of data. Statistical genetics lies at the intersection of human genetics and biostatistics, and research in this field has helped to augment our knowledge of how variation in traits within or across populations relates to variation among individuals and populations at the genetic level. Through the advent and prolific application of genome-wide association studies (GWAS), researchers have identified thousands of genetic polymorphisms associated with complex traits [1] and, in turn, great progress has been made in the clinical translation of these findings to improve public health.

However, due to the population-based nature of most GWAS cohorts, as opposed to family-based cohorts, it is implicitly assumed in analysis that the effect of a given allele on the phenotype under study is independent of the (unknown) parental ancestry of the inherited allele. This assumption is violated in the presence of parent-of-origin effects (POEs). A POE occurs when the effect of a maternally-inherited allele on expression of a trait differs from the effect of the paternally-inherited copy of the same allele [2]. Most existing methods that are used to detect POEs currently are limited by the requirement of familial genotype data, often of modest sample size, to determine maternal or paternal transmission of alleles in offspring [3–12]. One method proposed for GWAS-scale cohorts that does not require paternal genetics still requires a genealogy database containing data from more distant relatives to impute parental ancestry of haplotypes of the subjects under study [13]. Another recent POE method that does not require any familial genetic data is limited to the analysis of a single quantitative trait and does not leverage the pleiotropic nature of many genes [14]. In the second chapter of this dissertation, we describe our development of a powerful statistical method for detecting loci harboring POEs in samples of unrelated individuals that accommodates multiple phenotypes jointly.

Another challenge faced by researchers while interpreting GWAS results is that the vast majority of GWAS-identified risk variants fall in non-protein coding regions of the genome and thus lack an obvious mechanistic explanation by way of a direct effect on protein structure [15]. This realization has motivated considerable methodological and applied research in the field of transcriptome-wide association studies (TWAS). These studies aim to estimate the association between disease risk and genetically-regulated transcriptional activity and therefore improve our understanding of how the effects of risk variants are mediated by gene expression. While the catalog of susceptibility genes identified by TWAS for a wide range of complex traits is growing, further knowledge can be gained by exploring the trans-regulatory effects of common variants on gene expression [16, 17]. Further, most TWAS methods published to date (1) require individual-level transcriptomic and genetic data to train statistical models of gene expression in a reference cohort, which may be difficult to obtain and can be of limited sample size, and (2) are dedicated exclusively to ancestrally homogenous, non-admixed populations. In the next two chapters of this dissertation, we describe two projects that focus on the application and development, respectively, of innovative statistical methods for TWAS that help address these issues.

1.2 Outline of Research

In Chapter 2 of this dissertation, we describe our first project, a method we have termed POIROT (Parent-of-Origin Inference using Robust Omnibus Test) [18]. It is a powerful statistical test for detecting POEs in population-based samples of unrelated individuals that leverages multiple phenotypes simultaneously. In Section 2.1, we begin with a brief introduction to the biological phenomena giving rise to POEs in nature, describe known POE-trait associations, and describe existing methods currently employed for detecting these effects in the settings of both familial and population-based genotype data. We highlight the limitations of these approaches that motivate our proposed research. In Section 2.2, we define our statistical model for quantitative traits exhibiting POEs, outline our method for testing whether a POE exists at a given locus for one or more quantitative traits, and present the framework for our simulation and applied analyses. In Section 2.3, we describe the power and type I error of POIROT from performed simulation studies and the results from the application of our method to real-world data on three phenotypes from the UK Biobank (BMI, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol). We compare the performance of our method in both simulations and our applied analysis to a competing univariate method for detecting POEs in unrelated samples that does not utilize data across multiple phenotypes. We conclude this chapter with a discussion (Section 2.4) of the limitations of this work.

In Chapter 3 of this dissertation, we describe an applied TWAS analysis [19].

This work leverages expression quantitative trait loci (eQTL) information from both the variants located in close proximity to a given gene (cis-variants) and those located distal to the gene, often one megabase (Mb) or further on the same chromosome or on a separate chromosome (trans-variants). By first estimating the association of these genome-wide variants with expression using a reference panel from the Genotype-Tissue Expression (GTEx) project [20], we were able to detect genes whose genetically-regulated component of gene expression (GReX) in breast and ovarian tissue are associated with risk of overall breast and ovarian cancer. We also examine this gene-level association of GReX with five common histological subtypes of each of these cancers. Of note, we identify a subset of highly significant genes not previously detected in GWAS or cis-only TWAS of these cancers whose associations appear to be driven by strong trans-eQTL effects. In Section 3.1, we provide an introduction to the framework behind TWAS, discuss the limitations of TWAS that have been performed for breast and ovarian cancer thus far, and describe the recently published computationally-tractable statistical method for performing TWAS using both cis- and trans-eQTLs [21]. In Section 3.2, we describe the data we obtained for performing this analysis, including whole-genome sequencing, transcriptomic, and GWAS summary statistic data, as well as outline the methodology behind the method we apply. We also describe our extensive set of validation analyses performed to investigate how our putatively novel cancer-associated genes replicate using data from studies independent of our main analyses. In Section 3.3, we describe our estimated GReX models in both tissues and their corresponding eQTL architecture. We describe the findings from our TWAS of all 12 cancer phenotypes in detail and relate these to the results of all validation analyses. In Section 3.4, we conclude with an extensive discussion of the limitations of this work and compare our findings to genes previously implicated in other applied TWAS of breast and ovarian cancer.

In Chapter 4 of this dissertation, we describe our work on a statistical method for enhanced TWAS analysis in individuals of admixed ancestry, a group that has been largely underrepresented across the entire spectrum of genetic association studies. Our proposed method builds upon a recently published approach to TWAS that utilizes summary-level eQTL reference datasets and polygenic risk score (PRS) approaches to train GReX models [22]. We leverage local ancestry (LA) information within haplotypes to allow for possible LA-dependent eQTL architecture. In Section 4.1, we provide an overview of how PRS models can be used to perform TWAS, how performance of these PRS models varies greatly across different ancestral groups, and how LA deconvolution in recently admixed TWAS subjects could be used to yield improved estimates of the association between imputed gene expression and a phenotype of interest. In Section 4.2, we provide a statistical introduction to how we explicitly model gene expression in admixed subjects and our proposed TWAS method. We then describe our simulation analyses performed to evaluate the GReX imputation accuracy of our method, as well the power and type I error of the downstream gene-phenotype association test. We also discuss our applied analysis using real-world data from two-way African/European admixed individuals in the UK Biobank and phenotype data on 29 blood biomarker traits. In Section 4.3, we discuss the results of these analyses. We conclude this chapter in Section 4.4 with a discussion of the limitations of our proposed work.

In addition to the sections outlined above, each subsequent chapter includes an appendix. This appendix contains supplemental figures and tables not inlaid in the main sections, as well as any statistical proofs, where appropriate. Lastly, in Chapter 5, we briefly summarize a few possible extensions for each of the projects presented in this dissertation.

Chapter 2

Topic 1. POIROT: A powerful test for parent-of-origin effects in unrelated samples leveraging multiple phenotypes

2.1 Introduction

Most genome-wide association studies (GWAS) implicitly assume the magnitude and direction of effect of a genetic variant on expression of a phenotype is independent of whether the variant was maternally or paternally inherited. However, there exists a distinct class of genetic variants for which this assumption is violated. Such variants harbor a parent-of-origin effect (POE) wherein the effect of an allele on a trait depends on whether it was transmitted from the mother or the father [2]. A substantial proportion of the variation in complex traits is not explained by the additive effects of common single nucleotide polymorphisms (SNPs) across the genome. POEs may represent an important contribution to this missing heritability [23].

There are multiple cited biological mechanisms by which POEs can arise in mammals. These include maternal intrauterine environment effects and effects of the maternal mitochondrial genome. However, the most frequently considered mechanism is genomic imprinting [24]. This epigenetic phenomenon was formally discovered in the 1980s primarily through embryological experiments [25]. In imprinting, the maternal and paternal alleles undergo differential epigenetic modifications that leads to parent-of-origin-specific transcription of the gene copies. Many imprinted genes tend to be found in clusters across the genome. Regulation of the expression of these clustered genes is under control of an imprinting control region (ICR), the mechanisms of which are complex [26]. These ICR are often characterized by repetitive sequences and located near imprinted genes. It is estimated that only approximately 1% of mammalian genes are subject to imprinting. However, there has been growing evidence for the existence of POE variants for a wide range of hereditary traits [27]. Classic examples of POE-associated diseases include Prader-Willi syndrome and Angelman syndrome. These diseases result from imprinted genes at 15q11-15q13 when only maternal or paternal copies are expressed, respectively [13, 14, 24, 28–34].

To detect variants demonstrating POEs, studies have historically required genotype data from related individuals to ascertain parental ancestry of the inherited alleles. In the case of available parent-offspring trio or other forms of familial genomes, there are well-established methods to detect POEs [3–12]. These approaches often test for a mean difference in allele effect based on grouping offspring by parent-of-origin of the allele. These mean-based tests are intuitive and optimally powered given sample size. Yet, the requirement of trio or more general family data severely limits this sample size in practice. This, in consequence, limits genome-wide discovery of the modest genetic effects that we anticipate for complex human traits.

Rather than rely on family studies of limited sample size to detect POEs, researchers have recently transitioned to detecting the phenomenon in GWAS-scale cohorts. This practice requires innovative statistical methods to deal with missing parental ancestry information. For example, Kong et al. inferred parental origin of alleles when parental genotype data are not available by first phasing Icelandic probands. For each of the proband haplotypes, they searched a genealogy database for the closest typed maternal and paternal relatives carrying that haplotype [13]. For each haplotype, they constructed a robust score comparing the meiotic distances between the proband and these two relatives to quantify the likelihood of maternal or paternal transmission and ultimately assign parental origin. While this approach solves the issue of small sample sizes, power is still impacted by the potential inaccuracy or uncertainty in haplotypic reconstruction.

More recently, Hoggart et al. described a novel statistical method for detecting POEs for a single quantitative trait using GWAS data of unrelated individuals [14]. The authors illustrated that the existence of a POE results in increased phenotypic variance among heterozygotes compared to homozygotes. They tested for this variance inflation using a robust version of the Brown-Forsythe test. The method successfully identified previously undocumented POE associations of two SNPs with body mass index (BMI). This work has enabled POE analysis in population studies of biobank scale.

A sizable proportion of genes in the GWAS catalog are pleiotropic [35]. These genes affect more than one biological process, in turn associating with multiple (correlated) phenotypes [36]. Analyzing the joint effects of a gene on more than one trait can often result in a marked increase in power over univariate approaches [37–39]. Importantly, well-established POEs in humans are usually the result of embryonic silencing of one parental allele. As this silencing generally occurs early in development, its effects are likely to present in all or nearly all tissues expressing the gene. When differential silencing of this gene affects multiple tissues, this can yield POEs for several distinct phenotypes. Joint analysis of multiple traits can leverage this potential pleiotropy to improve power over univariate variance-based POE tests while simultaneously reducing multiple testing burden of multiple phenotypes.

Here, we expand on the concept initially suggested by Hoggart et al. to develop a test for POEs in genetic studies of unrelated individuals that considers multiple quantitative phenotypes. We show that a pleiotropic POE variant will not only induce differences in the variance of POE traits between heterozygotes and homozygotes, but also in their corresponding covariances. In our method, POIROT (Parent-of-Origin Inference using Robust Omnibus Test), we test for equality of phenotypic covariances matrices between heterozygous and homozygous groups. Specifically, we use the robust omnibus (R-Omnibus) test [40] to accommodate highly skewed traits. We first provide background on the statistical construction of our test statistic using the R-Omnibus framework. Next, through simulations, we demonstrate that our proposed method properly controls type I error and achieves superior power compared to the univariate approach of Hoggart et al. We also introduce a post-hoc test that can help distinguish variants with POE effects from variants demonstrating more general gene-gene/gene-environment effects (which also induce patterns of trait variance/covariance that differ by genotype). We apply our method to GWAS data of BMI, HDL cholesterol, and LDL cholesterol from the UK Biobank and identify 338 significant potential POE loci. We conclude with a discussion of our findings, limitations, and proposed research to extend this work.

2.2 Methods

2.2.1 Phenotype Model

Using the notation of Hoggart et al., consider one bi-allelic SNP with reference allele “A” and alternative allele “B” [14]. Assume that we have collected n_{AA} individuals who have the homozygous AA genotype, n_{BB} individuals who have the homozygous

BB genotype, and n_{AB} individuals who are heterozygous at the SNP of interest. Further assume we have $K > 1$ continuous phenotypes on all subjects and that we have already adjusted these phenotypes for the effects of non-genetic confounders like principal components of ancestry.

We first model phenotypes in homozygous AA subjects. We can model the kth phenotype of the ith individual with this genotype ($y_{i,k}^{(AA)}$) as follows:

$$y_{i,k}^{(AA)} = \mu_k + \epsilon_{i,k}, \quad i = 1, ..., n_{AA}, \; k = 1, ..., K \tag{2.1}$$

Here, μ_k is the mean of the kth phenotype among AA homozygotes and $\epsilon_{i,k}$ is a random error term. Let $\boldsymbol{y_i^{(AA)}} = (y_{i,1}^{(AA)}, y_{i,2}^{(AA)}, ..., y_{i,K}^{(AA)})^\top \in \mathbb{R}^K$ be the vector of phenotypes for the ith AA individual. We can then model this vector as below:

$$\boldsymbol{y_i^{(AA)}} = \boldsymbol{\mu} + \boldsymbol{\epsilon_i}, \quad i = 1, ..., n_{AA} \tag{2.2}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)^\top$ is the $K \times 1$ vector of phenotype means in AA subjects and $\boldsymbol{\epsilon_i} = (\epsilon_{i,1}, ...\epsilon_{i,K})^\top$ is the $K \times 1$ vector of error terms. We assume that $\mathrm{E}[\boldsymbol{\epsilon_i}] = \boldsymbol{0_K}$ and $\mathrm{Cov}[\boldsymbol{\epsilon_i}] = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the $K \times K$ variance-covariance matrix of the vector of error terms.

Now, let us next consider the individuals who carry two copies of the alternative allele (BB). We can model the kth phenotype of the ith individual with this genotype ($y_{i,k}^{(BB)}$) as follows:

$$y_{i,k}^{(BB)} = \mu_k + \beta_{Mk} + \beta_{Pk} + \epsilon_{i,k}, \quad i = 1, ..., n_{BB}, \; k = 1, ..., K \tag{2.3}$$

In the equation above, β_{Mk} represents the effect of the maternally-inherited B allele on the kth phenotype, and β_{Pk} represents the effect of the paternally-inherited B allele on the kth phenotype. If there is no association between this SNP and the kth phenotype, it follows that $\beta_{Mk} = \beta_{Pk} = 0$. If there is a marginal association between this SNP and the kth phenotype, but there is no POE present, then $\beta_{Mk} = \beta_{Pk} \neq 0$. Let $\boldsymbol{y_i^{(BB)}} = (y_{i,1}^{(BB)}, y_{i,2}^{(BB)}, ..., y_{i,K}^{(BB)})^\top \in \mathbb{R}^K$ be the vector of phenotypes for the ith BB individual. We can similarly rewrite the model for the phenotype vector as follows:

$$\boldsymbol{y_i^{(BB)}} = \boldsymbol{\mu} + \boldsymbol{\beta_M} + \boldsymbol{\beta_P} + \boldsymbol{\epsilon_i}, \quad i = 1, ..., n_{BB} \tag{2.4}$$

The $K \times 1$ vector $\boldsymbol{\mu}$ is as defined previously. $\boldsymbol{\beta_M} = (\beta_{M1}, ..., \beta_{MK})^\top$ is the $K \times 1$ vector of maternal effects of the B allele on each of the k phenotypes, and $\boldsymbol{\beta_P} = (\beta_{P1}, ..., \beta_{PK})^\top$ is the $K \times 1$ vector of corresponding paternal effects of the B allele. Each element of $\boldsymbol{\beta_M}$ and $\boldsymbol{\beta_P}$ is assumed to be a fixed effect. Just as for the AA subjects, we again assume that $\text{E}[\boldsymbol{\epsilon_i}] = \text{E}[(\epsilon_{i,1}, ..., \epsilon_{i,K})^\top] = \boldsymbol{0_K}$ and $\text{Cov}[\boldsymbol{\epsilon_i}] = \boldsymbol{\Sigma}$.

Lastly, let us consider the individuals who carry only one copy of the alternative allele (AB). We can model the kth phenotype of the ith heterozygous individual at this SNP ($y_{i,k}^{(AB)}$) as follows:

$$y_{i,k}^{(AB)} = \mu_k + \pi_i \beta_{Mk} + (1 - \pi_i)\beta_{Pk} + \epsilon_{i,k}, \quad i = 1, ..., n_{AB}, \ k = 1, ..., K \tag{2.5}$$

In the equation above, π_i is an indicator random variable, where $\pi_i = 1$ if individual i received the B allele from the mother and $\pi_i = 0$ if individual i received the B allele from the father. In other words, we assume $\pi_i \sim \text{Bernoulli}(\frac{1}{2})$. The parameter of this Bernoulli random variable takes value $\frac{1}{2}$ since we assume that half of heterozygotes will have maternally-derived B alleles. We can rewrite the equation in the following manner:

$$y_{i,k}^{(AB)} = \mu_k + \beta_{Pk} + (\beta_{Mk} - \beta_{Pk})\pi_i + \epsilon_{i,k}, \quad i = 1, ..., n_{AB}, \ k = 1, ..., K \tag{2.6}$$

Now, let $\boldsymbol{y_i^{(AB)}} = (y_{i,1}^{(AB)}, y_{i,2}^{(AB)}, ..., y_{i,K}^{(AB)})^\top \in \mathbb{R}^K$ be the vector of phenotypes for the ith AB individual. We can reformulate the model for the phenotype vector as follows:

$$\boldsymbol{y_i^{(AB)}} = \boldsymbol{\mu} + \boldsymbol{\beta_P} + (\boldsymbol{\beta_M} - \boldsymbol{\beta_P})\pi_i + \boldsymbol{\epsilon_i}, \quad i = 1, ..., n_{AB} \tag{2.7}$$

The maternal and paternal effect vectors are as defined as for the model of BB subjects. We also assume that $\mathrm{E}[\boldsymbol{\epsilon_i}] = \boldsymbol{0_K}$ and $\mathrm{Cov}[\boldsymbol{\epsilon_i}] = \boldsymbol{\Sigma}$. In other words, the covariance matrix of the error terms is the same within all three genotype groups.

We can easily calculate and compare the phenotypic covariance matrices across the three genotype groups assuming these models. For AA individuals, $\mathrm{Cov}(\boldsymbol{y_i^{(AA)}}) = \mathrm{Cov}(\boldsymbol{\epsilon_i}) = \boldsymbol{\Sigma},\ i = 1, ..., n_{AA}$. For BB individuals, $\mathrm{Cov}(\boldsymbol{y_i^{(BB)}}) = \mathrm{Cov}(\boldsymbol{\epsilon_i}) = \boldsymbol{\Sigma},\ i = 1, ..., n_{BB}$. We see that the phenotypic covariance matrices of the two homozygote groups (AA and BB) are both equal, and we define this matrix as $\boldsymbol{\Sigma}_{\text{Hom}} = \boldsymbol{\Sigma}$. For AB individuals, since we assume that $\pi_i \perp \boldsymbol{\epsilon_i}\ \forall\ i, i \in (1, ..., n_{AB})$, we can derive the phenotypic covariance matrix $\boldsymbol{\Sigma}_{\text{Het}} = \mathrm{Cov}(\boldsymbol{y_i^{(AB)}}) = \mathrm{Cov}[(\boldsymbol{\beta_M} - \boldsymbol{\beta_P})\pi_i + \boldsymbol{\epsilon_i}] = (\boldsymbol{\beta_M} - \boldsymbol{\beta_P})\mathrm{Var}(\pi_i)(\boldsymbol{\beta_M} - \boldsymbol{\beta_P})^\top + \boldsymbol{\Sigma} = \frac{1}{4}(\boldsymbol{\beta_M} - \boldsymbol{\beta_P})(\boldsymbol{\beta_M} - \boldsymbol{\beta_P})^\top + \boldsymbol{\Sigma}$. Let $b_k = \beta_{Mk} - \beta_{Pk}$ for $k = 1, ...K$. It can be shown that $\boldsymbol{\Sigma}_{\text{Het}} = \boldsymbol{\Sigma}_{\text{Hom}}$ if any only if:

$$\begin{bmatrix} b_1^2 & b_1 b_2 & \cdots & b_1 b_K \\ b_2 b_1 & b_2^2 & \cdots & b_2 b_K \\ \vdots & \vdots & \ddots & \vdots \\ b_K b_1 & b_K b_2 & \cdots & b_K^2 \end{bmatrix} = \boldsymbol{0}_{K \times K}$$

Thus, if a parent-of-origin effect exists for any phenotype k, then $\beta_{Mk} \neq \beta_{Pk}$, which implies $b_k \neq 0$ and $b_k^2 > 0$, and therefore the kth diagonal element of $\boldsymbol{\Sigma}_{\text{Het}}$ will be larger than the corresponding element of $\boldsymbol{\Sigma}_{\text{Hom}}$. Furthermore, if there exists POEs at this SNP for both phenotypes k and k', then $b_k b_{k'} \neq 0$ and the kk' element of $\boldsymbol{\Sigma}_{\text{Het}}$ will be different from the corresponding off-diagonal element of $\boldsymbol{\Sigma}_{\text{Hom}}$.

2.2.2 POIROT Method to Detect POE SNPs

We can test the null hypothesis that no POEs exist at a given SNP for any of the K phenotypes under study ($H_0 : \boldsymbol{\beta_M} = \boldsymbol{\beta_P}$) by equivalently testing $H_0 : \boldsymbol{\Sigma}_{\text{Het}} = \boldsymbol{\Sigma}_{\text{Hom}}$. In our proposed method POIROT, we test for equality of these phenotypic covariance matrices between homozygotes and heterozygotes using the robust omnibus (R-Omnibus) test of O'Brien [40]. POIROT uses R-Omnibus rather than the traditional Box's M test [41] to test covariance differences since the latter is highly sensitive to deviations of phenotypes from multivariate normality. This can lead to an undesirable elevation in type I error rates [42].

To derive the R-Omnibus test, we first center the phenotypes by the median within each genotype group (AA, AB, BB). This step is necessary if a marginal association exists between the alternative allele and a given phenotype. In that event, the variance of original phenotype values among aggregate homozygous subjects (AA, BB) would be erroneously inflated. We next group these centered phenotypes by homozygote versus heterozygote status. Let $x^{\text{het}}_{i,k}$ be the kth centered phenotype of the ith heterozygote ($i = 1, ..., n_{AB}$) and $x^{\text{hom}}_{j,k}$ be the kth centered penotype of the jth homozygous (AA and BB combined) individual ($j = 1, ..., n_{AA} + n_{BB}$). We then calculate the median of each phenotype k in heterozygotes and homozygotes separately. Let M^{het}_k be the median of the kth phenotype in the n_{AB} heterozygotes. Correspondingly, let M^{hom}_k be the median of the kth phenotype in the $n_{AA} + n_{BB}$ homozygotes. For heterozygotes and homozygotes separately, we then calculate for phenotypes k and k':

$$Z^{\text{het}}_{i,k,k'} = (x^{\text{het}}_{i,k} - M^{\text{het}}_k)(x^{\text{het}}_{i,k'} - M^{\text{het}}_{k'}) \tag{2.8}$$

$$Z^{\text{hom}}_{j,k,k'} = (x^{\text{hom}}_{j,k} - M^{\text{hom}}_k)(x^{\text{hom}}_{j,k'} - M^{\text{hom}}_{k'}) \tag{2.9}$$

$$W_{i,k,k'}^{\text{het}} = \frac{Z_{i,k,k'}^{\text{het}}}{|Z_{i,k,k'}^{\text{het}}|^{1/2}} \tag{2.10}$$

$$W_{j,k,k'}^{\text{hom}} = \frac{Z_{j,k,k'}^{\text{hom}}}{|Z_{j,k,k'}^{\text{hom}}|^{1/2}} \tag{2.11}$$

In Equation 2.10 and 2.11, we standardize the Z measures by dividing by the square root of their absolute values. We consider $\boldsymbol{W_i^{\text{het}}}$ to be the vector of W values for the ith heterozygous subject, and $\boldsymbol{W_j^{\text{hom}}}$ is the corresponding vector of W values for the jth homozygous subject. We then perform a two-sample Hotelling's T^2 test [43] comparing our two sets of $p = (K^2+K)/2$ samples means $\boldsymbol{(\overline{W}_{\text{het}}, \overline{W}_{\text{hom}})}$. There are p dependent variables being compared between heterozygotes and homozygotes as this corresponds to the number of upper-triangular elements in the phenotypic covariance matrix. We calculate the test statistic $t^2 = \frac{n_{\text{het}}n_{\text{hom}}}{n_{\text{het}}+n_{\text{hom}}}(\overline{\boldsymbol{W}}_{\text{het}} - \overline{\boldsymbol{W}}_{\text{hom}})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{W}}_{\text{het}} - \overline{\boldsymbol{W}}_{\text{hom}})$, where $\boldsymbol{S}^{-1}$ is the inverse of the pooled covariance matrix estimate. Under the null, our test statistic $t^2 \sim T^2(p, n_{\text{het}} + n_{\text{hom}} - 2)$ [43]. The test can also be viewed as a one-way multivariate analysis of variance test (MANOVA).

2.2.3 Post-Hoc Test for Interaction Effects

As detailed above, POIROT tests for a variant demonstrating POE by comparing/contrasting trait variances and covariances by genotype. However, trait variances can also differ by genotype when a variant exhibits a gene-gene (GxG) or gene-environment (GxE) interaction effect [44]. To increase confidence that a variant identified by POIROT demonstrates a POE rather than a more general interaction effect, we propose a post-hoc test that can be utilized to differentiate the two phenomena. The test is motivated by the observation that, for a general interaction effect, the variance of a quantitative phenotype among BB homozygous individuals is different from that of AA homozygotes. This observation is in contrast to the variance pattern expected under a POE, in which the variability of each homozygous group is the same after phenotype centering. Thus, we can craft a post-hoc test that assesses the null hypothesis of a POE (trait variance/covariances are the same between the two homozygous categories) versus the alternative of a general interaction effect (trait variance/covariances differ between the two homozygous categories). We create such a test by implementing the R-Omnibus framework as previously outlined but restricted to comparison of the two homozygous groups (AA, BB).

2.2.4 Simulation Study

We conducted a variety of simulation studies to determine POIROT's ability to detect POEs while maintaining proper rates of type I error. We considered $K = 3$, 6, or 10 phenotypes and $n = 3{,}000$, 5,000, or 10,000 unrelated individuals. To generate phenotypes for each round of simulation, we first randomly generate K intercepts from a standard normal distribution to form the $K \times 1$ vector $\boldsymbol{\mu}$. This corresponds to the mean vector of phenotypes among AA homozygotes. For simplicity, we assume the diagonal elements of the matrix $\mathbf{\Sigma}$ corresponding to the variances of the random error terms are all equal to one. We assume the K phenotypes exhibit one of three possible levels of pairwise correlation (low, medium, or high). We assume the pairwise trait correlations are randomly drawn from a uniform distribution. To simulate phenotypes exhibiting "low" correlation, we assume this is a Uniform(0,0.3) distribution. For phenotypes of "medium" and "high" correlation, we assume a Uniform(0.3,0.5) and Uniform(0.5,0.7) distribution, respectively. These random draws are used to populate the off-diagonal elements of $\mathbf{\Sigma}$.

Once we have constructed $\mathbf{\Sigma}$, we then randomly generate n maternal and paternal genotypes for a given SNP by sampling twice from a Bernoulli($p =$ MAF [minor allele frequency]) for each parent. To generate offspring genotypes, we sample from a Bernoulli($p = 0.5$) distribution to determine which maternal allele and which paternal allele is transmitted. Thus, we can now assign all n offspring to one of four genotype groups: (1) AB with maternal reference/paternal alternative, (2) AB with paternal reference/maternal alternative, (3) AA, and (4) BB. We then simulate the phenotypic error vector for all n unrelated offspring by drawing from a multivariate distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$. The respective fixed $K \times 1$ maternal and paternal effect vectors of the alternative allele $(\boldsymbol{\beta_M}, \boldsymbol{\beta_P})$ are constructed depending on the specific null or alternative scenario under consideration. We then add these vectors to the random error and intercept term in concordance with the genotype group of each individual, as described in Section 2.2.1

For type I error rate simulations, as described above, we assume these phenotypes have pairwise-trait correlation of levels low, medium, or high. To reflect the scenario where there exist no POEs or marginal effects of the alternative allele at the locus for any phenotype, we assume that $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$. We also considered a second null scenario wherein a marginal association exists for the variant that is not specific to the parent-of-origin, i.e., $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} \neq \mathbf{0}$. However, we note that if the same seeds are used in simulating the data, this marginal fixed effect is effectively removed when centering phenotypes by genotype group. The resulting test statistics are equivalent to the first null scenario. We first consider the circumstance where the random error terms are drawn from a normal distribution, i.e., the error follows $\text{MVN}_K(\mathbf{0}, \mathbf{\Sigma})$ and assume a MAF of 0.25. For each of the 27 combinations of number of phenotypes, sample size, and pairwise-trait correlation, we conducted 50,000 null simulations. To evaluate the robustness of our method to highly skewed phenotypes, we then repeated these parameter settings with non-normal random error terms. In particular, we utilize the method of Vale and Maurelli to simulate multivariate non-normal error terms assuming a skewness of two and excess kurtosis of two for each phenotype [45]. An example distribution of such a phenotype is illustrated in Appendix Figure 2.5.

Next, we investigated the power of our test when POEs do in fact exist for a locus.

We again considered $K = 3$, 6, or 10 normally distributed phenotypes. We assumed 1, 2, or 3 had parent-of-origin specific associations with the variant. When the number of affected phenotypes is greater than one, this corresponds to pleiotropy. For these scenarios, we assumed $\boldsymbol{\beta_P = 0}$ and $\beta_{Mk} = 0.5$, 0.6, or 0.75 for each phenotype k harboring a POE. All other elements of the maternal effect vector are 0 for the phenotypes with no POE associations. We again considered low, medium, and high pairwise-trait correlations. We assumed a MAF of 0.25 and sample sizes of 5,000, and 10,000. We applied our method to 5,000 simulated datasets for each of the 162 settings and calculated power at significance level $\alpha \in \{0.005, 5 \times 10^{-4}\}$. We also evaluated the power of POIROT when a locus is pleiotropic for POEs, but the magnitude of β_{Mk} varies by phenotype. For this power analysis, we again tested 3, 6, or 10 total normal phenotypes, of which 2 or 3 are harboring POEs. Since maternal effect sizes of $0.5 - 0.75$ were considered for the scenarios described above, we tested $\beta_{M1} = 0.5, \beta_{M2} = 0.75$ when two phenotypes have POEs. When 3 phenotypes have POEs, we tested power using 0.5, 0.6, and 0.75 as maternal effect sizes.

We also compared the performance of POIROT to the corresponding univariate test of Hoggart et al. [14]. For the univariate test, we first calculated power using standard Bonferroni correction. Power was calculated as the proportion of loci for which the minimum observed p-value across the K phenotypes tested was less than α/K. Given that these phenotypes are correlated and therefore may not reflect K independent tests, this approach can be overly conservative. Thus, we implemented a second more liberal approach that estimates the true number of independent tests, K_{eff}, which corresponds to the minimum number of principal components (PCs) explaining 90% of the variation in our K phenotypes. We then calculated power of the univariate approach as the proportion of loci for which the minimum observed p-value was less than α/K_{eff} [46, 47]. We then repeated these parameter settings for assessing power of POIROT with non-normal phenotypes, as described for null simulations.

Finally, we performed several simulations to investigate the performance of our proposed post-hoc test for distinguishing POEs from general interaction effects. Under the null hypothesis (i.e., there exist POEs but no interaction effects for any of the phenotypes considered), we looked at type I error of the R-Omnibus test comparing phenotypic covariances of the two homozygous groups. Similar to above, we considered a MAF of 0.25 and 3, 6, or 10 tested phenotypes, of which 1, 2, or 3 had POEs but no interaction effects. We considered sample sizes of 5,000 and 10,000, maternal POE effect sizes $\{0.5, 0.6, 0.75\}$, and low/medium/high trait correlation. We also evaluated the power of this post-hoc test to identify GxE effects when present. Simulation parameters were informed by prior work of Paré et al. [44]. We considered a single unmeasured covariate drawn from a standard normal distribution. Again, we considered 3, 6, or 10 total quantitative traits, of which 3 had a non-negligible covariate effect. Of these three phenotypes, 1, 2 or 3 had gene-covariate interaction effects. The covariate effect sizes ranged from 0.3 to 0.7. Among the phenotypes with gene-covariate interaction effects, we varied to the proportion of total variation of each phenotype explained by the interaction effects between 0.005 and 0.01. Again, we allowed traits to have varying pairwise correlation. We performed 5,000 simulations for each of the 216 power settings outlined for the post-hoc interaction test.

2.2.5 Application of POIROT to UK Biobank

To assess the utility of POIROT for detecting POEs on continuous phenotypes using published population-based GWAS data, we utilized genotype and phenotype data from the UK Biobank (UKB), a large-scale biomedical database housing data collected from approximately 500,000 individuals from the UK. This study allows for widespread investigation of the genetic variation associated with hundreds of lifestyle and health factors. To identify potential POE variants, we obtained data on three quantitative phenotypes (BMI [kg/m^2], high-density lipoprotein [HDL] cholesterol [mmol/L], and low-density lipoprotein [LDL] cholesterol [mmol/L]). Relevant covariates included genotyping array, PCs, sex, age at recruitment, and smoking status (prefer not to answer, never, previous, current). Prior to analysis, we removed all individuals identified as outliers according to pre-calculated metrics of genotype missingness, heterozygosity, and excess relatedness. We excluded those with putative sex chromosome aneuploidy and those who were not included in PCA calculation. We included individuals of self-reported white British ancestry only.

Subjects were genotyped using either the UK BiLEVE or UK Biobank Axiom arrays. We considered only autosomal variants with MAF > 0.05, Hardy-Weinberg equilibrium $p > 1 \times 10^{-8}$, and missingness rate < 0.02. After quality control and filtering, 330,801 SNPs remained for analysis across 292,779 unrelated individuals with complete phenotype and covariate information. There is moderate negative correlation between BMI and HDL cholesterol (Pearson's r = -0.35), low positive correlation between BMI and LDL ($r = 0.02$), and low positive correlation between LDL and HDL ($r = 0.10$). However, all estimated correlations are statistically significant ($p < 2.2 \times 10^{-16}$). Covariate adjustment was performed by first fitting a linear model for each phenotype and extracting the residuals as the new adjusted phenotypes. We then applied POIROT to these three adjusted phenotypes to jointly test for POEs across the genome. We compared the findings of our approach to those from the method of Hoggart et al. performed on each phenotype individually. For any variant identified by POIROT meeting the Bonferroni-adjusted genome-wide significance threshold, we applied our proposed post-hoc test to assess if the effect might be explained by a general interaction effect rather than a POE.

We concluded with a follow-up analysis to determine whether we see enrichment of variants in imprinting regions among those with lowest POIROT p-values for detecting POEs in the UKB cohort. We first downloaded genes of known imprinting and predicted imprinting status in humans from the GeneImprint database (https://www.geneimprint.com). We then determined which variants in the UKB dataset fell within 500kb of the starting and ending site of these genes. We defined these as our variant set of interest (comparable to a gene set in Gene Set Enrichment Analysis [GSEA]). We then utilized the GSEAPreranked tool to test for enrichment of variants in this set among those top ranked variants by $-\log_{10}$(POIROT p-value) [48, 49].

2.3 Results

2.3.1 Type I Error Rate

We summarize the type I error of null scenarios with a sample size of 5,000 individuals using Quantile-Quantile (QQ) plots in Figure 2.1 (normal traits) and Figure 2.2 (non-normal traits). Across the settings considered, our method yields the expected distribution of p-values under the null hypothesis of no POEs for any single phenotype. The distribution of the p-values is again as expected under the null when we have non-normality of phenotypes (Figure 2.2), suggesting our method remains robust. We summarize the empirical type I error rates of our proposed test and the competing univariate approach at significance level $\alpha \in \{0.05, 0.005, 5\times10^{-4}, 5\times10^{-5}\}$ in Tables 2.1, 2.2, 2.3, and 2.4 (Appendix). POIROT maintained appropriate type I error across all scenarios for normally distributed traits. We observed slightly higher error when 6 or 10 highly-skewed non-normal phenotypes were tested. The univariate approach with correction for K_{eff} tests showed minor inflation with 6 or 10 highly correlated phenotypes.

2.3.2 Power

Simulation results comparing the performance of POIROT to the competing univariate test under the assumption of true POE(s) are summarized in Figure 2.3. This

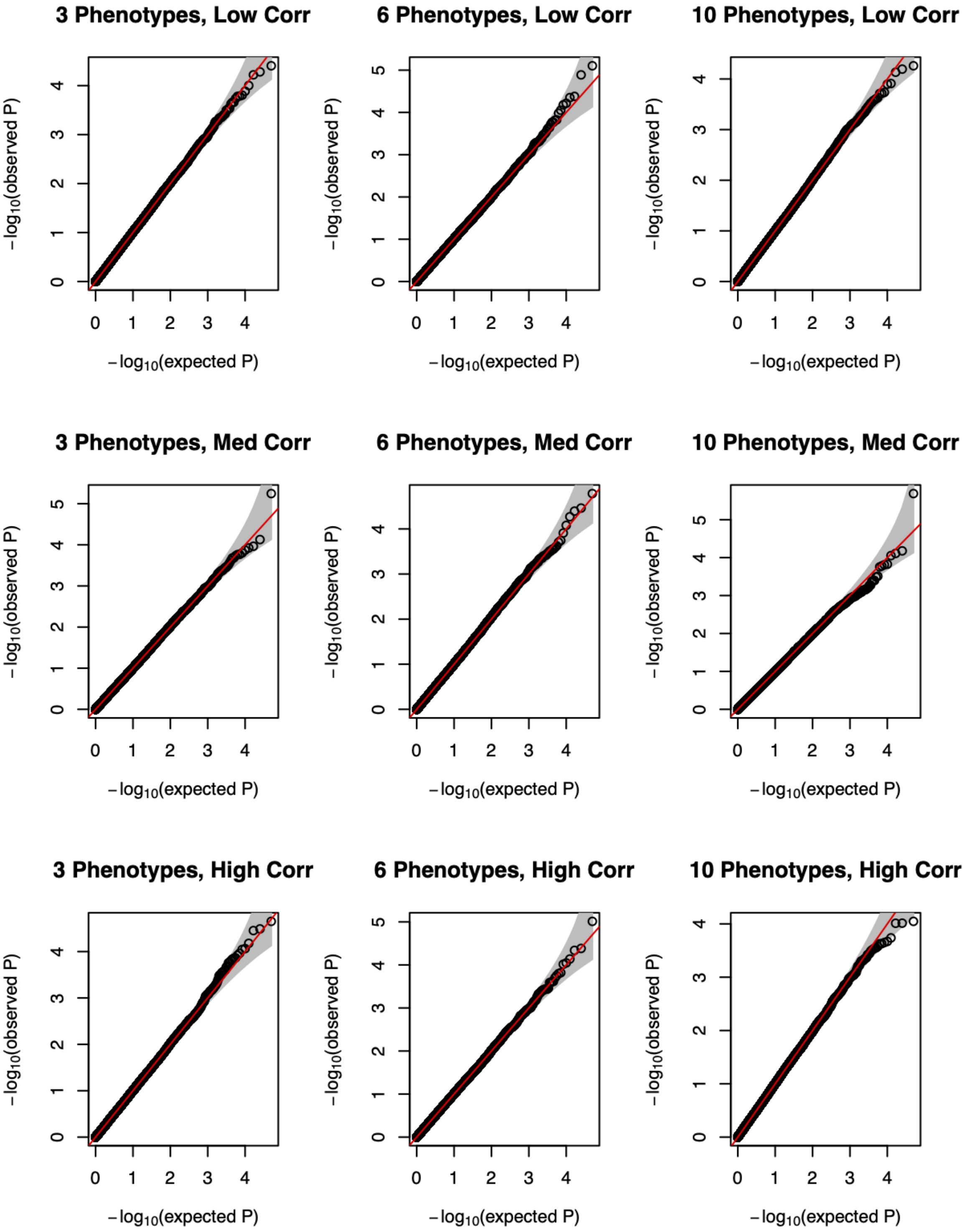

Figure 2.1: Q-Q plots of p-values for proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ using a series of 50,000 simulations of 5,000 individuals using 3 (left column), 6 (middle column) or 10 (right column) continuous normal phenotypes. Minor allele frequency is assumed to be 0.25. Horizontal panels depict level of pairwise-trait correlation (low, medium, high).

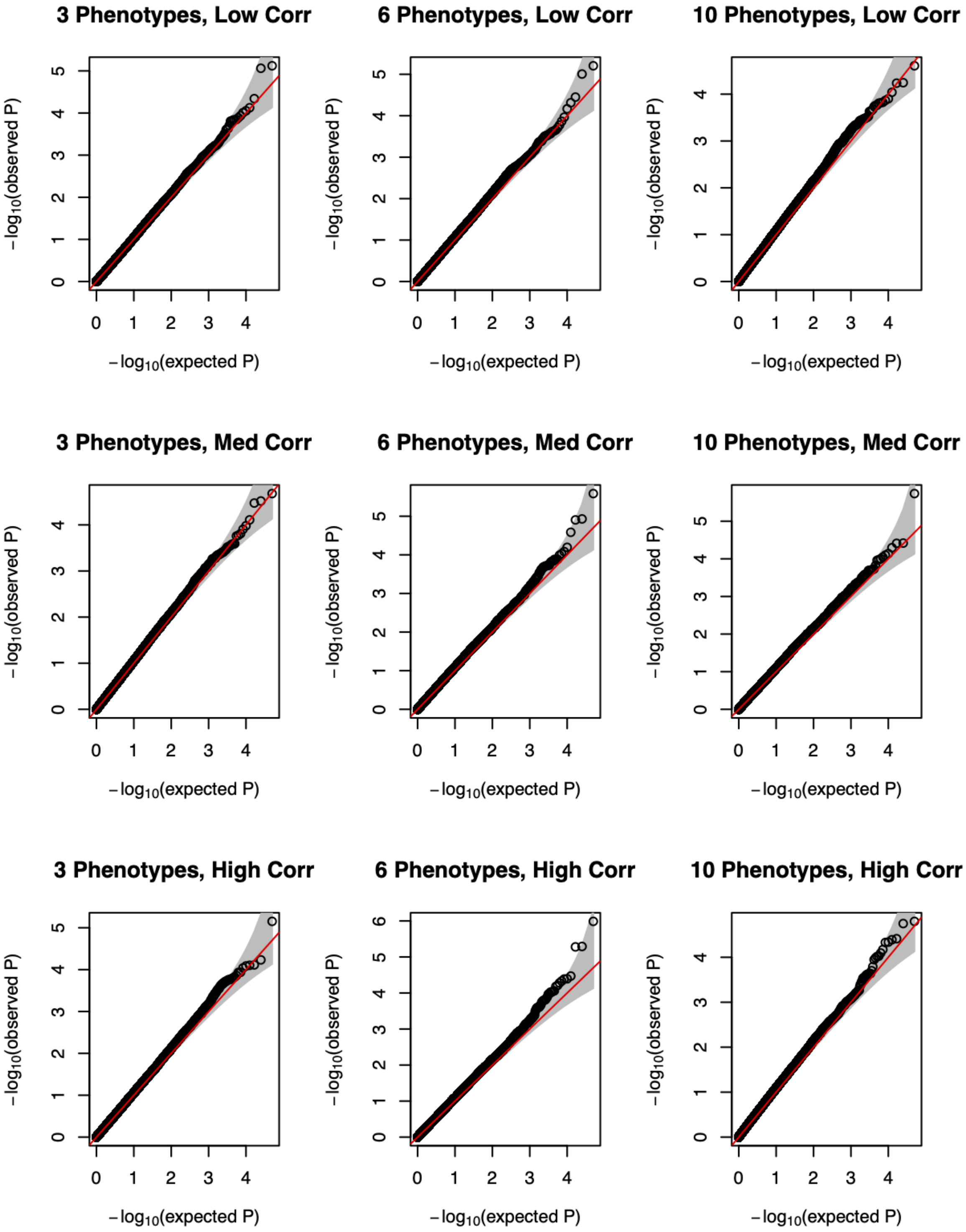

Figure 2.2: Q-Q plots of p-values for proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ using a series of 50,000 simulations of 5,000 individuals using 3 (left column), 6 (middle column) or 10 (right column) continuous non-normal phenotypes. Minor allele frequency is assumed to be 0.25. Horizontal panels depict level of pairwise-trait correlation (low, medium, high).

figure reflects normally distributed traits and sample size of 5,000 ($\alpha = 5 \times 10^{-4}$). Corresponding results from all other additional power settings, including both normal and non-normal traits, sample sizes of 5,000 and 10,000, and $\alpha = 0.005, 5 \times 10^{-4}$ are largely consistent with the results shown below.

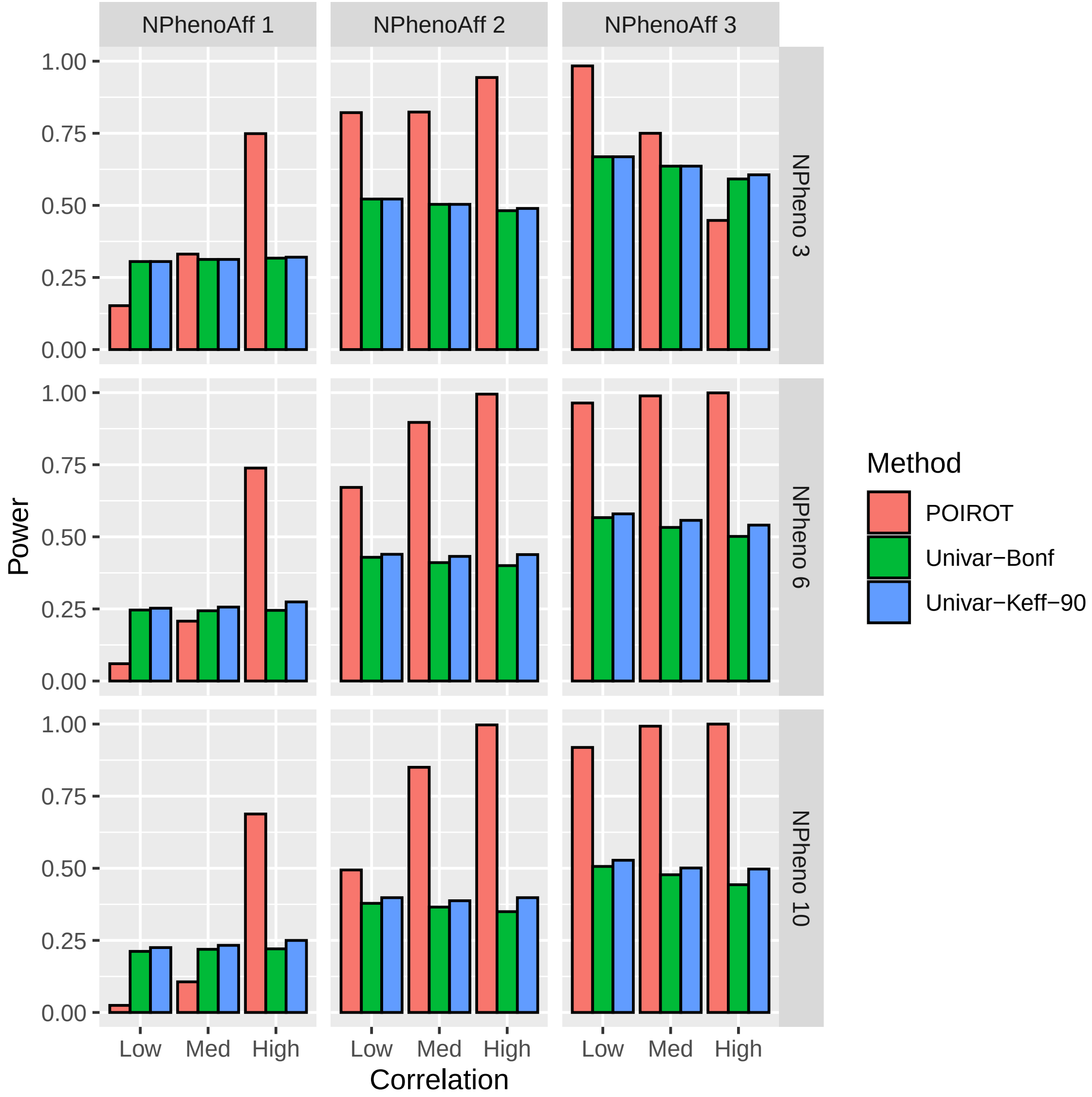

Figure 2.3: Power of POIROT to identify POEs assuming $K = 3$, 6, or 10 normal phenotypes (horizontal panels) compared to univariate test. We assume either 1, 2, or 3 of the phenotypes harbor POEs at the locus (vertical panels). We performed 5,000 simulations for each scenario. We calculated power at significance level 0.0005 for our multi-trait test and $0.0005/K$ (Bonferroni correction) and $0.0005/K_{\text{eff}}$ for the univariate test, where K_{eff} is the number of PCs needed to explain 90% phenotypic variation. $\boldsymbol{\beta}_{Mk} = 0.75$ for POE traits, MAF $= 0.25$, and sample size $=$ 5,000.

Simple Bonferroni correction tends to be overly conservative in the presence of correlated traits. We therefore used two multiple-testing correction approaches for the univariate method. As power generally increases with increasing sample size and POE magnitude, the scenarios shown in Figure 2.3 correspond to a β_{Mk} of 0.75 and sample size of 5,000. For almost all scenarios, we see three general trends. First, unlike the univariate method, our method successfully leverages the correlation among phenotypes. We see power increasing with increasing trait correlation. Second, when pleiotropy exists and more than one phenotype harbors a POE, our method outperforms the univariate approach regardless of the multiple testing correction strategy. Third, power of POIROT increases as the number of phenotypes associated with the maternally-transmitted alternative allele increases across all levels of phenotypic correlation. Under simulated pleiotropic POE loci with varying β_{Mk}, the power of POIROT tends to reflect the power assuming a constant β_{Mk} for POE phenotypes at the median effect size.

The one exception to these trends is the top right panel of Figure 2.3. This reflects the scenario where 3 of 3 phenotypes harbor POEs of the same magnitude and direction. We see here that power decreases going from low to medium correlation and from medium to high correlation. We also see lower power when 3 phenotypes are affected when compared to the corresponding settings when only 2 of 3 phenotypes have POEs. This pattern, although unusual, has been documented in previous cross-phenotype methodological studies [47, 50]. As described in Section 2.2.2, the R-Omnibus test for equality of covariance matrices used by POIROT ultimately employs a one-way MANOVA test to generate our test statistic. Ray et al. describe how when we have K correlated traits being tested and a SNP is associated with all K traits, utilizing a MANOVA to find marginal associations with multiple traits can result in an appreciable loss of power. In particular, the authors show how the power of MANOVA is asymptotically lower when all traits are associated with equal magnitude and direction than when fewer than K phenotypes are associated [50].

2.3.3 Post-Hoc Interaction Test

Type I error results of our post-hoc test for distinguishing POE (null) from general interaction effects (alternative) are shown in Figure 2.6 (Appendix). This is an illustrative example when only POEs exists for a sample size of 10,000 and the maternal POE effect size is 0.75. These results are indicative of all null simulation settings which show the test was well-calibrated under the null when the only effects were parent-of-origin-dependent. Under alternative simulations with a GxE interaction effect, our post-hoc test had the power to differentiate interaction effects from POEs (Figures 2.7, 2.8 [Appendix]). Power is increasing with increasing number of phenotypes with non-null interaction effects, sample size, strength of interaction effect, and generally, pairwise trait correlation.

2.3.4 Applied Data Analysis

We applied our method for detecting POEs to genotype and multivariate phenotype data of 292,779 individuals of European ancestry from the UK Biobank. Raw quantitative phenotype measures of interest were BMI, HDL cholesterol, and LDL direct cholesterol. Phenotypes were appropriately adjusted for the effects of genotype array, PCs, sex, age, and smoking status. For the 330,801 variants considered, the average computation time per POIROT test was 22.53 seconds. Analysis was run with parallel computation with the genome segmented into 793 blocks with a maximum block runtime of 4.7 hours (681 variants). We identified a total of 338 variants with POIROT p-values falling below the Bonferroni-adjusted genome-wide significance threshold of 1.5×10^{-7}. These suggestive POE variants are shown in the Manhattan plot in Figure 2.4.

We also saw a significant positive normalized enrichment score (nominal $p < 0.001$)

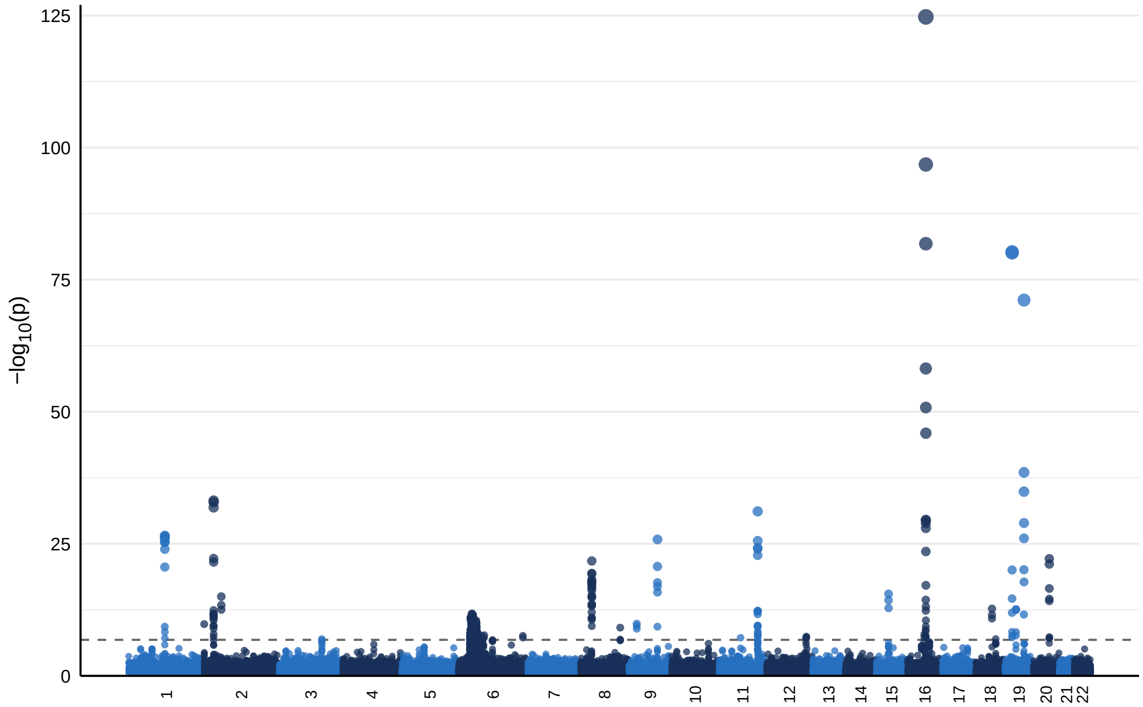

Figure 2.4: Manhattan plot of parent-of-origin effects analysis using POIROT with BMI, HDL cholesterol, and LDL cholesterol phenotypes from the UK Biobank. The dashed line represents Bonferroni-adjusted genome-wide significance of 1.5×10^{-7}.

from the GSEA follow-up test, indicating that variants within 500kb of imprinted or predicted-imprinted genes tended to lie at the top of our list ranked by increasing POIROT p-value. We next applied our post-hoc test to these 338 identified variants to evaluate whether any demonstrated general interaction effects and observed that approximately two-thirds (230) had $p > 0.05/338$ and failed to reject the null of a POE. We similarly applied the univariate test for POEs genome-wide using each individual phenotype separately.

In Table 2.7 (Appendix), we report on the 41 variants identified by POIROT as potential POE loci that were not identified by any of the three univariate tests for POEs and further were not significantly demonstrating general interaction effects based on our post-hoc test. These 41 variants thus represent the strongest evidence for novel POE effect(s) in our analysis. Among them, we highlight one exonic variant (Affx-20090007, POIROT $p = 9.7 \times 10^{-16}$) and one intronic variant (rs41360247, POIROT $p = 3.0 \times 10^{-13}$) on chromosome 2 for gene ABCG8. Polymorphisms in this gene have previously been associated with direct LDL in UKB samples [51, 52]. Variants within this gene have additionally been associated with cholesterol phenotypes in analyses outside of the UK Biobank dataset [53]. Of particular note, ABCG8 has been shown by prior research to be a high-confidence gene for maternal imprinting [54]. We also wish to highlight variants identified by POIROT around the gene APOB on chromosome 2. Of 14 POIROT-identified variants mapping to this gene, two failed to show evidence of significant interaction effects by our post-hoc test (rs550619 [intronic, POIROT $p = 3.1 \times 10^{-10}$], rs74629722 [intergenic, POIROT $p = 3.3 \times 10^{-10}$]). In particular, rs550619 lies 3,299bp from a previously-published POE variant for BMI (rs1367117) [55] and has significant GWAS associations with direct LDL levels and total cholesterol phenotypes [51, 52]. Neither of these variants were identified for any of the three tested phenotypes using the existing univariate approach to detect POEs.

2.4 Discussion

In this project, we introduce a multivariate method, POIROT, for identifying common variants exhibiting POEs on one or more quantitative phenotypes in unrelated subjects. This work is motivated dually by the widespread evidence of pleiotropy in the genetics literature, as well as the limited statistical options for detecting POEs in unrelated cohorts. Our proposed method is an inherently simple statistical test of whether the phenotypic covariance matrix of heterozygotes is equal to that of homozygotes at a given locus. It represents a multivariate extension of the POE test of a single continuous phenotype proposed by Hoggart et al. [14]. It allows for appropriate adjustment for the effects of important covariates on the phenotypes under study and is also computationally efficient for application to biobank-scale datasets

(Appendix Tables 2.5, 2.6).

Through simulations, we demonstrate POIROT achieves appropriate type I error under the null. It further displays superior power to detect POEs than the competing univariate approach under most settings. Our method is indeed robust to non-normality of phenotypes across several simulation scenarios. We further applied our method to real GWAS data on white individuals of European ancestry from the UKB. In this analysis, we considered BMI as well as HDL and direct LDL cholesterol as potential imprinted phenotypes. The analysis revealed 338 variants meeting the stringent genome-wide significance threshold. Of these, 41 may warrant particular focus in future investigations. They were not identified by the existing univariate approach to detect POEs and did not show evidence of significant gene-gene or gene-environment interaction effects using our proposed post-hoc test. Two of these variants map to gene ABCG8, a gene with high confidence of maternal imprinting in humans based on previously published work, and another lies nearby a known POE variant for BMI in the gene APOB.

While the results presented here are promising for the utility of our proposed multivariate method for POE detection in practice, there are inherent limitations that we must address. Firstly, we propose POIROT as a method to detect SNPs wherein the effect of the variant allele in offspring differs according to which parent transmitted it. We do not evaluate the ability of our method to detect other trans-generational effects that may appear as imprinting effects at surface evaluation [3]. Furthermore, we acknowledge that our method to detect POEs by evaluation of differing phenotypic covariance matrices by genotype groups may lead to false positive identifications at loci where gene-environment or gene-gene interaction effects exist. We have proposed a two-stage screening procedure to combat this: first by implementing POIROT as described, and second, by following up with our proposed test to distinguish which findings may be the result of more general interaction effects. We also note if a trans-generational effect exists by which, for example, the maternal genotype is affecting the offspring phenotype in a manner that is not completely independent of offspring genotype (in other words, there are maternal-fetal genotype interaction effects), we do believe we would be able to detect these in our post-hoc test for interaction effects.

POIROT is a variance/covariance-based test for detecting POEs applicable to large population samples where allelic parental origin is unknown. If parental genetic information is known (i.e., through collection of parent-offspring trios), then it is well-established that variance-based tests within the offspring are often considerably less powerful than mean-based tests (like those described in the Introduction) that leverage allelic parental origin and look for differences in phenotypic means between heterozygous offspring with maternally- vs. paternally-inherited effect alleles [56]. We performed additional simulations comparing the power of the two strategies at different sample sizes. Specifically, we simulated parent-offspring trio genotype data, restricted samples for analysis to include only heterozygous offspring, and tested for mean-based differences in phenotypes between offspring who inherited the variant alleles maternally versus those who inherited it paternally via one-way MANOVA. We assumed 2 out of 3, 6, or 10 phenotypes harbored a POE. We generally found that variance/covariance methods require approximately 13 times as many observations as familial mean-based tests for equivalent power. The trio-based simulations assumed full knowledge of parental transmission of the variant allele in heterozygous offspring, when in reality, parent-of-origin may be ambiguous in certain cases. For details, please see Figure 2.9 (Appendix). Thus, if family-based data are available, we recommend the use of mean-based tests for POE detection rather than variance-based tests. For population studies, variance-based tests remain the only option for POE analysis.

2.5 Appendix

2.5.1 Proofs

Proof of equivalence of null simulations with and without marginal effects

Consider one phenotype. Let us assume there are n heterozygotes. Let $\boldsymbol{x}$ be the $n \times 1$ vector of simulated phenotypes for AB individuals when there are no maternal or paternal effects of the B allele. When we center these phenotpes by the sample average of AB individuals, we are left with the new vector $\boldsymbol{x} - \frac{\sum x_i}{n}$. When we assume a marginal effect of the B allele with no POE ($\beta_M = \beta_P = \beta \neq 0$), then the vector of simulated phenotypes for the n heterozygotes will now be $\boldsymbol{x}+\beta$. When we again center by the sample average of AB heterozygotes, we are left with $\boldsymbol{x}+\beta-\frac{\sum(x_i+\beta)}{n} = \boldsymbol{x}-\frac{\sum x_i}{n}$. Thus, the centered phenotypes are the same in both scenarios, and the downstream p-value will be the same regardless of the number of phenotypes with marginal effects (no POEs) and the magnitude of those effects.

2.5.2 Tables

Table 2.1: Empirical type I error rates at significance level $\alpha = 0.05$ of proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ for 50,000 simulations performed per scenario.

N	NPheno	Error	Correl	POIROT	Univar-Bonf	Univar-Keff-90
3000	3	Normal	Low	0.050	0.050	0.050
3000	3	Normal	Medium	0.050	0.047	0.047
3000	3	Normal	High	0.051	0.046	0.050
3000	3	Non-normal	Low	0.055	0.051	0.051
3000	3	Non-normal	Medium	0.057	0.048	0.048
3000	3	Non-normal	High	0.059	0.045	0.045
3000	6	Normal	Low	0.050	0.051	0.056
3000	6	Normal	Medium	0.050	0.047	0.057
3000	6	Normal	High	0.050	0.044	0.062
3000	6	Non-normal	Low	0.058	0.050	0.051
3000	6	Non-normal	Medium	0.060	0.047	0.056
3000	6	Non-normal	High	0.062	0.044	0.052
3000	10	Normal	Low	0.048	0.052	0.060
3000	10	Normal	Medium	0.048	0.049	0.061
3000	10	Normal	High	0.048	0.042	0.064
3000	10	Non-normal	Low	0.058	0.049	0.054
3000	10	Non-normal	Medium	0.061	0.048	0.057
3000	10	Non-normal	High	0.061	0.042	0.052
5000	3	Normal	Low	0.051	0.050	0.050
5000	3	Normal	Medium	0.050	0.050	0.050
5000	3	Normal	High	0.050	0.047	0.051
5000	3	Non-normal	Low	0.057	0.051	0.051
5000	3	Non-normal	Medium	0.058	0.048	0.048
5000	3	Non-normal	High	0.059	0.046	0.046
5000	6	Normal	Low	0.051	0.051	0.056
5000	6	Normal	Medium	0.051	0.048	0.057
5000	6	Normal	High	0.050	0.045	0.062
5000	6	Non-normal	Low	0.059	0.049	0.049
5000	6	Non-normal	Medium	0.063	0.048	0.057
5000	6	Non-normal	High	0.063	0.043	0.051
5000	10	Normal	Low	0.049	0.050	0.057
5000	10	Normal	Medium	0.049	0.049	0.060
5000	10	Normal	High	0.051	0.043	0.065
5000	10	Non-normal	Low	0.062	0.050	0.056
5000	10	Non-normal	Medium	0.064	0.048	0.058

5000	10	Non-normal	High	0.063	0.044	0.053
10000	3	Normal	Low	0.049	0.050	0.050
10000	3	Normal	Medium	0.049	0.048	0.048
10000	3	Normal	High	0.050	0.046	0.050
10000	3	Non-normal	Low	0.055	0.050	0.050
10000	3	Non-normal	Medium	0.056	0.047	0.047
10000	3	Non-normal	High	0.058	0.046	0.046
10000	6	Normal	Low	0.050	0.049	0.053
10000	6	Normal	Medium	0.051	0.048	0.056
10000	6	Normal	High	0.051	0.044	0.061
10000	6	Non-normal	Low	0.059	0.049	0.049
10000	6	Non-normal	Medium	0.064	0.047	0.055
10000	6	Non-normal	High	0.062	0.044	0.051
10000	10	Normal	Low	0.051	0.050	0.058
10000	10	Normal	Medium	0.050	0.048	0.058
10000	10	Normal	High	0.050	0.042	0.064
10000	10	Non-normal	Low	0.060	0.048	0.053
10000	10	Non-normal	Medium	0.065	0.045	0.053
10000	10	Non-normal	High	0.062	0.040	0.050

Table 2.2: Empirical type I error rates at significance level $\alpha = 0.005$ of proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ for 50,000 simulations performed per scenario.

N	NPheno	Error	Correl	POIROT	Univar-Bonf	Univar-Keff-90
3000	3	Normal	Low	0.005	0.005	0.005
3000	3	Normal	Medium	0.005	0.005	0.005
3000	3	Normal	High	0.005	0.005	0.005
3000	3	Non-normal	Low	0.006	0.005	0.005
3000	3	Non-normal	Medium	0.007	0.005	0.005
3000	3	Non-normal	High	0.007	0.005	0.005
3000	6	Normal	Low	0.005	0.005	0.005
3000	6	Normal	Medium	0.004	0.004	0.005
3000	6	Normal	High	0.004	0.005	0.007
3000	6	Non-normal	Low	0.007	0.005	0.005
3000	6	Non-normal	Medium	0.007	0.005	0.007
3000	6	Non-normal	High	0.007	0.005	0.006
3000	10	Normal	Low	0.005	0.005	0.007
3000	10	Normal	Medium	0.005	0.005	0.006
3000	10	Normal	High	0.005	0.005	0.008
3000	10	Non-normal	Low	0.007	0.005	0.006

3000	10	Non-normal	Medium	0.007	0.005	0.007
3000	10	Non-normal	High	0.007	0.005	0.006
5000	3	Normal	Low	0.005	0.005	0.005
5000	3	Normal	Medium	0.005	0.005	0.005
5000	3	Normal	High	0.005	0.005	0.005
5000	3	Non-normal	Low	0.006	0.005	0.005
5000	3	Non-normal	Medium	0.006	0.005	0.005
5000	3	Non-normal	High	0.007	0.005	0.005
5000	6	Normal	Low	0.005	0.005	0.005
5000	6	Normal	Medium	0.006	0.005	0.006
5000	6	Normal	High	0.005	0.004	0.007
5000	6	Non-normal	Low	0.007	0.005	0.005
5000	6	Non-normal	Medium	0.007	0.005	0.006
5000	6	Non-normal	High	0.007	0.005	0.006
5000	10	Normal	Low	0.005	0.005	0.006
5000	10	Normal	Medium	0.005	0.005	0.007
5000	10	Normal	High	0.005	0.005	0.008
5000	10	Non-normal	Low	0.007	0.005	0.006
5000	10	Non-normal	Medium	0.007	0.006	0.007
5000	10	Non-normal	High	0.007	0.005	0.006
10000	3	Normal	Low	0.004	0.005	0.005
10000	3	Normal	Medium	0.005	0.005	0.005
10000	3	Normal	High	0.005	0.005	0.005
10000	3	Non-normal	Low	0.006	0.005	0.005
10000	3	Non-normal	Medium	0.007	0.005	0.005
10000	3	Non-normal	High	0.007	0.005	0.005
10000	6	Normal	Low	0.005	0.005	0.005
10000	6	Normal	Medium	0.005	0.005	0.006
10000	6	Normal	High	0.005	0.005	0.007
10000	6	Non-normal	Low	0.007	0.005	0.005
10000	6	Non-normal	Medium	0.007	0.005	0.006
10000	6	Non-normal	High	0.007	0.005	0.006
10000	10	Normal	Low	0.005	0.005	0.006
10000	10	Normal	Medium	0.005	0.006	0.007
10000	10	Normal	High	0.005	0.005	0.008
10000	10	Non-normal	Low	0.007	0.005	0.006
10000	10	Non-normal	Medium	0.007	0.005	0.007
10000	10	Non-normal	High	0.007	0.005	0.006

Table 2.3: Empirical type I error rates at significance level $\alpha = 5 \times 10^{-4}$ of proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ for 50,000 simulations performed per scenario.

N	NPheno	Error	Correl	POIROT	Univar-Bonf	Univar-Keff-90
3000	3	Normal	Low	0.00044	0.00058	0.00058
3000	3	Normal	Medium	0.00044	0.00042	0.00042
3000	3	Normal	High	0.00048	0.00046	0.00050
3000	3	Non-normal	Low	0.00062	0.00060	0.00060
3000	3	Non-normal	Medium	0.00084	0.00050	0.00050
3000	3	Non-normal	High	0.00078	0.00056	0.00056
3000	6	Normal	Low	0.00038	0.00052	0.00058
3000	6	Normal	Medium	0.00040	0.00052	0.00064
3000	6	Normal	High	0.00042	0.00056	0.00078
3000	6	Non-normal	Low	0.00060	0.00060	0.00060
3000	6	Non-normal	Medium	0.00102	0.00064	0.00072
3000	6	Non-normal	High	0.00086	0.00054	0.00062
3000	10	Normal	Low	0.00036	0.00076	0.00086
3000	10	Normal	Medium	0.00044	0.00044	0.00054
3000	10	Normal	High	0.00046	0.00050	0.00076
3000	10	Non-normal	Low	0.00082	0.00062	0.00068
3000	10	Non-normal	Medium	0.00088	0.00066	0.00084
3000	10	Non-normal	High	0.00086	0.00054	0.00066
5000	3	Normal	Low	0.00052	0.00072	0.00072
5000	3	Normal	Medium	0.00052	0.00060	0.00060
5000	3	Normal	High	0.00058	0.00042	0.00046
5000	3	Non-normal	Low	0.00052	0.00060	0.00060
5000	3	Non-normal	Medium	0.00064	0.00058	0.00058
5000	3	Non-normal	High	0.00080	0.00050	0.00050
5000	6	Normal	Low	0.00056	0.00042	0.00050
5000	6	Normal	Medium	0.00054	0.00046	0.00052
5000	6	Normal	High	0.00056	0.00050	0.00078
5000	6	Non-normal	Low	0.00060	0.00054	0.00054
5000	6	Non-normal	Medium	0.00076	0.00044	0.00058
5000	6	Non-normal	High	0.00086	0.00064	0.00076
5000	10	Normal	Low	0.00046	0.00064	0.00070
5000	10	Normal	Medium	0.00026	0.00050	0.00064
5000	10	Normal	High	0.00050	0.00066	0.00096
5000	10	Non-normal	Low	0.00084	0.00072	0.00076
5000	10	Non-normal	Medium	0.00078	0.00064	0.00074
5000	10	Non-normal	High	0.00058	0.00064	0.00072
10000	3	Normal	Low	0.00048	0.00054	0.00054
10000	3	Normal	Medium	0.00050	0.00048	0.00048
10000	3	Normal	High	0.00046	0.00054	0.00058

10000	3	Non-normal	Low	0.00048	0.00038	0.00038
10000	3	Non-normal	Medium	0.00056	0.00050	0.00050
10000	3	Non-normal	High	0.00052	0.00058	0.00058
10000	6	Normal	Low	0.00054	0.00054	0.00060
10000	6	Normal	Medium	0.00052	0.00062	0.00074
10000	6	Normal	High	0.00036	0.00052	0.00070
10000	6	Non-normal	Low	0.00064	0.00072	0.00074
10000	6	Non-normal	Medium	0.00094	0.00058	0.00066
10000	6	Non-normal	High	0.00062	0.00056	0.00070
10000	10	Normal	Low	0.00036	0.00060	0.00064
10000	10	Normal	Medium	0.00050	0.00054	0.00088
10000	10	Normal	High	0.00040	0.00050	0.00090
10000	10	Non-normal	Low	0.00086	0.00050	0.00052
10000	10	Non-normal	Medium	0.00078	0.00034	0.00050
10000	10	Non-normal	High	0.00068	0.00048	0.00066

Table 2.4: Empirical type I error rates at significance level $\alpha = 5 \times 10^{-5}$ of proposed POE test under the null hypothesis $\boldsymbol{\beta_M} = \boldsymbol{\beta_P} = \mathbf{0}$ for 50,000 simulations performed per scenario.

N	NPheno	Error	Correl	POIROT	Univar-Bonf	Univar-Keff-90
3000	3	Normal	Low	4.0E-05	2.0E-05	2.0E-05
3000	3	Normal	Medium	6.0E-05	2.0E-05	2.0E-05
3000	3	Normal	High	8.0E-05	2.0E-05	6.0E-05
3000	3	Non-normal	Low	1.4E-04	1.6E-04	1.6E-04
3000	3	Non-normal	Medium	4.0E-05	6.0E-05	6.0E-05
3000	3	Non-normal	High	4.0E-05	8.0E-05	8.0E-05
3000	6	Normal	Low	2.0E-05	4.0E-05	4.0E-05
3000	6	Normal	Medium	4.0E-05	8.0E-05	1.0E-04
3000	6	Normal	High	4.0E-05	2.0E-05	4.0E-05
3000	6	Non-normal	Low	6.0E-05	8.0E-05	8.0E-05
3000	6	Non-normal	Medium	1.0E-04	1.0E-04	1.2E-04
3000	6	Non-normal	High	2.0E-04	1.2E-04	1.2E-04
3000	10	Normal	Low	2.0E-05	8.0E-05	8.0E-05
3000	10	Normal	Medium	6.0E-05	8.0E-05	8.0E-05
3000	10	Normal	High	8.0E-05	8.0E-05	1.6E-04
3000	10	Non-normal	Low	4.0E-05	6.0E-05	6.0E-05
3000	10	Non-normal	Medium	8.0E-05	6.0E-05	6.0E-05
3000	10	Non-normal	High	6.0E-05	2.0E-05	2.0E-05
5000	3	Normal	Low	2.0E-05	1.0E-04	1.0E-04
5000	3	Normal	Medium	2.0E-05	4.0E-05	4.0E-05

5000	3	Normal	High	6.0E-05	8.0E-05	8.0E-05
5000	3	Non-normal	Low	6.0E-05	1.0E-04	1.0E-04
5000	3	Non-normal	Medium	6.0E-05	6.0E-05	6.0E-05
5000	3	Non-normal	High	2.0E-05	4.0E-05	4.0E-05
5000	6	Normal	Low	8.0E-05	2.0E-05	2.0E-05
5000	6	Normal	Medium	6.0E-05	6.0E-05	6.0E-05
5000	6	Normal	High	6.0E-05	6.0E-05	1.0E-04
5000	6	Non-normal	Low	8.0E-05	4.0E-05	4.0E-05
5000	6	Non-normal	Medium	8.0E-05	2.0E-05	4.0E-05
5000	6	Non-normal	High	1.4E-04	4.0E-05	8.0E-05
5000	10	Normal	Low	0.0E+00	2.0E-05	2.0E-05
5000	10	Normal	Medium	2.0E-05	4.0E-05	6.0E-05
5000	10	Normal	High	0.0E+00	4.0E-05	8.0E-05
5000	10	Non-normal	Low	2.0E-05	2.0E-05	2.0E-05
5000	10	Non-normal	Medium	6.0E-05	6.0E-05	8.0E-05
5000	10	Non-normal	High	1.2E-04	1.4E-04	1.4E-04
10000	3	Normal	Low	4.0E-05	4.0E-05	4.0E-05
10000	3	Normal	Medium	2.0E-05	1.2E-04	1.2E-04
10000	3	Normal	High	2.0E-05	6.0E-05	6.0E-05
10000	3	Non-normal	Low	4.0E-05	2.0E-05	2.0E-05
10000	3	Non-normal	Medium	2.0E-05	4.0E-05	4.0E-05
10000	3	Non-normal	High	6.0E-05	4.0E-05	4.0E-05
10000	6	Normal	Low	2.0E-05	8.0E-05	8.0E-05
10000	6	Normal	Medium	4.0E-05	2.0E-05	2.0E-05
10000	6	Normal	High	2.0E-05	0.0E+00	0.0E+00
10000	6	Non-normal	Low	1.0E-04	8.0E-05	8.0E-05
10000	6	Non-normal	Medium	1.6E-04	4.0E-05	8.0E-05
10000	6	Non-normal	High	1.0E-04	6.0E-05	6.0E-05
10000	10	Normal	Low	0.0E+00	6.0E-05	8.0E-05
10000	10	Normal	Medium	4.0E-05	2.0E-05	4.0E-05
10000	10	Normal	High	4.0E-05	2.0E-05	8.0E-05
10000	10	Non-normal	Low	8.0E-05	2.0E-05	2.0E-05
10000	10	Non-normal	Medium	1.0E-04	4.0E-05	4.0E-05
10000	10	Non-normal	High	8.0E-05	4.0E-05	4.0E-05

Table 2.5: Median computational time per POIROT test by sample size and number of phenotypes (K). Each row is based on 45,000 simulations.

K	Sample Size N	Median Time (s)	IQR Time (s)
3	3000	0.550	0.139
6	3000	1.278	0.458

10	3000	2.499	0.544
3	5000	1.026	0.216
6	5000	2.169	0.135
10	5000	4.264	0.946
3	10000	2.020	0.373
6	10000	4.319	0.214
10	10000	5.194	2.841

Table 2.6: Median computational time per proposed POE test for larger-scale datasets. Each row is based on 50 null simulations assuming normal random error.

Sample Size N	K	Median Time (s)	IQR Time (s)
20000	3	3.173	2.082
20000	6	8.587	0.261
20000	10	13.497	7.966
40000	3	8.181	0.149
40000	6	9.431	0.563
40000	10	31.411	6.611

Table 2.7: Significant variants (41) identified by POIROT in UKB that were not identified by any univariate POE test or post-hoc GxE test.

rs	Chr	Pos	Ref	Alt	Function	POIROT p
rs4970829	1	109757295	A	G	intronic	4.9E-10
rs585362	1	109789795	T	C	intergenic	7.3E-08
rs11577931	1	109820884	G	A	intergenic	2.5E-21
rs55660224	1	109839400	T	C	intronic	4.9E-09
rs62104180	2	466003	A	G	intergenic	1.6E-10
rs74629722	2	21127044	C	G	intergenic	3.3E-10
rs550619	2	21260601	A	G	intronic	3.1E-10
Affx-20089987	2	44065090	A	G	exonic	4.1E-14
Affx-20090007	2	44066247	C	G	exonic	9.7E-16
rs41360247	2	44073656	C	T	intronic	3.0E-13
rs61789562	3	135926784	C	T	intergenic	1.1E-07
rs3128987	6	31434198	C	T	downstream	1.2E-07
rs9276689	6	32751962	T	C	intergenic	5.7E-08
rs74617384	6	160997118	T	A	intronic	2.5E-08

rs10455872	6	161010118	G	A	intronic	6.0E-08
rs471364	9	15289578	T	C	intronic	3.5E-10
rs643531	9	15296034	A	C	intronic	1.3E-10
rs12686004	9	107653426	A	G	intronic	1.2E-17
rs28927680	11	116619073	G	C	UTR3	5.7E-10
rs11825181	11	116626258	A	G	intronic	3.3E-10
rs11820589	11	116633862	A	G	exonic	3.0E-10
rs11216185	11	116782974	G	T	intronic	4.0E-09
Affx-52324980	11	117030633	TC	T	intronic	5.2E-08
rs45574931	11	117076972	A	C	exonic	1.4E-08
rs74580294	12	122622795	G	A	exonic	7.4E-08
rs4759377	12	123796849	T	C	intronic	3.5E-08
rs11057273	12	123814466	C	T	intronic	4.8E-08
rs28660993	12	123875394	T	C	intronic	5.9E-08
rs12445698	16	56928216	T	C	intronic	5.2E-08
rs8063291	16	56930251	C	T	intronic	3.3E-08
rs2217332	16	56969148	A	G	exonic	4.5E-10
rs12934552	16	57021433	G	A	intergenic	7.9E-14
rs11542916	19	10694720	A	G	exonic	4.7E-09
rs4425006	19	10813364	C	T	intronic	1.1E-12
rs73013159	19	11122710	T	G	intronic	2.4E-15
rs2228603	19	19329924	T	C	exonic	3.3E-13
rs7259004	19	45432557	C	G	ncRNA_intronic	1.7E-18
rs77617917	20	44563217	A	G	upstream	6.5E-15
rs2274755	20	44639692	T	G	intronic	7.5E-08
Affx-16780572	20	44643111	A	G	exonic	4.4E-08
rs10432735	20	44650318	T	A	upstream	7.1E-08

2.5.3 Figures

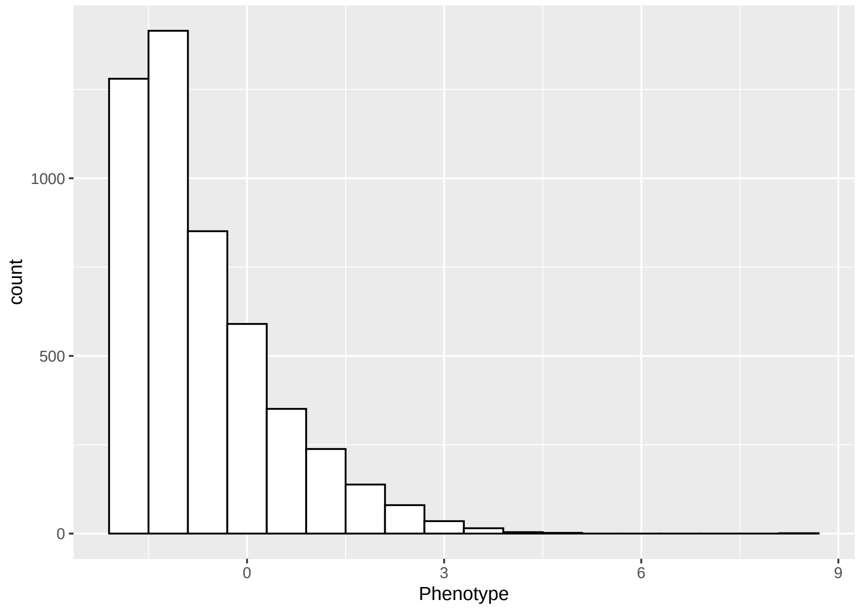

Figure 2.5: Histogram of example simulated non-normal phenotypes assuming skewness = 2 and excess kurtosis = 2. Data shown here corresponds to a sample size of 5,000 for a single phenotype with no parent-of-origin effects.

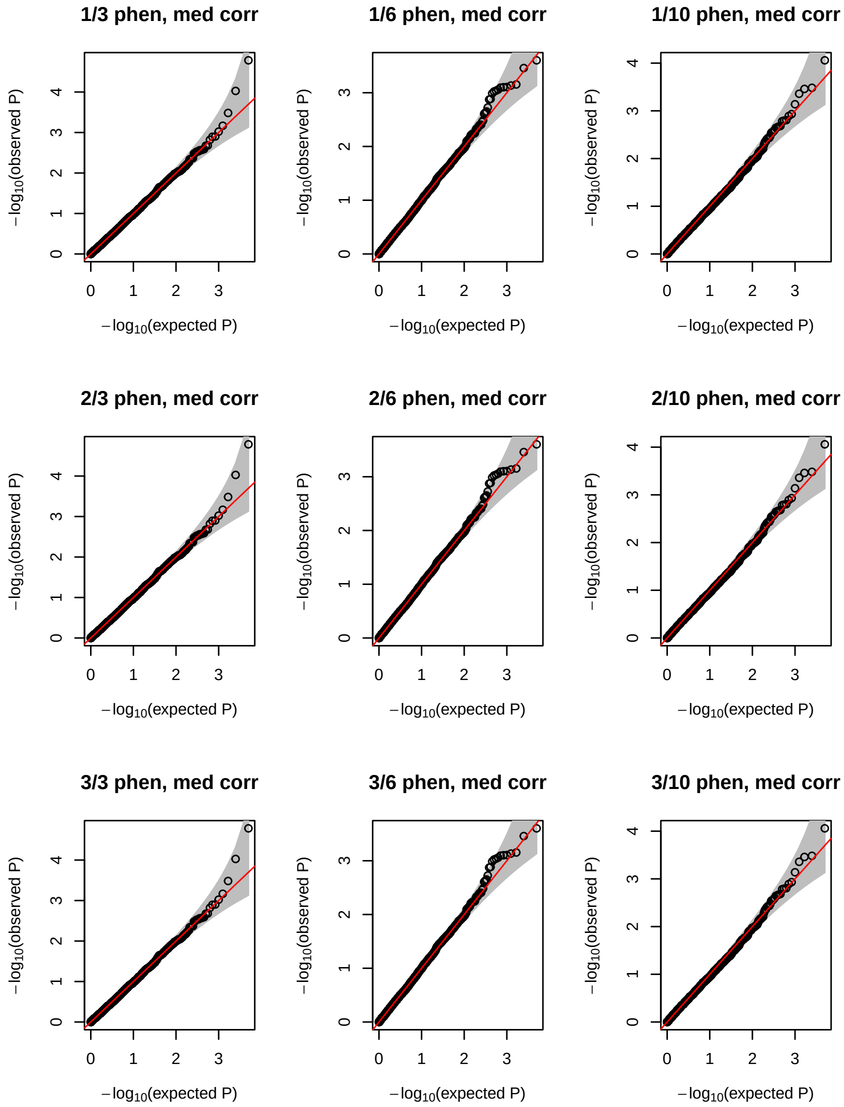

Figure 2.6: QQ plots of p-values for proposed post-hoc test for gene-gene or gene-environment interaction effects under the null hypothesis of no interactions effects but under the presence of POEs. Simulations used 10,000 individuals with 3 (left column), 6 (middle column) or 10 (right column) continuous normal phenotypes with medium correlation. MAF is assumed to be 0.25. Horizontal panels depict number of phenotypes with POE (1, 2, or 3), and maternal POE effect size of 0.75.

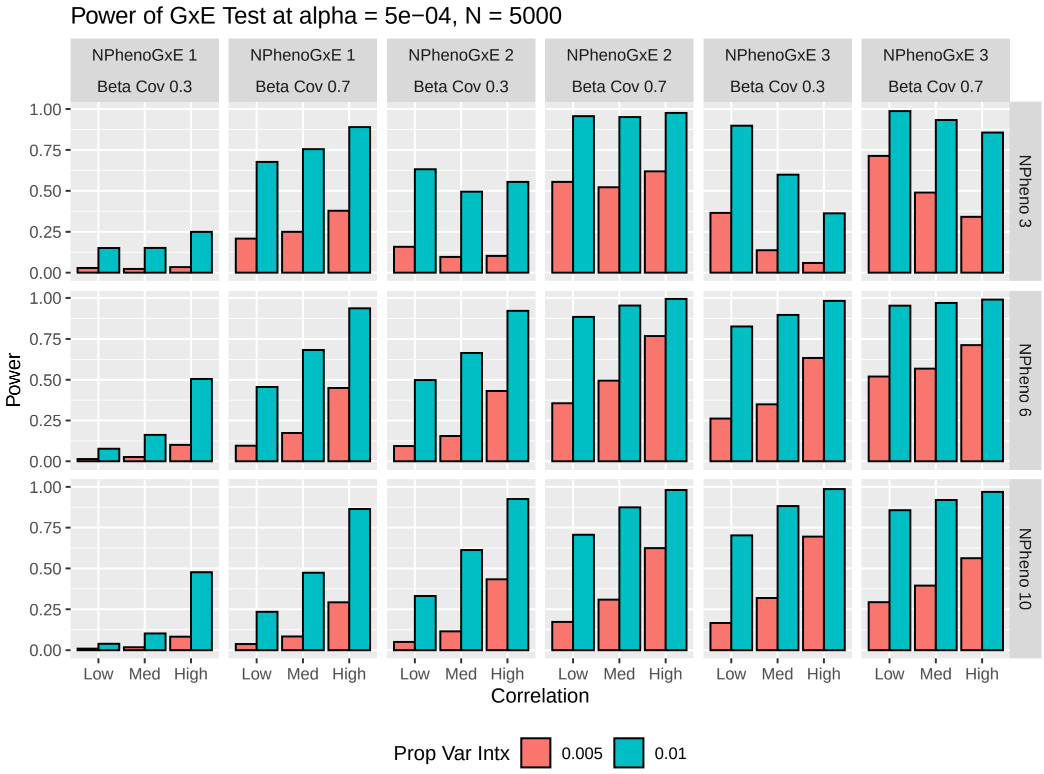

Figure 2.7: Power of post-hoc test for interaction effects assuming K = 3, 6, or 10 normal phenotypes (horizontal panels). We assume either 1, 2, or 3 of the phenotypes harbor gene-environment interaction effects at the locus with varying magnitude of covariate effect size (vertical panels). Color corresponds to proportion of phenotypic variation explained by interaction effects for an affected phenotype. We performed 5,000 simulations for each scenario. We calculated power at significance level 5×10^{-4}. We assume MAF = 0.25 and sample size = 5,000.

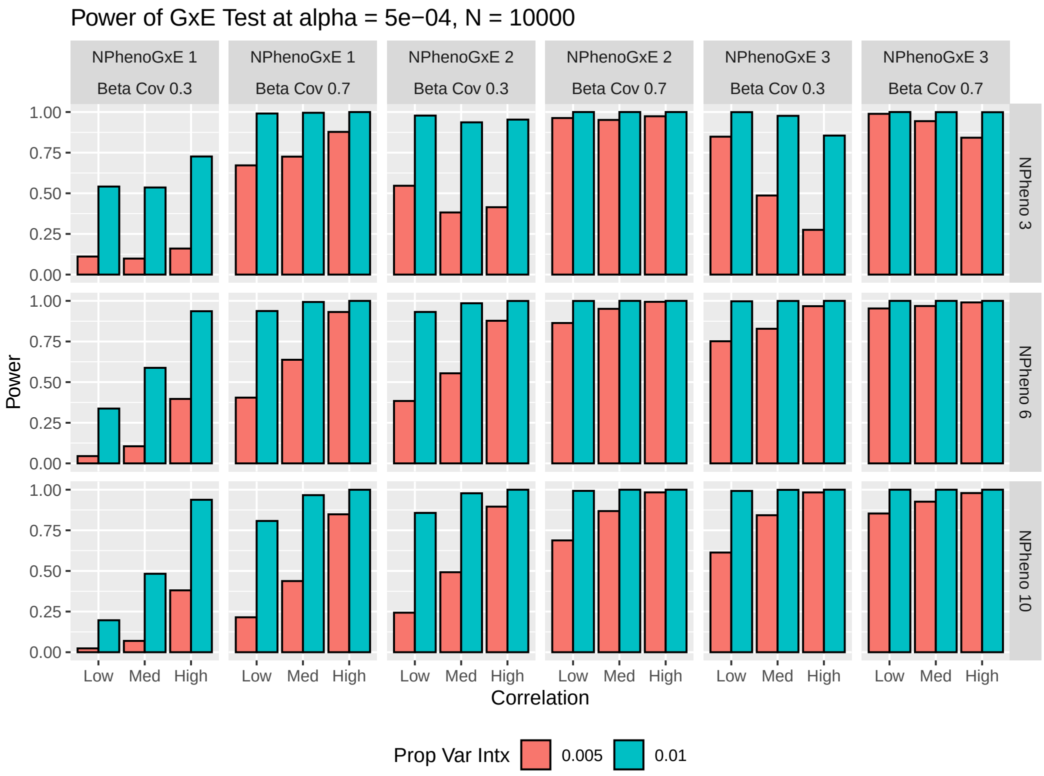

Figure 2.8: Power of post-hoc test for interaction effects assuming K = 3, 6, or 10 normal phenotypes (horizontal panels). We assume either 1, 2, or 3 of the phenotypes harbor gene-environment interaction effects at the locus with varying magnitude of covariate effect size (vertical panels). Color corresponds to proportion of phenotypic variation explained by interaction effects for an affected phenotype. We performed 5,000 simulations for each scenario. We calculated power at significance level 5×10^{-4}. We assume MAF = 0.25 and sample size = 10,000.

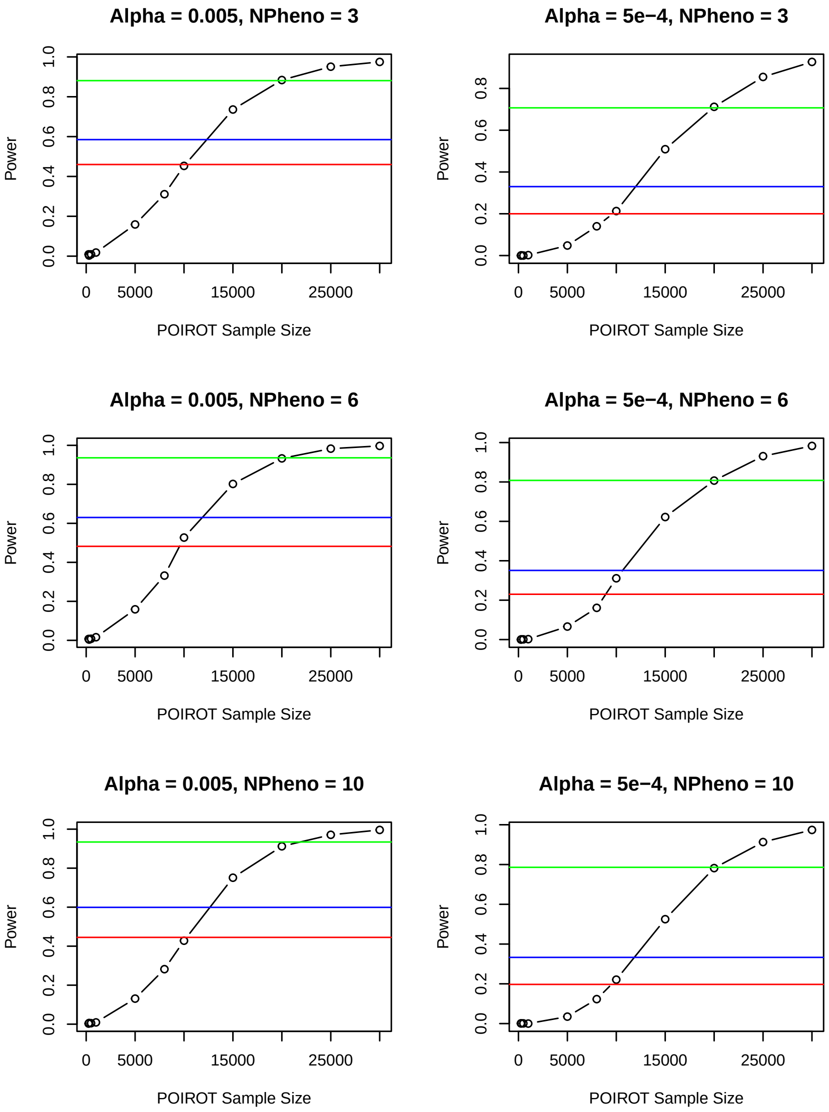

Figure 2.9: Power comparison of POIROT to a family-based approach using trio genotype data. Black line represents the power of POIROT at given sample size of unrelated individuals with no family genotype data (x-axis). Horizontal lines represent power of one-way MANOVA comparing phenotypic means of heterozygous offspring with maternally inherited minor allele versus heterozygous offspring with a paternally inherited minor allele. Green line represents family-based approach power at trio size 500, blue corresponds to trio size 300, and red represents trio size 250. Given trio size N, this corresponds to approximately 2(MAF)(1-MAF)N heterozygous offspring. Simulation parameters included MAF $= 0.25$, $\beta_M = 0.5$, normal error distribution, and medium pairwise phenotype correlation. Of the 3, 6, or 10 total phenotypes tested in each analysis, we assumed 2 harbored parent-of-origin effects.

Chapter 3

Topic 2. Cis- and trans-eQTL TWAS of breast and ovarian cancer identify more than 100 risk associated genes in the BCAC and OCAC consortia

3.1 Introduction

Both breast and ovarian cancer carry a significant global burden. The estimated numbers of new cases of female breast cancer and ovarian cancer each year exceed 2.2 million and 310,000, respectively [57]. Genome-wide association studies (GWAS) have identified a growing catalog of validated common risk variants for breast and ovarian cancer [58–65]. Further research has helped define risk variants that are unique to distinct subtypes of breast cancer (for example, hormone receptor positive tumors) and ovarian cancer (for example, high grade and low grade serous histotypes) [61, 62, 64, 66, 67]. While pleiotropic and subtype-specific GWAS have helped delineate the germline genetic architecture of these cancers, most GWAS-derived risk variants for complex traits lie in non-coding regions of the genome [15, 68]. This suggests that considerable disease risk may stem from variation in regulatory elements that affect gene transcription [69].

Transcriptome-wide association studies (TWAS) are a powerful approach to identifying genes that are associated with risks for complex diseases with genetic effects mediated through genetically regulated transcriptional activity. In a training dataset, TWAS studies first build a statistical regression model of gene expression in a specific tissue by selecting those genetic variants having non-zero effect sizes on gene expression; we refer to such genetic variants as expression quantitative trait loci (eQTLs) of a broad sense for that gene. Using these models, TWAS then imputes the genetically regulated expression (GReX) levels of the gene in a target GWAS dataset where transcriptomic data are absent but disease outcome data are available. TWAS then tests for association between imputed gene expression and phenotype. Equivalent TWAS tests also can be conducted using only GWAS summary data with estimated eQTL effect sizes from the expression imputation models. TWAS have successfully identified novel candidate susceptibility genes for not only overall breast cancer and ovarian cancer risk, but, more recently, for specific subtypes of breast and ovarian cancer [70–72].

To date, standard TWAS methods employ training models that only consider the regulatory effects of variants located in close proximity to the target gene (cis-SNPs) [73–79]. These variants reside within a small (e.g., 1Mb) window around the target gene. However, recent work has estimated the average proportion of heritability of gene expression estimated from mapped SNPs (mostly cis) to be modest, with reported values ranging between 0.2 and 0.38 [16]. One potential source for the remaining heritability of gene expression may be the aggregated effects of trans-eQTLs, which are defined as those variants that influence transcriptional activity that reside 1Mb or further away from the transcription start/end site of the target gene [16, 17]. With growing evidence of distal regulatory effects of common variants, Luningham et al. [21] developed a Bayesian genome-wide TWAS (BGW-TWAS) method, which trains expression prediction models considering both cis- and trans-

SNPs. This approach both improves prediction of GReX as well as enhances detection of genes that influence phenotype through distal transcriptional regulation.

In light of established shared genetic etiology between breast cancer and ovarian cancer [68, 80–84], here we apply BGW-TWAS to conduct TWAS of each disease (and various subtypes) that consider the regulatory activity of both distal and proximal germline variants for a target gene. We first constructed gene expression imputation models using BGW-TWAS in normal breast and ovarian tissue from the Genotype-Tissue Expression (GTEx) project and subsequently imputed GReX into large-scale GWAS summary data from the Breast Cancer Association Consortium (BCAC) and Ovarian Cancer Association Consortium (OCAC) to identify genes associated with risk of overall breast cancer, five breast cancer subtypes (luminal A-like, luminal B-like, luminal B/HER2-negative-like, HER2-enriched-like, triple-negative), non-mucinous ovarian cancer, and five ovarian cancer subtypes (high grade serous, low grade serous, endometrioid, mucinous, clear cell). Our findings replicate several established cancer risk loci and suggest several novel candidate trans-eQTL driven genes not discovered by a standard TWAS approach that models cis-SNPs only. We then used independent GWAS summary data and matched genotype/gene expression data in breast tissue from the Cancer Genome Atlas to validate several of our top identified genes. This work provides new insight into the eQTL genetic architecture of breast and ovarian tissue and leverages trans-genome regulation of expression in these tissues for improved TWAS of breast and ovarian cancer.

3.2 Materials and Methods

3.2.1 GTEx V8 Training Dataset

In order to train our imputation models for gene expression levels, we first obtained whole-genome sequencing (WGS) and RNA sequencing data on breast mammary tissue and ovarian tissue from the Genotype-Tissue Expression (GTEx) project V8 (dbGaP accession number phs000424.v8.p2). Data were available for 337 (125 female, 212 male) White individuals in breast tissue and 140 White females in ovarian tissue. We obtained gene expression levels as transcripts per million (TPM) per sample per tissue. We focused exclusively on autosomal genes for our analyses. We adjusted raw transcript data for the effects of age, body mass index, top five principal components to account for ancestry, and top probabilistic estimation of expression residuals (PEER) factors. In the gene expression data from mammary tissue, we also adjusted for Estrogen Receptor 1 (ESR1) expression in accordance with previous studies [68, 78]. This gene encodes estrogen receptor α, a transcription factor that plays a critical role in regulating gene expression and cell division in mammary glands [85].

3.2.2 Breast Cancer GWAS Summary Data

We obtained recently published summary-level GWAS data from BCAC (see Web Resources). The summary statistics for variant-level associations with breast cancer risk and risk of specific breast cancer subtypes were the result of a large multi-study GWAS of women of European ancestry [61]. The overall breast cancer analysis used genotype data from cases (invasive, in situ, unknown invasiveness) and controls across 82 BCAC studies that were genotyped using either the iCOGS or OncoArray Illumina genome-wide custom arrays. For this overall analysis, data from 11 other breast cancer GWAS were incorporated. This yielded a total sample size of 133,384 cases and 113,789 controls. The authors estimated SNP-disease associations using standard logistic regression, adjusting for country of origin and top principal components. Results were obtained for iCOGS subjects, OncoArray subjects, and additional GWAS separately and then combined via fixed-effects meta-analysis.

In addition to the summary statistics for the outcome of overall breast cancer risk, this study also published summary statistics for the association of variants with risk of specific intrinsic-like subtypes of breast cancer: luminal A-like cancer, luminal B-like cancer, luminal B/HER2-negative-like cancer, HER2-enriched-like, and triple-negative cancer. These subtype analyses were performed by fitting two-stage polytomous logistic regression models. Full details on these models are described elsewhere [86, 87]. Only invasive cases were considered for this analysis, and samples from the 11 additional GWAS were not included due to missing tumor marker information. The final sample for the GWAS subtype analyses included 106,278 cases and 91,477 controls.

Lastly, a meta-analysis was performed combining results from the analysis of triple-negative breast cancer cases from BCAC and the separate analysis of cases and controls with a pathogenic BRCA1 variant from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). The CIMBA participants were also of European ancestry. Authors performed a fixed-effects meta-analysis, combining odds ratio estimates from the BCAC study and hazard ratio estimates from the CIMBA study. As the majority of breast cancer cases in BRCA1 mutation carriers are triple-negative [88], our "triple-negative" TWAS results utilized summary statistics from this meta-analysis for greatest power. Quality control and imputation protocols for all sets of genotype data used in the study are described separately [58, 66, 89, 90]. Additional details on the study samples and statistical methods used in the breast cancer GWAS are provided by Zhang et al [61].

Once we obtained the summary data from this GWAS, we performed liftover to map these GWAS variants to Human Genome Assembly GRCh38. We then harmonized and imputed missing variants' summary statistics using the MetaXcan suite of tools (see Web Resources) and a European reference panel from the 1000 Genomes Project [91].

3.2.3 Ovarian Cancer GWAS Summary Data

We obtained summary-level GWAS data from a large-scale fine-mapping project of epithelial ovarian cancer [92]. This study used genotype data from six OCAC projects and two BCAC projects. The final sample contained a total of 26,151 ovarian cancer cases, 40,138 controls from OCAC, and 65,586 controls from BCAC. This project provided summary statistics for variant-level associations with risk of each of the five main subtypes of ovarian cancer: high grade serous ovarian cancer (HGSOC, 13,609 cases), low grade serous ovarian cancer (LGSOC, 2,749 cases), mucinous ovarian cancer (MOC, 2,587 cases), endometrioid ovarian cancer (ENOC, 2,877 cases), and clear cell ovarian cancer (CCOC, 1,247 cases). Additionally, the authors also performed a GWAS for the aggregate non-mucinous subtype, which excludes MOC cases. All participants were of European ancestry. Details on the genotyping and imputation procedures for this data are available elsewhere [92].

Authors used logistic regression models to generate association statistics for SNP genotypes and the five subtype outcomes, as well as the non-mucinous epithelial ovarian cancer analysis. For each analysis, authors fit separate models for OncoArray data, COGS data, and five additional GWAS datasets. Results from each group were then combined via fixed effects meta-analysis. Analyses adjusted for the effects of study of origin and possible population stratification by way of top principal components.

We performed liftover, harmonization, and summary statistic imputation as outlined for the breast cancer GWAS summary data.

3.2.4 Model Training and Association Test

After we obtained the training dataset from GTEx V8 individuals with both WGS and RNA-seq transcriptomic data, we proceeded to train genome-wide expression prediction models separately in breast mammary tissue (N = 337) and ovarian tissue (N = 140) using BGW-TWAS. To briefly summarize, the method first calculates genome-wide single variant eQTL summary statistics from the simple linear regression of adjusted expression against genotype at a given SNP. Cis-SNPs were defined as those within 1Mb of the flanking 5' and 3' ends of the target gene. Trans-SNPs were those falling outside of the 1Mb window. Genotype files were then segmented into approximately independent genome blocks using LDetect [93]. We excluded rare variants with minor allele frequency (MAF) $< 0.5\%$ from consideration.

Next, we pruned genome-wide blocks to select a subset of blocks containing cis-SNPs and a subset of trans blocks with a minimum single-variant eQTL p-value less than 0.00001 to fit the following Bayesian variable selection regression (BVSR) model:

$$\boldsymbol{E_g} = \boldsymbol{X}_{cis}\boldsymbol{w}_{cis} + \boldsymbol{X}_{trans}\boldsymbol{w}_{trans} + \boldsymbol{\epsilon} \tag{3.1}$$

Here, $\boldsymbol{E_g}$ is the vector of expression levels in the training (GTEx) dataset, $\boldsymbol{X}_{cis}$ and $\boldsymbol{X}_{trans}$ are the cis and trans genotype matrices from the pruned genome blocks, and the $\boldsymbol{w}$ terms correspond to the eQTL effect sizes of the considered SNPs. The BGW-TWAS model then assumes a spike-and-slab prior distribution for the eQTL effect sizes, allowing these distributions to be different for cis and trans SNPs. Using an adapted expectation-maximization Markov Chain Monte Carlo (EM-MCMC) algorithm, BGW-TWAS estimates $(\boldsymbol{w}, \boldsymbol{PP})$, where, for selected SNPs, $\boldsymbol{w}$ is the vector of eQTL effects and $\boldsymbol{PP}$ is the vector of posterior causal probabilities (PP) of the selected SNPs beings true eQTLs (with non-zero effect size). Only selected SNPs with estimated PP greater than 0.0001 were retained in the imputation models for each respective gene. We did not make any restrictions on the number of models in which a certain trans- or cis-variant can be included. If a variant is used to predict expression of more than one gene and has high probability of being a true eQTL for these genes, this does not impact the accuracy of the corresponding GReX models or interpretation. Full details on the statistical methodology and computational algorithms of BGW-TWAS are provided by Luningham et al [21].

Once we have trained the GReX imputation models, we then performed TWAS using the breast cancer and ovarian cancer GWAS summary statistics by calculating the following burden BGW-TWAS Z-score statistic for each gene g:

$$Z_g = \sum_{l \in Model_g} \frac{w_l^* \widehat{\sigma}_l Z_l}{\sqrt{\widehat{\boldsymbol{w}}' \boldsymbol{V} \widehat{\boldsymbol{w}}}} \tag{3.2}$$

Here, $w_l^* = \widehat{PP}_l \widehat{w}_l$, the expected eQTL effect size estimated from the BVSR model above. $\widehat{\sigma}_l$ is a reference-derived estimate of standard deviation (SD) of genotype data for variant l, Z_l is the variant's corresponding Z-score statistic from the GWAS, and $\boldsymbol{V}$ is the reference-derived covariance of the genotype data of selected SNPs for this gene. We used GTEx V8 samples with available genotype information as reference for $\widehat{\sigma}_l$ and $\boldsymbol{V}$. We calculated this burden BGW-TWAS Z-score test statistic for each of the six breast cancer GWAS phenotypes and the six ovarian cancer GWAS phenotypes to test for significant candidate risk genes.

Transcriptome-wide significant genes were those passing Bonferroni correction at $0.05/M$. Here, M is the total number of gene-level association tests performed across all six analyses for breast cancer and ovarian cancer, respectively. We note that this is a strict multiple test correction, as we expect the gene-level test statistics to be correlated across certain breast cancer phenotypes and ovarian cancer phenotypes (such that the number of effectively independent tests is smaller than M). We further compared the performance of BGW-TWAS in identifying breast and ovarian cancer susceptibility genes to sPrediXcan with pre-computed GTEx models that only consider cis-eQTLs [76]. We used the published GTEx V8 multivariate adaptive shrinkage in R (MASHR-M) models available at https://predictdb.org/. These models leverage information across multiple tissues and incorporate posterior causal probability of variants from fine-mapping. These models have been described in detail previously [94]. We applied sPrediXcan to all 12 cancer phenotypes described above.

3.2.5 Validation Analyses

Independent Genome-wide Association Dataset

To evaluate the robustness of our findings, we determined which genes identified by BGW-TWAS in the primary analyses using BCAC and OCAC GWAS data replicated in a similar analysis using summary statistic data from an independent GWAS of multiple cancers [95]. This GWAS aimed to identify common germline genetic variants associated with 18 types of cancer and interrogate possible pleiotropy among these identified variants. The study sample comprised of individuals of European ancestry from the UK Biobank (UKB) and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. A total of 17,881 breast cancer cases (13,903 UKB, 3,978 GERA) and 1,259 ovarian cancer cases (1,006 UKB, 253 GERA) were considered, as well as a total of 219,656 controls (189,855 UKB, 29,801 GERA). Following quality control, imputation was performed using 1000 Genomes reference data. Authors fitted cohort-specific logistic regression models. These models adjusted for the effects of age, top 10 PCs, and genotyping array and genotype reagent kit (where applicable). Results were then combined via meta-analysis for variants present in both cohorts. We obtained the publicly available summary statistics for the fixed-effect meta-analysis of breast and ovarian cancer (see Web Resources). Again, we performed variant harmonization and liftover to GRCh38 using the MetaXcan suite of tools. We repeated BGW-TWAS with this data for significant genes identified in the primary BCAC/OCAC analyses. We note that this validation GWAS data reflects odds ratio estimates for overall breast cancer and ovarian cancer, as the GWAS study did not include subtype-specific regression models.

GReX Prediction in Independent Breast Tissue Samples

As a second validation analysis, we investigated how the GReX imputation models trained by BGW-TWAS in GTEx data (normal breast tissue) would predict gene expression in an independent set of breast cancer cases. Specifically, we looked at prediction performance in tumor-adjacent normal breast tissue (NAT) and tumor tissue samples. We obtained individual-level germline genotype data and matched gene expression data in breast cancer cases from The Cancer Genome Atlas (TCGA). Genotypes were called from Affymetrix SNP Array 6.0. We restricted consideration to individuals of consensus European ancestry as defined by Carrot-Zhang et al. [96]. In this paper, authors performed four methods of ancestry determination in TCGA individuals. Consensus ancestry refers to the majority ancestry assignment across the employed methods. We also retrieved ancestral PCs for TCGA samples from this study. To maximize sample size in TCGA, we downloaded both blood-derived genotype data and solid normal tissue-derived (adjacent to tumor) genotype data. For individuals who had both, we preferentially used blood-derived genotypes as germline. We set as missing genotypes with birdseed confidence scores exceeding 0.1. We retained SNPs with call rate greater than 95% and samples with call rate greater than 95%. We aligned alleles to agree with Affymetrix SNP Array 6.0 annotation. We then excluded ambiguous SNPs and duplicates with identical chromosome and position and removed samples with high heterozygosity. We defined this when the absolute value of the inbreeding coefficient exceeded 0.2. We also pruned sample pairs with high estimated relatedness. We defined this by an estimated KING kinship coefficient exceeding 0.0884 (2nd degree relatedness or higher) [97]. We excluded SNPs with MAF < 0.005 and HWE $p < 1 \times 10^{-6}$. We aligned alleles with data from the 1000 Genomes. We performed imputation using the Michigan Imputation Server with 1000 Genomes Phase 3 V5 data and applied a Rsq threshold of 0.3. We then lifted variants over to GRCh38.

We downloaded RNA sequencing data (Illumina TruSeq) in NAT breast tissue and tumor tissue from TCGA. We ultimately had 786 individuals (779 females, 7 males) with germline genotype data, complete covariate information, and gene expression data in breast tumor samples. We had 101 individuals (100 female, 1 male) with complete genotype data, covariate data, and expression data in NAT breast tissue samples. We selected genes as having TPM ≥ 0.1 in 20% more of samples and those having 6 or more reads in 20% or more of samples. To calculate PEER factors, we then normalized read counts using the trimmed means of M values (TMM) method [98] and applied inverse normal transformation. These factors were calculated using `peer` package in R [99]. For both models (NAT and tumor), we determined the number of PEER factors to adjust for by examining the plot of the posterior variance of factor weights for a clear elbow. Expression levels in tumor were adjusted for the effects of age, first 10 PCs to account for ancestry, first 5 PEER factors, and ESR1 expression via linear regression. For NAT samples, we adjusted for age, first 3 PCs, first 3 PEER factors, and ESR1 expression.

Once all data were processed, we used our GTEx-derived cis- and trans-eQTL gene expression imputation models constructed for the main analysis to predict gene expression using TCGA germline variants. For genes that we had successfully trained a model for in Section 3.2.4 and were present in the TCGA expression datasets, we estimated the correlation between predicted GReX and observed, adjusted expression levels in both NAT and tumor samples. For this, we used Spearman's rank correlation coefficient.

BGW Modeling of Gene Expression in Breast Tumor Tissue

We were also interested in comparing the discovery yield of BGW-TWAS using GReX imputation models trained in normal breast tissue (GTEx) to imputation models trained in tumor tissue. Specifically, we trained genome-wide models of gene expression using BGW-TWAS and the adjusted tumor expression data and matched germline genotype data from TCGA described in the above section. We restricted model training to only those genes that were identified as significantly associated with one or more breast cancer phenotypes in the original BGW-TWAS analyses using BCAC summary statistics (Section 3.2.4). For this analysis, we further restricted our attention to females only (N = 779). We used the same BGW training protocol as described in Section 3.2.4. Once the models were trained, we similarly calculated the burden BGW-TWAS Z-score statistics for each of the six breast cancer BCAC GWAS phenotypes.

Colocalization of Top Trans-SNPs and GWAS Variants

For those genes that were identified as significantly associated with one or more cancer phenotypes in the main BGW-TWAS analysis (Section 3.2.4), we performed downstream colocalization tests using BCAC and OCAC GWAS loci. Guided by protocols in a recent study of trans-eQTLs [100], we first defined our set of highest confidence trans-eQTLs as those with BGW-TWAS $PP > 0.001$ and LRT $p < 0.05$. We then defined 200kb upstream and downstream of these variants as our "trans" regions. We performed two rounds of Bayesian colocalization analyses of each region/gene pair with minimum GWAS $p < 5 \times 10^{-5}$ using the R package `coloc` [101]. In the first round, we used LRT p-values of SNPs in the corresponding GReX imputation model detected in each trans region. In the second round, we used simple linear regression eQTL summary statistics of all SNPs in the trans region, regardless of whether they were selected in the GReX model of a given gene. We defined significant colocalizations as those with posterior probability of one common causal variant $(PPH4) > 0.75$.

3.3 Results

3.3.1 Fitted GReX Models and eQTL Architecture

In normal GTEx breast tissue samples, we successfully trained 24,833 autosomal gene expression models using Bayesian variable selection regression (BVSR) within BGW-TWAS. For each variant included in the BGW-TWAS models, we estimated the posterior probability (PP) of the variant being a true eQTL for the target gene. The sum of these PP estimates across all SNPs included as predictors in the fitted model corresponds to the estimated number of eQTLs per gene. These quantities can help improve our understanding of the location and distribution of eQTL effects genome-wide. For quality control, we excluded 102 outlier gene models with estimated PP summation exceeding 35 and/or maximum ($|w|$) (eQTL effect size) greater than 1000, indicating poor model fit. Across the remaining genes passing these quality control filters, the median number of SNPs with estimated $PP > 0.0001$ per model was 868 (interquartile range [IQR] = 1074). For all breast tissue models, the average training R^2 (squared correlation between imputed GReX and observed gene expression in the training GTEx dataset) was 0.30 (SD = 0.11). The median proportion of variants included as predictors per model that was located in trans to the target gene was 0.98 (IQR = 0.04). The median number of estimated eQTLs per gene was 1.14 (IQR = 2.04), while the median estimated eQTLs from trans regions across all breast models was 0.74 (IQR = 1.70). The distribution of estimated total eQTLs in breast tissue expression models is shown in Figure 3.4 (Appendix). The median (IQR) of estimated eQTLs located in trans regions on the same chromosome as the target gene was 0.02 (0.05). The median (IQR) of estimated eQTLs located in trans regions on a different chromosome from the target gene was 0.69 (1.60). This distribution indicates most SNPs used to model gene expression in the trans genome were located on a different chromosome than the target gene. The median genome-wide, cis-region, and trans-region estimates of total eQTLs according to model training R^2 are provided in Table 3.5 (Appendix).

In normal ovarian tissue samples from GTEx, we trained 22,584 autosomal GReX models using the BVSR framework of BGW-TWAS. All models had cumulative PP below 35, but 68 models with large maximum ($|w|$) were excluded. Across our fitted BGW-TWAS models, the number of genome-wide variants included as predictor variables was generally smaller than for the breast tissue models, with a median of 608 (IQR = 983). The median proportion of variants included as predictors per model that were located in trans to the target gene was 0.98 (IQR = 0.03). The median number of estimated eQTLs per gene was lower for genes in ovarian tissue at 0.40 (IQR = 1.27), while the median estimated eQTLs from trans regions across models was 0.30 (IQR = 1.14). The distribution of estimated total eQTLs in ovarian tissue expression models is also provided in Figure 3.5 (Appendix). The median (IQR) of estimated eQTLs located in trans regions on the same chromosome as the target gene was 0.01 (0.04). The corresponding median (IQR) of estimated eQTLs residing on a different chromosome as the target gene was 0.28 (1.06). The median genome-wide, cis-region, and trans-region estimates of total eQTLs according to model training R^2 are similarly provided in Table 3.5 (Appendix). The average training R^2 for fitted ovarian tissue models was 0.49 (SD = 0.13). We note that a larger proportion of genes had training R^2 exceeding 0.5 in ovarian tissue compared to breast tissue. This inflation is likely a result of limited ovarian tissue samples with RNA sequencing data for model training, as observed in previous simulations for this method [21].

3.3.2 Breast Cancer TWAS

We performed a total of 148,929 tests for BGW-TWAS across the six breast cancer phenotypes. This corresponds to a Bonferroni-adjusted p-value threshold of 3.36×10^{-7} for transcriptome-wide significance. We note again this correction is stringent as all tests are not expected to be independent across phenotypes. We further obtained sPrediXcan results based on cis-eQTLs only using the GTEx V8-derived MASHR models for 14,145 genes for each of the breast cancer phenotypes. We performed a total of 84,870 tests with sPrediXcan, corresponding to a Bonferroni threshold of 5.90×10^{-7}. Manhattan plots and quantile-quantile (QQ) plots of the BGW-TWAS results for the analysis of overall breast cancer risk and risk of the five subtypes of breast cancer are included in Figures 3.6-3.12 (Appendix). We see that the BGW-TWAS p-values do appear to suffer from inflation for the overall and luminal A phenotypes. As an illustrative example, for overall breast cancer, we have a genomic inflation factor of $\lambda = 1.34$, likely the result of the large GWAS sample sizes and phenotype polygenicity. Thus, as an alternative measure, we calculated the genomic inflation factor scaled to a study of 1000 cases and 1000 controls ($\lambda_{1000} = 1.003$) [102–104]. This measure was acceptable and reflective of all our BGW-TWAS results for breast cancer phenotypes.

Across the six TWAS performed for breast cancer phenotypes, BGW-TWAS identified 101 unique genes significantly associated with at least one of the six phenotypes considered. Location and phenotypes associated with these genes are shown in Figure 3.1. This figure also illustrates how the location of these genes relate to the location of significant GWAS variants from the original BCAC analysis of overall breast cancer risk. Using the `coloc` $PPH4$ cut-off of 0.75, we identified a sizable subset of genes/trans-eQTL regions with high probability of a single shared causal variant (eQTL and GWAS variant). For the 101 identified breast cancer genes, we found 21 harboring at least one significant colocalization with a trans region. It is important to further note that, although the GReX imputation models of these genes were trained in samples composed of both males and females, imputation accuracy (training R^2) of these genes were highly concordant when stratified by sex. Using our BGW-TWAS models of these 101 genes and observed gene expression levels in GTEx, we calculated sex-stratified training R^2 values (R^2_{female}, R^2_{male}) and compared them to the training R^2 we obtained originally using the combined samples (R^2_{all}). Training R^2_{female} is concordant with training R^2_{all} (correlation $r = 0.92, p < 2.2 \times 10^{-16}$), and training R^2_{male} is concordant with training R^2_{all} (correlation $r = 0.97, p < 2.2 \times 10^{-16}$). Further, the two sex-stratified training R^2 results are significantly correlated (correlation $r = 0.80, p < 2.2 \times 10^{-16}$). All results are shown in Figure 3.13 (Appendix).

We then performed a validation TWAS of these 101 genes using independent GWAS summary statistics of breast cancer from Rashkin et al [95]. For overall breast cancer risk, we identified 87 significant genes initially from BCAC and, of these, 31 further validated in the BGW-TWAS of Rashkin et al. data ($p < 0.05/101$). These 31 genes are provided in Table 3.1 along with sPrediXcan p-values. Of the 31 genes, 23 either could not be fit using MASHR-M models with sPrediXcan or failed to reach statistical significance in the cis-eQTL only approach. These 23 genes likely either failed to show a strong cis-eQTL in GTEx and thus did not have a publicly available MASHR-M model, or the associations were driven by trans-eQTL effects and thus were non-significant when modeled using a cis-eQTL-only approach.

For this overall analysis, the most significant gene identified by BGW-TWAS was ACAP3 on chromosome 1 (BCAC $p = 2.3 \times 10^{-34}$). We do note that this gene was found to be similarly associated with risk of luminal-A-like breast cancer (BCAC $p = 3.4 \times 10^{-34}$). The BVSR genome-wide model contained 18 cis-SNPs and 1,052 trans-SNPs. As the upper plot of Figure 3.2 illustrates, the SNPs with highest posterior probability of being eQTLs and largest expected eQTL weights lie on chromosomes 10 and 5, and several of these SNPs on chromosome 10 additionally have highly significant GWAS p-values (Figure 3.2, bottom plot). Given that the association of this gene is predominantly driven by trans-eQTL effects, it was not identified by sPrediXcan in either the overall analysis or the luminal-A-like analysis.

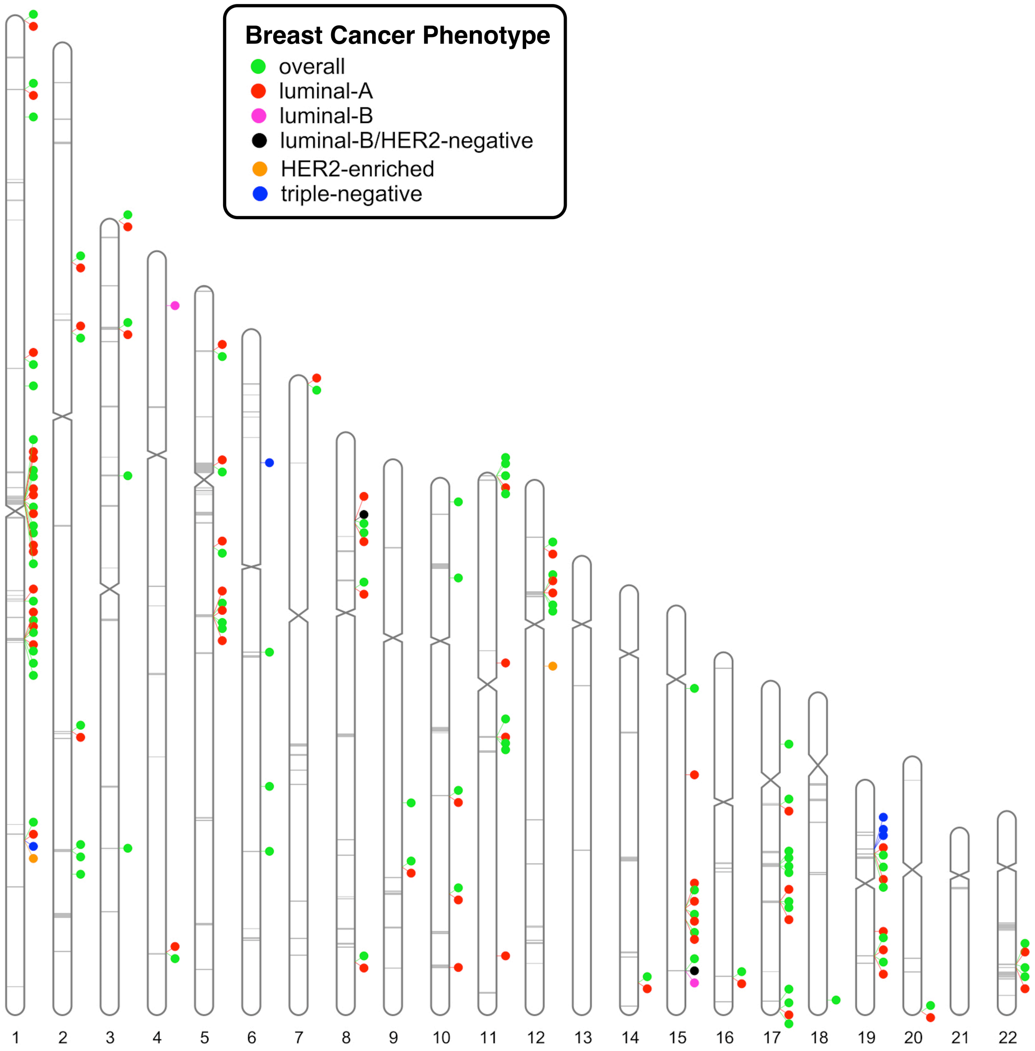

Figure 3.1: Ideogram of BCAC-derived BGW- TWAS results for overall breast cancer and breast cancer subtypes. 101 genes shown meet transcriptome-wide Bonferroni-adjusted p-value threshold for one or more phenotypes. Gray lines indicate position of genetic variants with BCAC GWAS $p < 5 \times 10^{-8}$ for association with overall breast cancer risk.

Table 3.1: BGW-TWAS identified genes associated with risk of overall breast cancer in BCAC analysis that validated in Rashkin et al. GWAS analysis (31).

		BGW				sPred				
		BCAC		Rashkin[a]		BCAC			TCGA $\rho^{b,c}$	
Gene	Chr	Z	p	Z	p	Z	p	TransPP	Tumor	NAT
ACAP3	1	12.23	2.3E-34	5.8	6.7E-09	-0.124	9.0E-01	0.996	0.031	-0.048
BLACAT1	1	9.14	6.5E-20	3.65	2.7E-04	0.523	6.0E-01	0.999	-0.081	0.135
ITGA10	1	6.91	4.7E-12	3.58	3.4E-04			0.995	0.064	-0.13
TMEM50A	1	-5.91	3.5E-09	-4.66	3.2E-06	-0.603	5.5E-01	0.992	0.031	-0.011
AC093110.3	2	-7.98	1.4E-15	-3.95	7.8E-05			0.882		
ALS2CR12	2	6.15	7.7E-10	5.06	4.1E-07	6.074	1.3E-09	0.705		
CYBRD1	2	-5.77	7.8E-09	-4.27	1.9E-05	-5.165	2.4E-07	0.3	0.058	**0.538**
KLF7-IT1	2	5.54	3.0E-08	4.39	1.1E-05			0.92	0.002	0.105
SLC4A7	3	10.44	1.7E-25	4.94	7.8E-07	1.914	5.6E-02	0.838	**0.116**	0.112
GLRA3	4	-8.09	6.0E-16	-3.96	7.4E-05			0.362	0.043	-0.051
RPS23	5	6.81	9.8E-12	5.15	2.7E-07	6.255	4.0E-10	0.042	**0.395**	**0.553**
ATP6AP1L	5	-6.8	1.1E-11	-5.14	2.8E-07	-5.586	2.3E-08	0.288	**0.296**	**0.533**
MRPS30	5	6.17	6.9E-10	3.74	1.8E-04	9.24	2.5E-20	0.804	**0.1**	0.066
ATG10	5	-6.14	8.0E-10	-4.48	7.6E-06	-4.835	1.3E-06	0.383	**0.474**	**0.608**
NUDT1	7	6.18	6.3E-10	6.06	1.4E-09	0.072	9.4E-01	0.918	-0.065	-0.028
NPM2	8	-9.48	2.5E-21	-5.32	1.0E-07	-0.276	7.8E-01	0.999	0.036	-0.176
EFR3A	8	-7.51	6.2E-14	-4.63	3.6E-06	-0.555	5.8E-01	0.988	-0.014	0.085
RP11-723D22.3	8	5.33	9.9E-08	3.53	4.2E-04			0.995		
IL2RA	10	-6.73	1.7E-11	-3.71	2.1E-04	-1.084	2.8E-01	0.93	-0.009	-0.169
RP11-165A20.3	10	6.66	2.8E-11	4.7	2.7E-06			0.999		
PIDD1	11	7.33	2.3E-13	3.79	1.5E-04	7.249	4.2E-13	0.25	**0.083**	0.189
CCDC91	12	-6.78	1.2E-11	-3.64	2.7E-04	-4.391	1.1E-05	0.641	0.035	0.178
RCCD1	15	-8.58	9.5E-18	-4.85	1.2E-06	-8.208	2.3E-16	0.023	**0.111**	**0.345**
RP11-467L19.16	15	-5.87	4.4E-09	-4.06	4.8E-05			1		
LINC02210	17	-5.99	2.1E-09	-6	2.0E-09			0.501	**0.322**	**0.686**
RP11-259G18.1	17	-5.93	3.0E-09	-5.9	3.6E-09			0.002		
KANSL1-AS1	17	-5.5	3.8E-08	-5.99	2.1E-09			0.508	**0.583**	**0.699**
CTD-3157E16.1	17	-5.42	6.1E-08	-4.02	5.7E-05			0.799		
PLEKHM1	17	-5.2	2.0E-07	-5.79	7.1E-09			0.089	**0.231**	**0.367**
LINC00683	18	8.78	1.6E-18	3.92	9.0E-05	-1.967	4.9E-02	0.994	0.017	-0.151
APOBEC3B	22	-5.93	3.1E-09	-4.45	8.7E-06	-6.871	6.4E-12	0.718	**0.101**	0.137

[a] BGW-TWAS results using summary statistics from independent GWAS of overall breast cancer.

[b] Correlation between observed expression and predicted GReX by BGW-TWAS model.

[c] Bold values indicate significant, positive correlation ($p < 0.05$).

The SNP with the highest estimated probability of being an eQTL for ACAP3 was rs1268974 on chromosome 10, and we further observe that many of the SNPs most likely to be eQTLs on chromosome 10 are intron variants for ZMIZ1. We see significant colocalization of eQTL signal in GTEx with breast cancer GWAS risk loci in this region (Chr10:78892621-79292621, Figure 3.14-3.15, $PPH4 = 0.99$ for overall phenotype). Notably, rs1268974 (eQTL PP = 0.1153) has been associated with breast cancer in European ancestry [58]. Another top predictor at this locus is rs704010 (PP = 0.0394), which has been associated with overall breast cancer in European ancestry [58] and Han Chinese [105] populations, as well as overall and ER-/PR- breast cancer in an African American cohort [106]. However, we do not see replication of this trans-signal in tumor or NAT tissue (Appendix Figures 3.16-3.17).

For our overall breast cancer analysis, several of the other top BGW-TWAS genes for this phenotype are supported by the literature. The majority of genes (64/87) lie within 1Mb of one or more curated breast-cancer associated variants, 25 of which were identified by sPrediXcan [107]. However, we consider 10 of the BGW-TWAS genes to represent potentially novel candidate risk loci, as they do not lie near these breast cancer variants or even a set of candidate susceptibility loci for ovarian cancer [64]. The expression models fitted for these genes (ACAP3, CTD-3157E16.1, EFR3A, KLF7-IT1, LINC00683, NPM2, NUDT1, RP11-467L19.16, RP11-723D22.3, TMEM50A) all show predominantly trans-SNP effects on regulation. The proportion of estimated eQTLs from trans regions exceeds 0.79 for each gene, and none were identified by sPrediXcan.

Beyond ACAP3, we identified 60 additional genes in the subtype-specific analysis of luminal A-like cancer using BGW-TWAS with BCAC data. Of these, 53 were also among those identified above for overall breast cancer risk. It is important to note that the independent GWAS summary data that we used for validation analyses of our BCAC-derived findings did not distinguish cases by cancer subtype. As such, we are

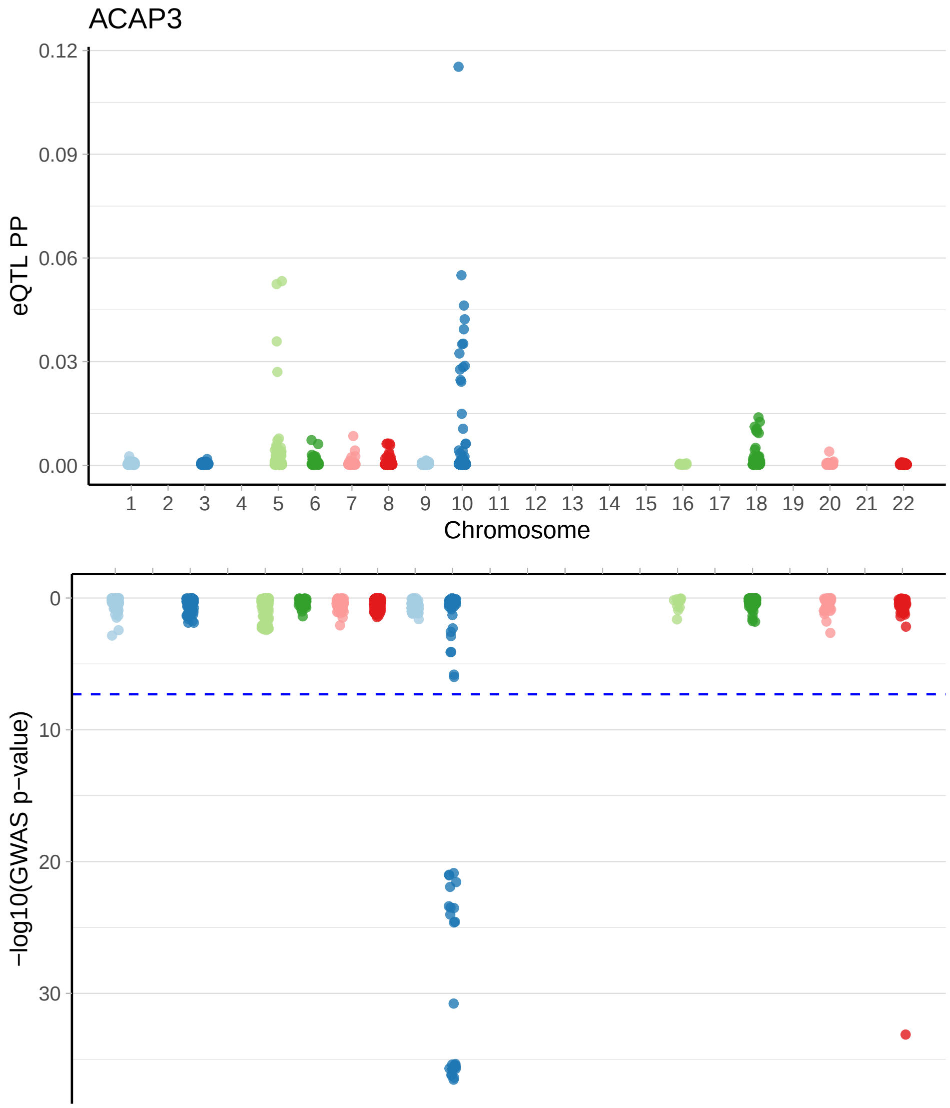

Figure 3.2: Estimated posterior probability (PP) of non-zero eQTL effects sizes from BGW-TWAS -selected SNPs for ACAP3 on chromosome 1 in breast tissue (top), and negative logarithm of the overall breast cancer GWAS p-values for these selected SNPs (bottom). Blue dotted line indicates genome-wide significance threshold for GWAS (5×10^{-8}).

less powered to validate certain subtype-specific associations, particularly for those of rarer cancer types. However, given the high prevalence of luminal A-like cancers globally, we assume the GWAS phenotype of the validation data (overall cancer) to be a reasonably proxy for the luminal-A-like phenotype. In Table 3.2, we present the 18 luminal-A-like genes that further validated when we applied our BGW-TWAS models to the Rashkin et al. data ($p < 0.05/101$). Of these 18 genes, ACAP3, EFR3A, NPM2, NUDT1, and RP11-723D22.3 are also not in regions of candidate breast cancer susceptibility loci [107].

Table 3.2: BGW-TWAS identified genes associated with risk of luminal-A-like breast cancer in BCAC analysis that validated in Rashkin et al. GWAS analysis (18).

		BGW				sPred				
		BCAC		Rashkin[a]		BCAC			TCGA $\rho^{b,c}$	
Gene	Chr	Z	p	Z	p	Z	p	TransPP	Tumor	NAT
BLACAT1	1	5.42	6.1E-08	3.65	2.7E-04	0.87	3.8E-01	0.882	-0.081	0.135
ITGA10	1	7.14	9.4E-13	3.58	3.4E-04			0.996	0.064	-0.13
ACAP3	1	12.19	3.4E-34	5.8	6.7E-09	-0.2	8.4E-01	0.718	0.031	-0.048
AC093110.3	2	-6.2	5.8E-10	-3.95	7.8E-05			0.383		
CYBRD1	2	-5.59	2.3E-08	-4.27	1.9E-05	-5.95	2.7E-09	0.288	0.058	**0.538**
SLC4A7	3	9.88	5.1E-23	4.94	7.8E-07	2.11	3.5E-02	0.999	**0.116**	0.112
GLRA3	4	-8.96	3.2E-19	-3.96	7.4E-05			0.641	0.043	-0.051
ATP6AP1L	5	-6.08	1.2E-09	-5.14	2.8E-07	-5.1	3.4E-07	0.3	**0.296**	**0.533**
ATG10	5	-5.53	3.3E-08	-4.48	7.6E-06	-4.8	1.6E-06	0.988	**0.474**	**0.608**
RPS23	5	6.17	6.9E-10	5.15	2.7E-07	5.81	6.2E-09	0.362	**0.395**	**0.553**
MRPS30	5	6.63	3.3E-11	3.74	1.8E-04	10.7	1.1E-26	0.995	**0.1**	0.066
NUDT1	7	6.07	1.3E-09	6.06	1.4E-09	-1.7	8.9E-02	0.804	-0.065	-0.028
NPM2	8	-8.72	2.9E-18	-5.32	1.0E-07	-0.23	8.2E-01	0.999	0.036	-0.176
EFR3A	8	-6.84	8.0E-12	-4.63	3.6E-06	-0.07	9.4E-01	0.918	-0.014	0.085
RP11-723D22.3	8	5.19	2.1E-07	3.53	4.2E-04			0.25		
PIDD1	11	5.33	9.7E-08	3.79	1.5E-04	5.16	2.5E-07	0.995	**0.083**	0.189
CCDC91	12	-5.4	6.8E-08	-3.64	2.7E-04	-3.57	3.6E-04	0.042	0.035	0.178
APOBEC3B	22	-5.41	6.2E-08	-4.45	8.7E-06	-6.1	1.1E-09	0.838	**0.101**	0.137

[a] BGW-TWAS results using summary statistics from independent GWAS of overall breast cancer.
[b] Correlation between observed expression and predicted GReX by BGW-TWAS model.
[c] Bold values indicate significant, positive correlation ($p < 0.05$).

Results from the TWAS of all other breast cancer subtypes (luminal B-like, luminal B/HER2-negative-like, HER2-enriched-like, and triple-negative) are shown in Table 3.3. We do not restrict this table to those genes that additionally had a significant association using the validation data since, here, we do not expect an independent GWAS of overall breast cancer to be an ideal validation study for these subtype-specific findings. However, the corresponding p-values from application of

BGW-TWAS to the validation summary data are provided in Table 3.3. The subtype-specific risk genes we identified also reflect prior candidate loci. For example, previous work has suggested that BLACAT1 has a role in breast cancer metastasis [108]. Expression of RCCD1 on chromosome 15 has also been shown to be associated with breast cancer risk in a large trans-ethnic TWAS [109]. However, BOD1L1 (luminal B-like), NPM2 (luminal B/HER2-negative-like), RP11-474P2.6 (HER2-enriched-like) and RPS18 (triple-negative) do not fall within 1Mb of curated candidate breast cancer risk loci [107]. For a global comparison of findings across all our analyses in breast cancer, we provide a correlation plot of BGW-TWAS Z-scores across all genes between subtypes in Figure 3.18 (Appendix).

Table 3.3: BGW-TWAS identified genes associated with specific risk of other breast cancer subtypes in BCAC analysis with corresponding PrediXcan and validation results.

				BGW p		sPred p	TCGA $\rho^{b,c}$	
Subtype	Gene	Chr	TransPP	BCAC	Rashkin^a	BCAC	Tumor	NAT
Luminal B	BOD1L1	4	0.217	2.7E-09	3.5E-01	2.1E-09	0.05	0.158
	RCCD1	15	0.023	8.3E-08	1.2E-06	2.5E-07	**0.111**	**0.345**
Luminal B/HER2-neg	NPM2	8	0.999	3.8E-10	1.0E-07	5.9E-01	0.036	-0.176
	RCCD1	15	0.023	1.6E-12	1.2E-06	1.9E-12	**0.111**	**0.345**
HER2-enriched	BLACAT1	1	0.999	2.6E-08	2.7E-04	3.6E-01	-0.081	0.135
	RP11-474P2.6	12	0.998	1.1E-07	7.5E-01	6.4E-01		
Triple-negative	BLACAT1	1	0.999	4.5E-10	2.7E-04	6.1E-01	-0.081	0.135
	RPS18	6	0.391	5.9E-10	7.7E-01	1.4E-12	**0.114**	0.106
	ANKLE1	19	0.378	1.3E-41	4.1E-03	4.2E-30	9.62E-05	-0.018
	OCEL1	19	0.19	2.1E-14	4.1E-03	9.7E-12	**0.205**	**0.303**
	ABHD8	19	0.001	4.7E-09	7.4E-01	1.3E-20	-0.015	**0.282**

[a] BGW-TWAS results using summary statistics from independent GWAS of overall breast cancer.

[b] Correlation between observed expression and predicted GReX by BGW-TWAS model.

[c] Bold values indicate significant, positive correlation ($p < 0.05$).

For breast cancer genes, we additionally used the GTEx-derived models of genetically regulated gene expression to predict gene expression levels in tumor tissue samples of breast cancer patients from TCGA, as well as normal tumor-adjacent breast tissue samples. This analysis was performed to gauge how well the models trained in GTEx among individuals without breast cancer accurately predict transcription in individuals with breast cancer. Tables 3.1 -3.3 provide an estimate of the correlation between imputed gene expression in TCGA samples that used the GTEx-derived BGW-TWAS models and observed expression levels in both tumor and NAT breast tissue. Among the genes from the overall and luminal A-like analyses that validated using external GWAS data (Table 3.1-3.2), seven (RPS23, ATP6AP1L, ATG10, RCCD1, KANSL1-AS1, LINC02210, PLEKHM1) showed nominally significant correlation between imputed and observed expression levels in both NAT and tumor TCGA samples ($p < 0.05$), and the estimated Spearman correlation coefficient was positive between the two vectors. For Table 3.1 and 3.2, we saw validation of SLC4A7, MRPS30, PIDD1, and APOBEC3B in tumor tissue but not NAT. Further, the BGW model of CYBRD1 showed significant correlation between predicted and observed expression in NAT tissue only. When looking at the genes identified by BGW-TWAS in the remaining subtypes (Table 3.3), we further saw RPS18, ABHD8, and OCEL1 (triple-negative) validate in tumor only, NAT only, and both tissue types, respectively.

Additionally, we performed a second round of BGW-TWAS with the six breast cancer phenotypes using genome-wide cis- and trans-eQTL models trained using tumor expression data of female cases in TCGA. Of the 101 genes originally identified in the GTEx-derived analyses, 67 had available expression data in TCGA and successfully passed BGW-TWAS model QC (PP summation below 35 and maximum $(|w|) < 1000$). Here, we identified 15 genes that further showed significant association ($p < 0.05/[6 \times 67]$) with one or more breast cancer phenotypes using these tumor-derived models. 11 and 10 genes showed significant association with overall and luminal A-like cancer, respectively. KLHDC7A, which was identified originally for overall and luminal A-like cancer, was similarly identified in the subsequent tumor analysis for not only these two phenotypes, but also the luminal B and luminal B/HER2-negative-like subtypes.

3.3.3 Ovarian Cancer TWAS

We performed a total of 135,474 tests using BGW-TWAS across the six ovarian cancer phenotypes. This corresponds to a Bonferroni threshold of 3.69×10^{-7} for transcriptome-wide significance. We obtained sPrediXcan results for 13,109 genes for each of the six ovarian cancer phenotypes, corresponding to a total of 78,654 tests and a Bonferroni threshold of 6.36×10^{-7}. We present genome-wide findings for non-mucinous ovarian cancer and the five main subtypes of ovarian cancer using BGW-TWAS and OCAC-derived GWAS summary statistics in Figures 3.19-3.25 (Appendix). The BGW-TWAS p-values do show evidence of some inflation for the non-mucinous and high grade serous phenotypes. However, as with the breast cancer results, this is corrected when considering the GWAS sample size used in construction of the test statistics ($\lambda = 1.06, \lambda_{1000} = 1.002$ for NMOC).

Eight unique significant genes were identified by BGW-TWAS when applied to the summary statistic data on risk of NMOC and HGSOC (Table 3.4). No genes meeting the multi-trait adjusted Bonferroni threshold were identified for LGSOC, EOC, MOC, or CCOC. A correlation plot of BGW-TWAS Z-scores across all genes between ovarian cancer subtypes is included in Figure 3.26 (Appendix). All eight significant genes had model training $R^2 > 0.1$ in ovarian tissue. Of the eight genes shown, sPrediXcan fit models for ANKLE1 and CCDC106. The most significant gene identified by BGW-TWAS across all analyses was ANKLE1 at 19p13 for HGSOC (BCAC $p = 4.4 \times 10^{-21}$), but it was also identified in the non-mucinous analysis (BCAC $p = 8.25 \times 10^{-13}$). ANKLE1 is a well-established candidate susceptibility locus for both breast cancer and ovarian cancer [83, 110, 111]. In the sPrediXcan model of ANKLE1, three SNPs were used to model expression, one of which (rs67412075) was also included in the BGW-TWAS model for ANKLE1 in ovarian tissue (PP = 0.0169). Although 389 out of 448 selected SNPs were located in trans to ANKLE1, all SNPs driving the association (with highest eQTL PP) are cis-SNPs located on chromosome 19, and thus it is not surprising that this gene was similarly identified by sPrediXcan in both phenotypes.

Table 3.4: Significant genes identified by BGW-TWAS for non-mucinous and high grade serous ovarian cancer in OCAC analyses with validation p-value for overall ovarian cancer using Rashkin et al. GWAS summary statistics.

			NMOC		HGSOC		Rashkin[a]	
Gene	Chr	TransPP	Z	p[b]	Z	p[b]	Z	p[b]
RP11-455G16.1	4	1	-7.43	**1.1E-13**	-8.53	**1.5E-17**	-1.58	1.1E-01
PRC1-AS1	15	0.17	4.91	9.3E-07	5.8	**6.7E-09**	0.72	4.7E-01
LRRC37A2	17	0.78	4.71	2.5E-06	5.28	**1.3E-07**	-0.13	9.0E-01
NSF	17	0.06	-4.31	1.7E-05	-5.3	**1.1E-07**	0.31	7.6E-01
ANKLE1[c]	19	0.39	7.16	**8.3E-13**	9.42	**4.4E-21**	0.96	3.4E-01
CCDC106	19	1	-7.61	**2.8E-14**	-7.1	**1.2E-12**	-2.47	1.4E-02
ZNF551	19	1	-4.92	8.6E-07	-5.25	**1.5E-07**	-0.86	3.9E-01
MLLT10P1	20	0.93	-6.2	**5.8E-10**	-5.43	**5.6E-08**	-2.78	**5.0E-03**

[a] BGW-TWAS results using summary statistics from independent GWAS of overall ovarian cancer.
[b] Bold indicates statistically significant results.
[c] Identified by PrediXcan for both HGSOC and NMOC.

In contrast, another highly significant gene from the BGW-TWAS analysis of both NMOC and HGSOC that appears to be largely driven by trans-eQTL effects is CCDC106 at 19q13. The BVSR genome-wide model had 21 cis-SNPs and 2016 trans-SNPs. As we can see in the top plot of Figure 3.3, the SNPs with highest PP of being eQTLs are located in trans on chromosomes 17 and 15. The bottom plot of Figure 3.3, which shows the corresponding OCAC GWAS p-values for these SNPs in the non-mucinous analysis, indicates that a subset of SNPs on chromosome 17 furthermore have the most significant GWAS associations for this phenotype. Therefore, this gene has not been implicated in previous cis-only TWAS of ovarian cancer. We see significant colocalization of eQTL signal in GTEx with GWAS loci in this region (Chr17:48211237-48611237, Figure 3.27-3.28, $PPH4 = 0.87$ for HGSOC phenotype). CCDC106 was not identified as significant by sPrediXcan for either HGSOC or NMOC ($p > 0.2$), as the association appears to be driven by distal GWAS-significant loci on chromosome 17 correlated with CCDC106 expression. The top SNP by eQTL PP is rs1979858 on chromosome 15, an intergenic variant for ARRDC4. The second top SNP is rs9898988 on chromosome 17, an intron variant for SKAP1. Intron variants of SKAP1 have previously been associated with overall epithelial ovarian cancer [64, 112, 113] and HGSOC [64].

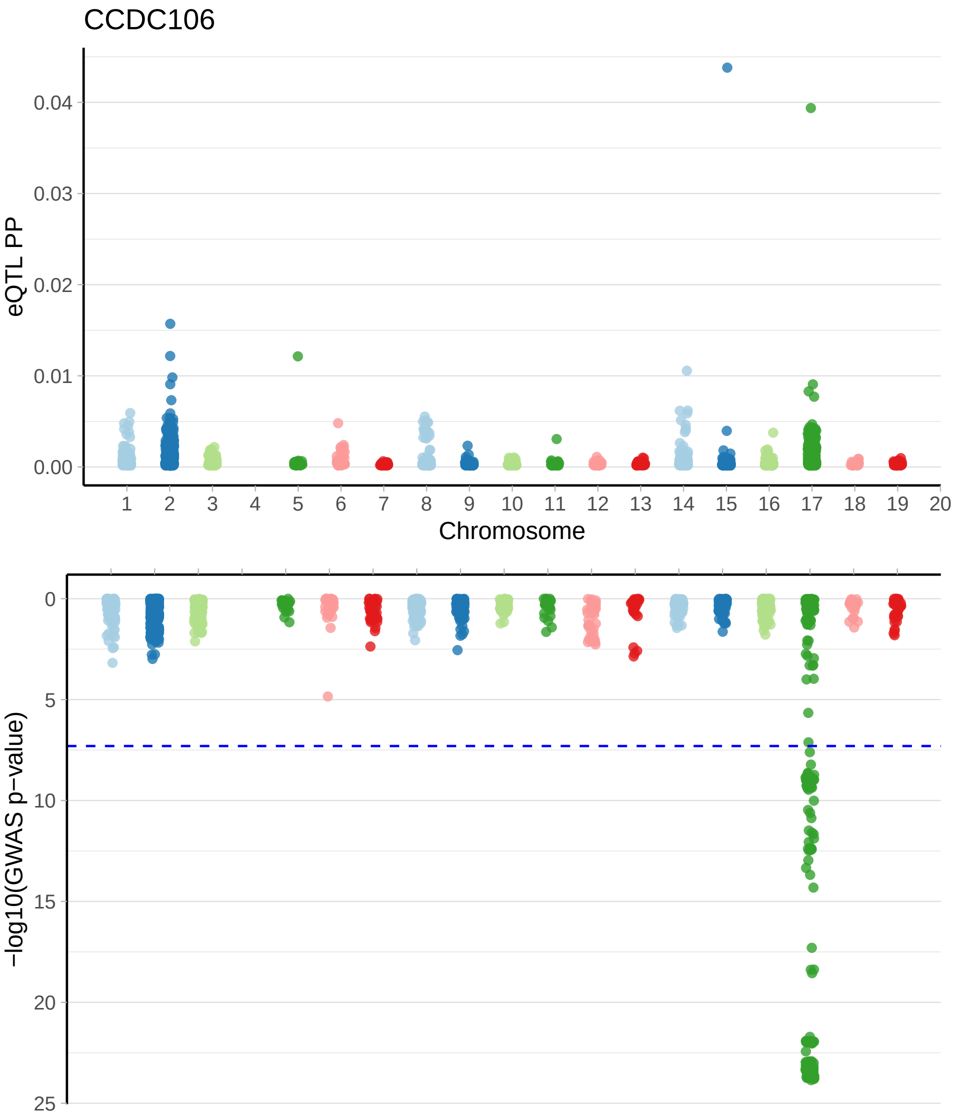

Figure 3.3: Estimated posterior probability (PP) of non-zero eQTL effects sizes from BGW-TWAS -selected SNPs for CCDC106 on chromosome 19 in ovarian tissue (top), and negative logarithm of the non-mucinous ovarian cancer GWAS p-values for these selected SNPs (bottom). Blue dotted line indicates genome-wide significance threshold for GWAS (5×10^{-8}).

Other top genes identified by BGW-TWAS have been implicated by previous work. For example, PRC1-AS1 was identified for HGSOC and is a candidate risk locus for breast cancer previously identified by GWAS in Europeans and East Asians [114, 115] and also identified by a previous cis-only TWAS of ovarian cancer [78]. Additionally, for the HGSOC analysis, our genome-wide model implicates NSF, which has documented GWAS and TWAS associations with risk of ovarian cancer [78, 110]. For R11-455G15.1, the top SNP by eQTL PP was rs10738466 on chromosome 9. This SNP is 32,525 bp to BNC2, a locus already implicated in ovarian cancer [113, 116–118].

However, of the eight ovarian cancer genes identified across phenotypes, LRRC37A2, CCDC106, ZNF551, and MLLT10P1 do not lie within 1 Mb of reported ovarian susceptibility loci [64]. CCDC106 and MLLT10P1 are additionally not within regions of curated breast cancer susceptibility loci [107]. For MLLT10P1, the top SNP most likely to be an eQTL was rs2229304 on chromosome 17. This is a missense variant for HOXB23. MLLT10P1 was additionally the only gene to validate when BGW-TWAS was applied to independent GWAS summary statistics for overall ovarian cancer risk by Rashkin et al. ($p < 0.05/8$). We note that failure of all other genes to replicate is likely due in part to the limited sample size for ovarian cancer cases in this validation GWAS itself (1,006 UKB, 253 GERA). These counts are considerably smaller than the corresponding sample size of breast cancer cases used (13,903 UKB, 3,978 GERA) and overall controls used (189,855 UKB, 29,801 GERA). Although only one gene achieved significance in this validation, six out of the eight genes originally identified by BGW-TWAS with OCAC summary data showed the same estimated direction of effect in the follow-up validation analysis (Table 3.4). For the two genes with differing effect directions between the OCAC and Rashkin et al. analyses, we emphasize that the latter Z-scores were close to zero ($p > 0.7$).

3.4 Discussion

In this work, we conducted the first TWAS of breast cancer and ovarian cancer that uses not only cis-SNPs, but both intra- and inter-chromosomal trans-SNPs, to model genetically regulated transcription. This genome-wide modeling approach, BGW-TWAS, stems from the growing catalog of trans-eQTL effects identified across a wide range of tissue types [17, 20, 119, 120]. We applied this method to train gene expression models in GTEx mammary and ovarian tissues and tested for association with risk of breast and non-mucinous ovarian cancer using summary GWAS data from recent large-scale meta-analyses. We further investigated how the landscape of identified risk genes for these diseases varied across cancer subtypes. We identified 101 significant genes across the overall and subtype-specific breast cancer analyses and eight for the corresponding ovarian cancer analyses.

Many of these genes have been implicated in recent cis-only TWAS of breast cancer and ovarian cancer [78, 121]. Of our eight ovarian cancer genes, four (50%) were found in these previous studies (PRC1-AS1, LRRC37A2, NSF, ANKLE1), and 36 of 101 (36%) breast cancer genes were similarly identified. However, several genes appear to be potentially novel associations that are driven largely by trans-eQTL effects. ACAP3, EFR3A, NPM2, and NUDT1 were (1) identified in our TWAS for both luminal A-like breast cancer and overall breast cancer, (2) did not lie near curated sets of candidate susceptibility variants for either breast or ovarian cancer, (3) further validated using an independent GWAS dataset, and (4) were not identified in the two recent TWAS of breast cancer. KLF7-IT1, LINC00683, and TMEM50A further met this criteria for overall breast cancer only. ACAP3 is predicted to play a role in GTPase activator activity, but the gene's possible role in tumorigenesis is unknown. EFR3A protein, however, has been implicated in oncogenic signaling and tumorigenic activity [122]. NUDT1 overexpression has been observed in several cancers, including breast [123, 124]. NPM2 is located near a well-studied tumor suppressor gene, DOK2, on 8p21.3, which is theorized to play a role in several cancers [125]. NPM2 further showed association with risk of luminal B/HER2-negative-like breast cancer using both BCAC and validation GWAS data.

While less powered to detect novel genes associated with risk of ovarian cancer and its main subtypes due to the limited number of ovarian tissue samples available in GTEx, we did identify two genes that may warrant further investigation. CCDC106 and MLLT10P1 were strongly associated with both HGSOC and NMOC with trans-driven GReX. They are not located near sets of curated candidate breast cancer and ovarian cancer risk variants and were not identified in recent cis-only TWAS of ovarian cancer [78, 121]. However, recent work in mutant p53 ovarian cancer cells has shown that overexpression of CCDC106 in particular leads to inhibition of p21 transcription and, ultimately, proliferation of the cancer cells [126]. MLLT10P1 was the only risk gene to validate using independent GWAS data in our ovarian cancer analyses. While it is a pseudogene, and therefore the biological mechanisms behind this association are unclear, functional research on pseudogenes has indicated that they can indeed play a role in tumorigenesis and are dysregulated in many cancers [127].

In this study, the abundance of non-trivial trans-SNP effects on gene expression that we observed in both mammary and ovarian tissues opened the door for identification of new potential risk genes and underscores the importance of including trans-SNPs in TWAS. However, we note there are several limitations to this work. Firstly, the samples used for training the genome-wide expression imputation models were limited in number, particularly for ovarian tissue (N = 140). One possible consequence of such a modest training size is overfitting, as reflected by inflated R^2 in the training samples. While BGW-TWAS is the first method of its kind to be computationally tractable enough to fit trans-eQTL models, its computational requirements do prevent cross-validation analysis during model training. Also, the lower prevalence of ovarian cancer relative to breast cancer makes identification of suitable validation

GWAS datasets and transcriptomic panels in ovarian tissue of sufficient sample size difficult to obtain.

Furthermore, we note that while breast cancer is a disease predominantly occurring among females, the majority of the GTEx samples for which RNA sequencing data was available in breast tissue came from men (63%). While training gene expression models using only female samples is ideal, the subsequent drop in sample size would have negatively impacted model predication accuracy. To assess any impact of sex bias in our analyses, we calculated sex-specific GReX imputation accuracy in the GTEx training samples of our BGW-TWAS genes identified for breast cancer. Training R^2 results were highly concordant between males and females. We also limited our scope to autosomal genes only and did not consider the important role of sex-chromosome genes in sex-biased diseases like breast cancer. Our study also used data only from individuals of European ancestry for the expression model training, gene-level association tests, and validation analyses. However, there are considerable disparities in clinical outcomes of these cancers across racial groups. Research has also reported unique gene expression profiles across non-European ancestries in breast cancer tumors [128], which motivates that the application of trans-eQTL TWAS to underrepresented populations.

Lastly, we note that the expression models of our trans-eQTL-driven TWAS genes were trained using normal breast and ovarian tissue from GTEx rather than tumor adjacent normal tissue or tissue with precursor lesions that are disease relevant. Our expression models showed little validation in tumor adjacent normal tissue (NAT) in TCGA, the closest independent surrogate samples for the GTEx normal breast tissue used to train our models. This may be a result of sample size or due to altered regulatory effects of these SNPs in these samples caused by their proximity to tumors. Indeed, recent work comparing GTEx tissues and the corresponding TCGA NAT across cancer types suggests that the NAT transcriptome is not “normal” but represents an intermediate gene expression state between normal and tumor with multiple pathway-level perturbations differentiating NAT from GTEx [129]. Further, some models validated in NAT tissue but failed to validate in the TCGA tumor samples. The association between predicted and observed expression in NAT here may be spurious considering the small sample size of NAT and that significance was assessed at the nominal level. However, this may otherwise be indicative of altered regulatory activity in tumor vs. non-tumor tissue whereby germline control of somatic gene expression may be lost during the oncogenic transition from normal/NAT to tumor tissue [130]. Lastly, some models validated in tumor tissue but not NAT. This is likely due to the difference in sample size between tumor tissue (N = 786) and tumor-adjacent normal tissue (N = 101). We have higher power to detect significant correlation between imputed GReX and observed GReX in tumor tissue samples. In fact, for most of these models, the estimated correlation coefficients were quite similar in both NAT and tumor, with a median difference near zero (0.03).

3.5 Web Resources

The BCAC GWAS summary statistic data for risk of breast cancer phenotypes are publicly available and can be accessed here. The manuscript for the OCAC GWAS of ovarian cancer phenotypes is currently under review, and the GWAS summary data will be made available for download upon publication. The Rashkin et al. pan-cancer GWAS summary statistics used in validation analyses are publicly available and can be accessed here. The MetaXcan suite of tools are available for download on Github here. Code and a corresponding tutorial for fitting BGW-TWAS models are available here.

3.6 Appendix

3.6.1 Tables

Table 3.5: Median (IQR) of sums of posterior probabilities of having non-zero eQTL effect sizes by range of training R^2 in fitted BGW-TWAS models.

Tissue	Training R^2	Num. Genes	Genome-wide	Cis-Region	Trans-Region
Breast	(0,0.05]	295	0.683 (1.838)	0.004 (0.011)	0.443 (1.462)
Breast	(0.05,0.1]	326	0.396 (0.761)	0.077 (0.592)	0.069 (0.252)
Breast	(0.1,0.25]	7719	0.982 (1.2)	0.005 (0.689)	0.428 (0.87)
Breast	(0.25,0.5]	15678	1.292 (2.546)	0.004 (0.013)	1.061 (2.238)
Breast	(0.5,1]	756	2.7 (4.44)	0.698 (1.191)	1.584 (3.35)
Ovary	(0,0.05]	262	0.336 (1.139)	0.004 (0.004)	0.285 (1.007)
Ovary	(0.05,0.1]	70	0.25 (1.113)	0.004 (0.003)	0.226 (0.905)
Ovary	(0.1,0.25]	581	0.36 (0.686)	0.007 (0.581)	0.165 (0.214)
Ovary	(0.25,0.5]	9702	0.423 (1.133)	0.004 (0.006)	0.282 (0.925)
Ovary	(0.5,1]	11940	0.388 (1.502)	0.004 (0.004)	0.362 (1.422)

3.6.2 Figures

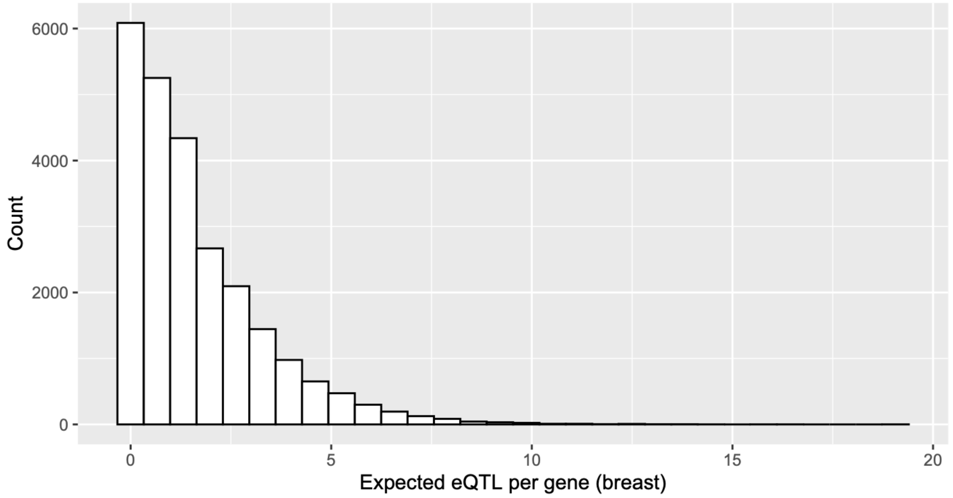

Figure 3.4: Histogram of expected number of eQTLs (cumulative posterior causal probability) per gene from BGW-TWAS imputation models in breast tissue.

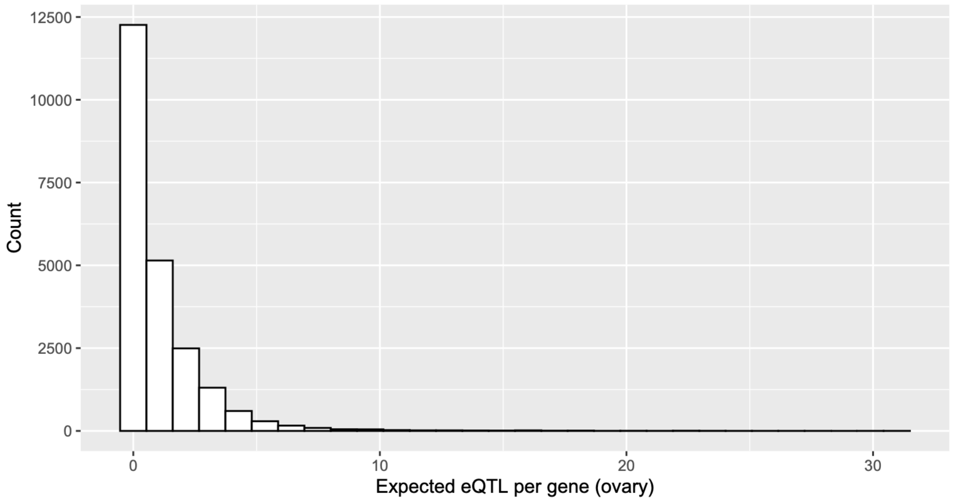

Figure 3.5: Histogram of expected number of eQTLs (cumulative posterior causal probability) per gene from BGW-TWAS imputation models in ovarian tissue.

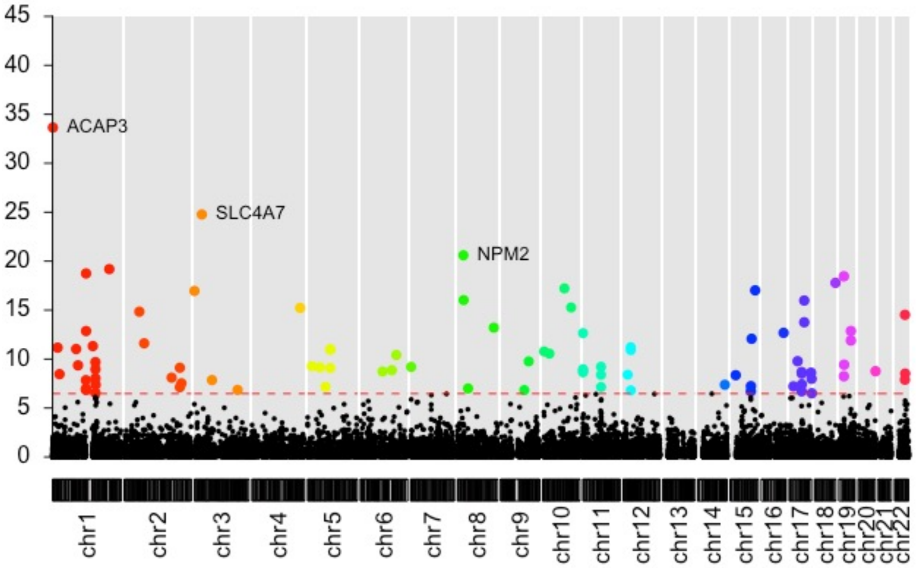

Figure 3.6: Manhattan plot of BGW-TWAS results for overall breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

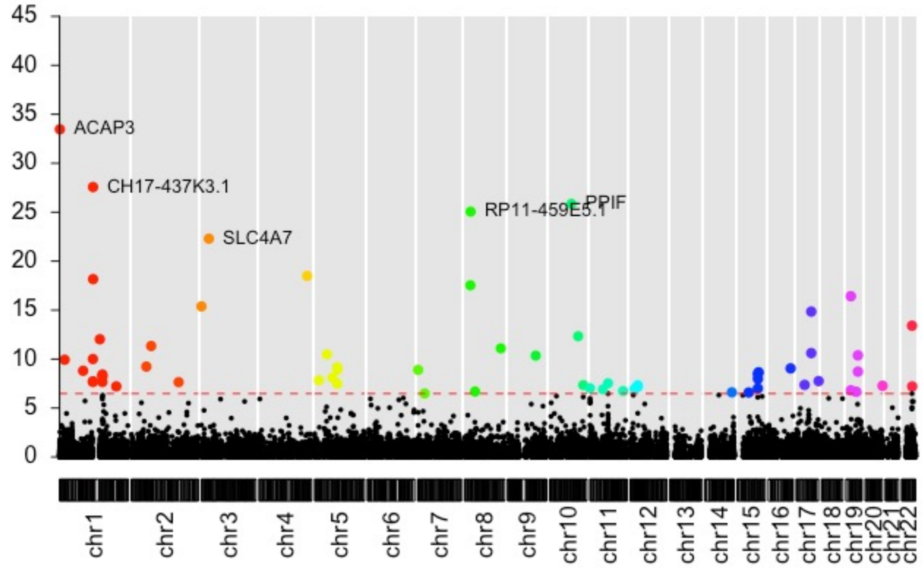

Figure 3.7: Manhattan plot of BGW-TWAS results for luminal A-like breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

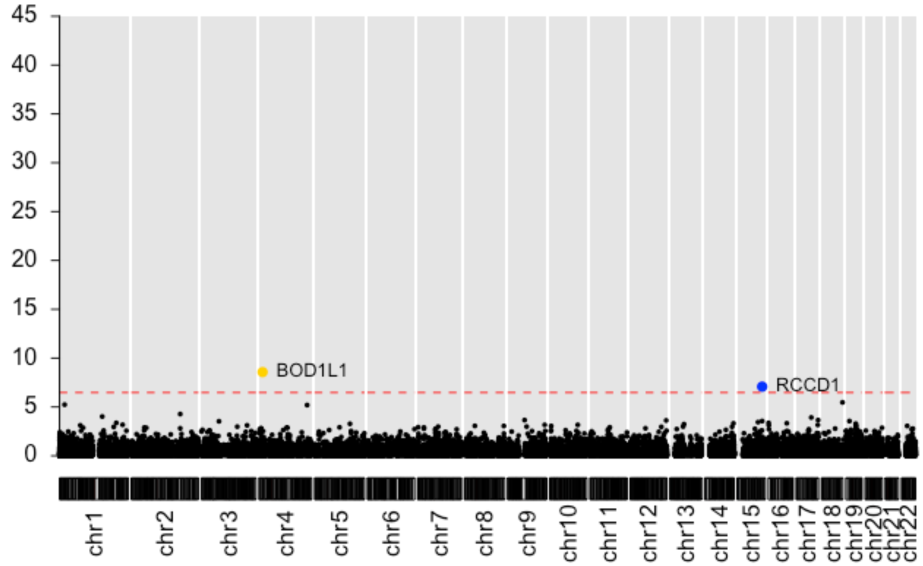

Figure 3.8: Manhattan plot of BGW-TWAS results for luminal B-like breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

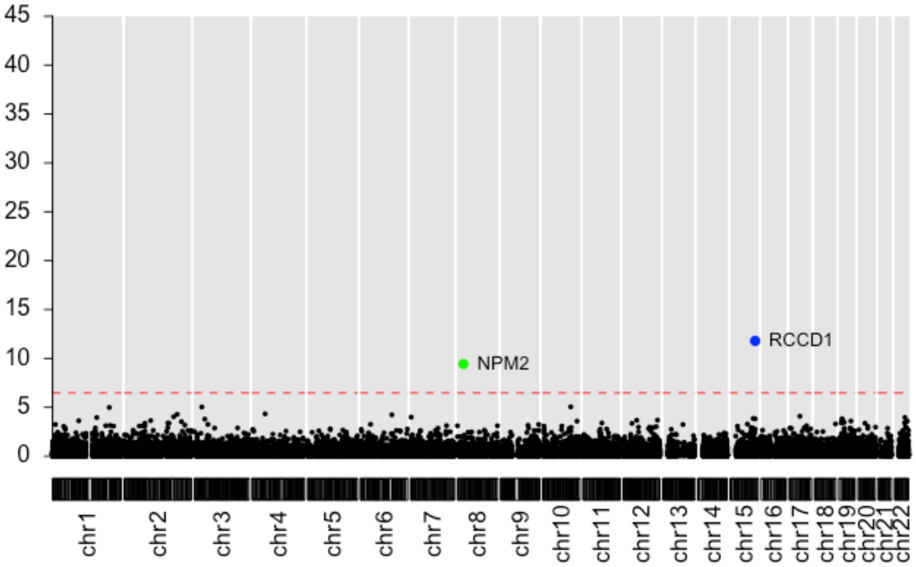

Figure 3.9: Manhattan plot of BGW-TWAS results for luminal B/HER2 negative-like breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

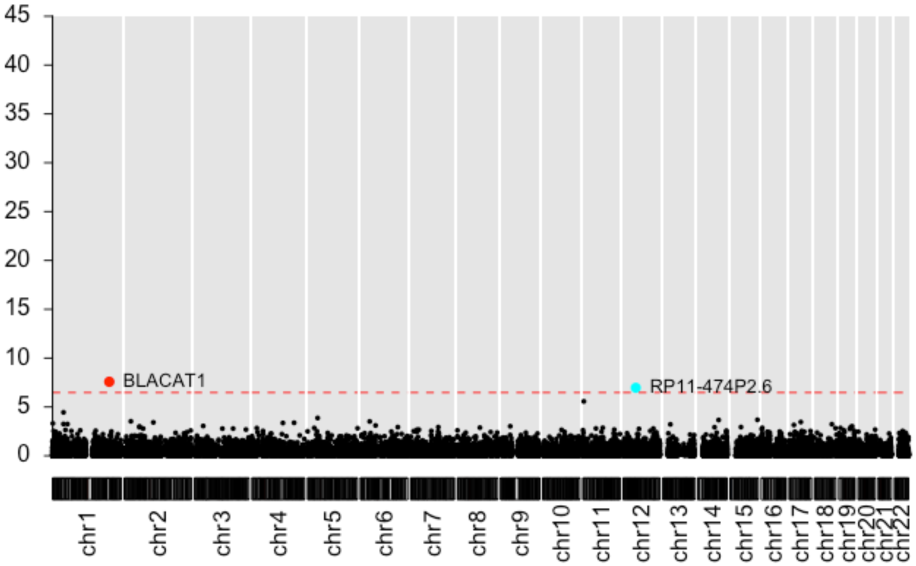

Figure 3.10: Manhattan plot of BGW-TWAS results for HER2 enriched-like breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

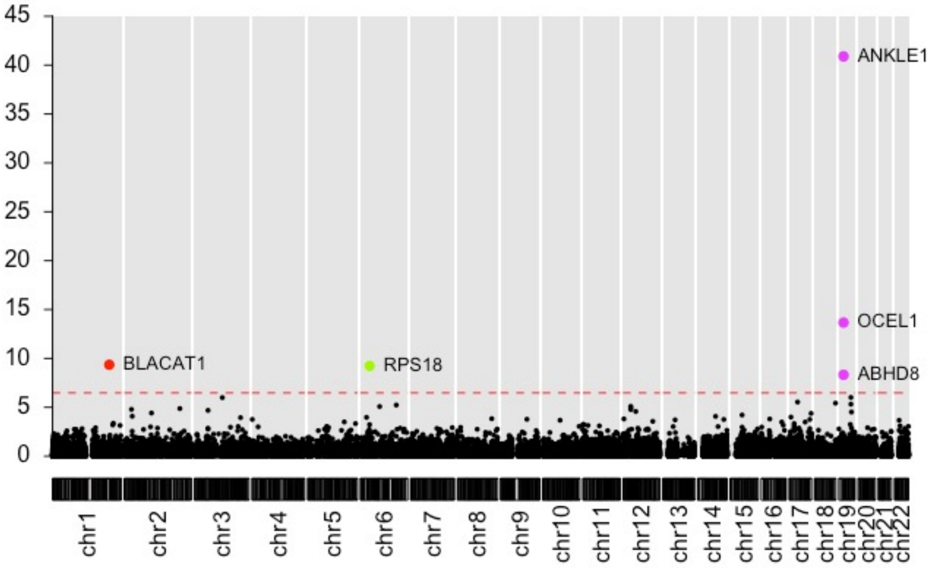

Figure 3.11: Manhattan plot of BGW-TWAS results for triple negative-like breast cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

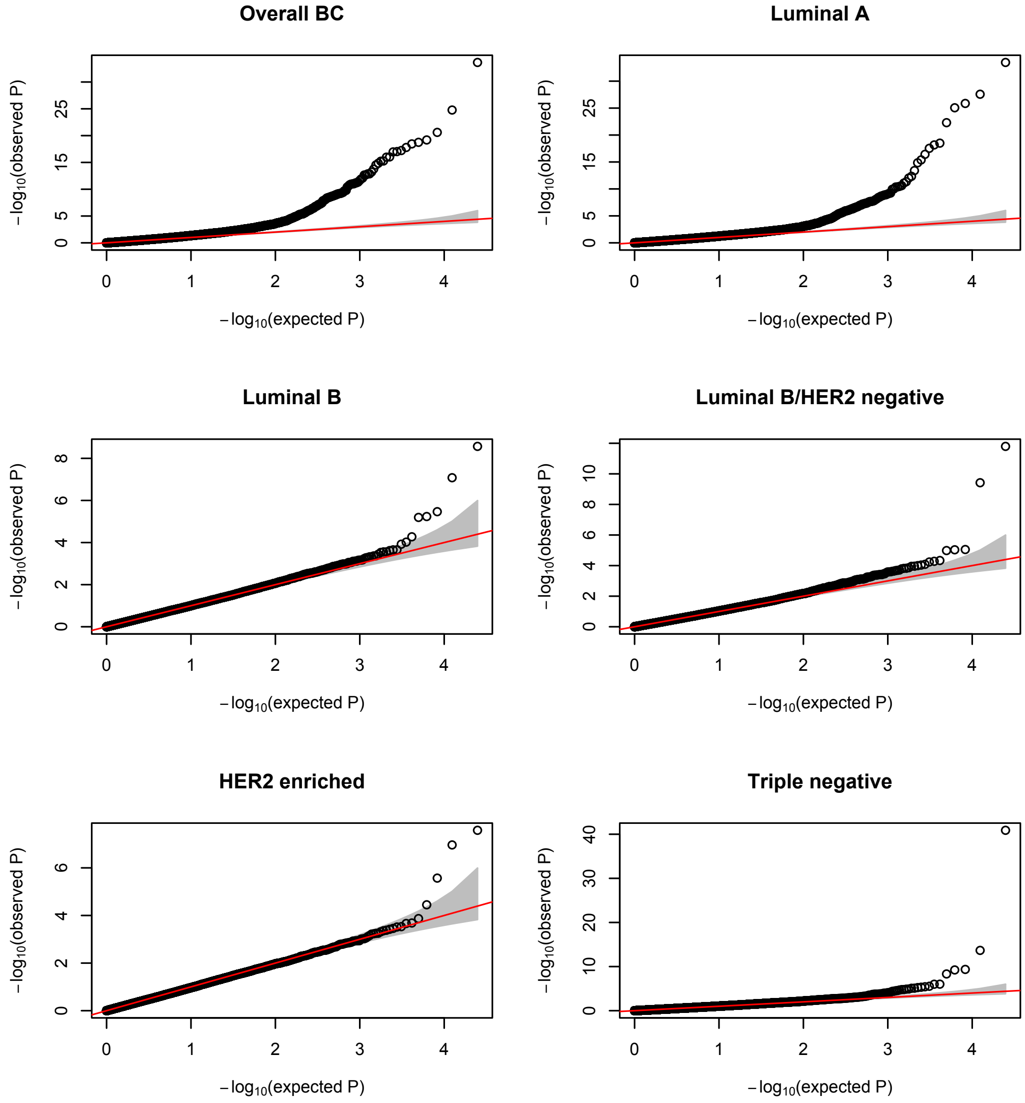

Figure 3.12: Quantile-quantile plots of BGW-TWAS results for six breast cancer phenotypes using BCAC GWAS summary statistics.

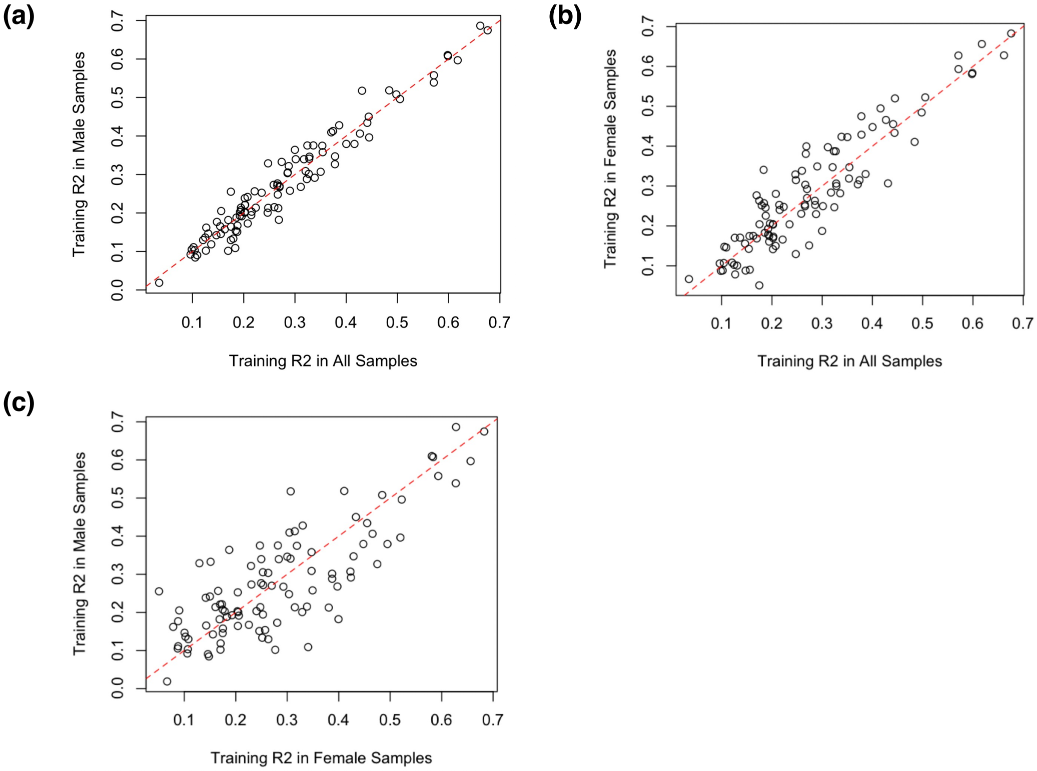

Figure 3.13: Training R^2 (squared correlation between imputed GReX and observed gene expression in the training GTEx dataset) of 101 breast cancer genes identified by BGW-TWAS. Panel (a) compares training R^2 in male subjects compared to all subjects. Panel (b) compares training R^2 in female subjects compared to all subjects. Panel (c) compares training R^2 in female subjects compared to male subjects.

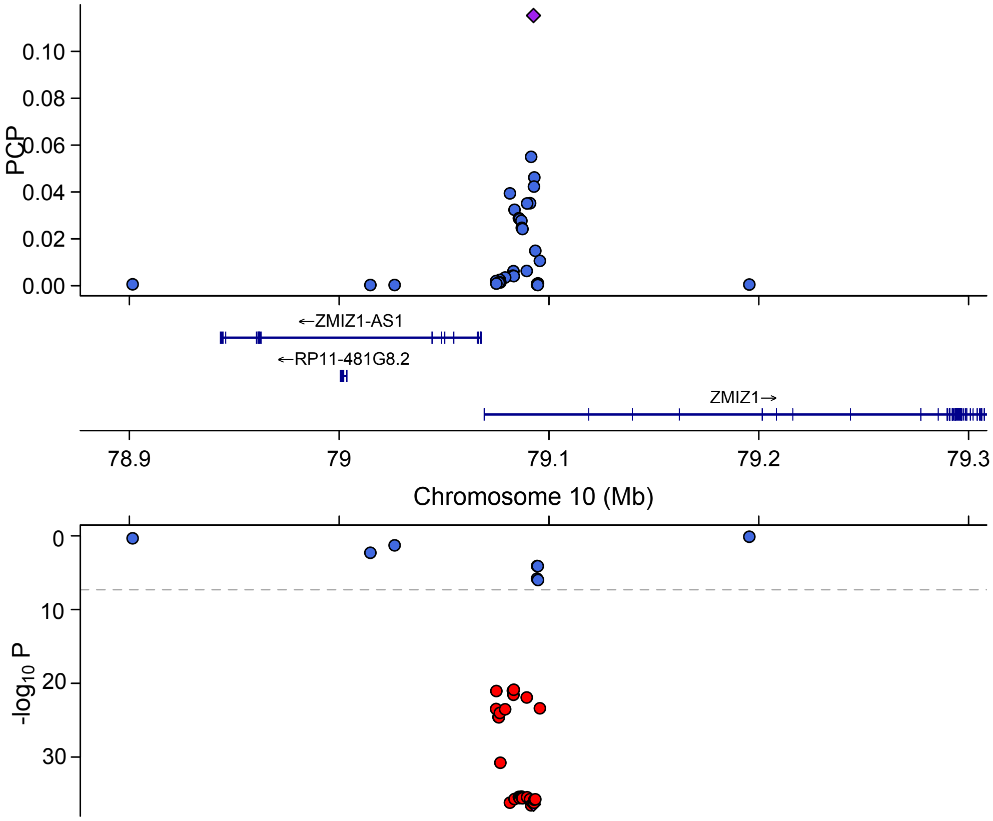

Figure 3.14: The upper figure shows the posterior causal probability of SNPs being eQTLs in the BGW-TWAS GReX imputation model for ACAP3 in breast tissue in the region Chr10:78892621-79292621. The bottom figure shows the corresponding -log(p) of GWAS BCAC p-values for these variants in the overall phenotype analysis.

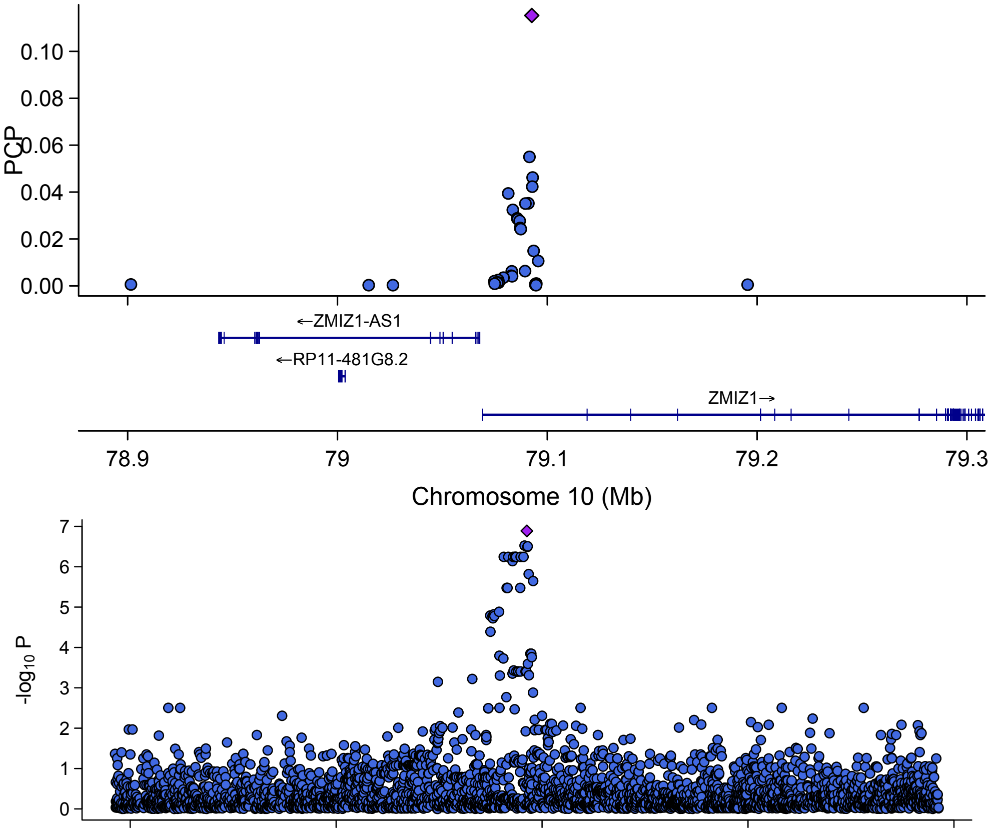

Figure 3.15: The upper figure shows the posterior causal probability of SNPs being eQTLs in the BGW-TWAS GReX imputation model for ACAP3 in breast tissue in the region Chr10:78892621-79292621. The bottom figure shows the corresponding -log(p) from single-variant eQTL analysis of all SNPs in this region in GTEx.

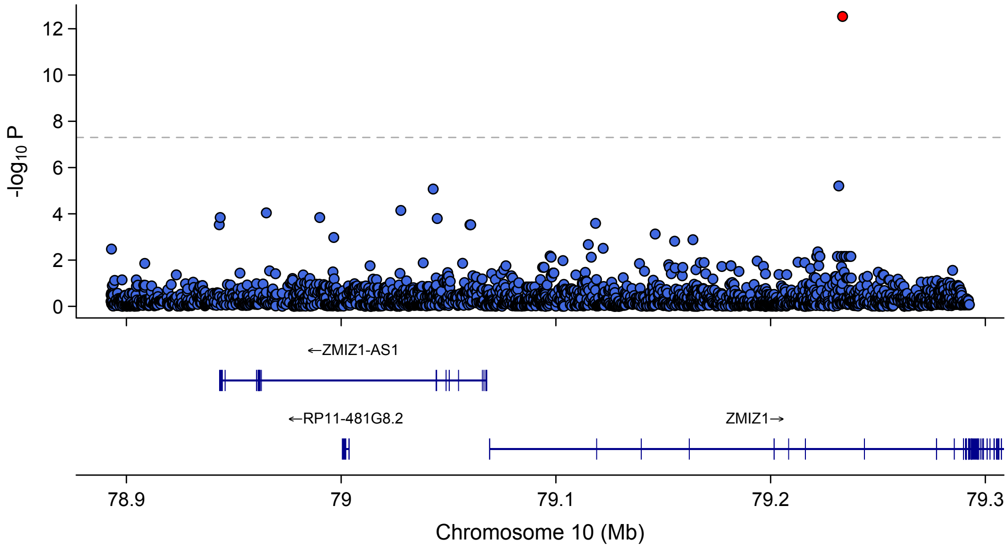

Figure 3.16: This figure shows -log(p) from single-variant eQTL analysis of all SNPs in Chr10:78892621-79292621 (top trans region for ACAP3) in TCGA breast tumor.

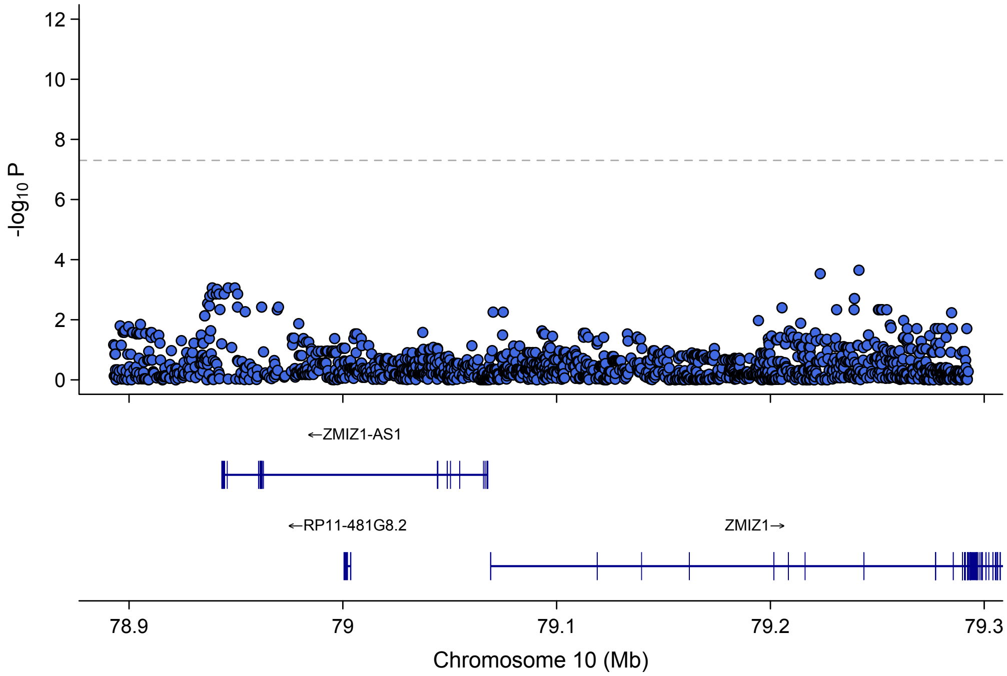

Figure 3.17: This figure shows -log(p) from single-variant eQTL analysis of all SNPs in Chr10:78892621-79292621 (top trans region for ACAP3) in TCGA breast tumor-adjacent normal tissue.

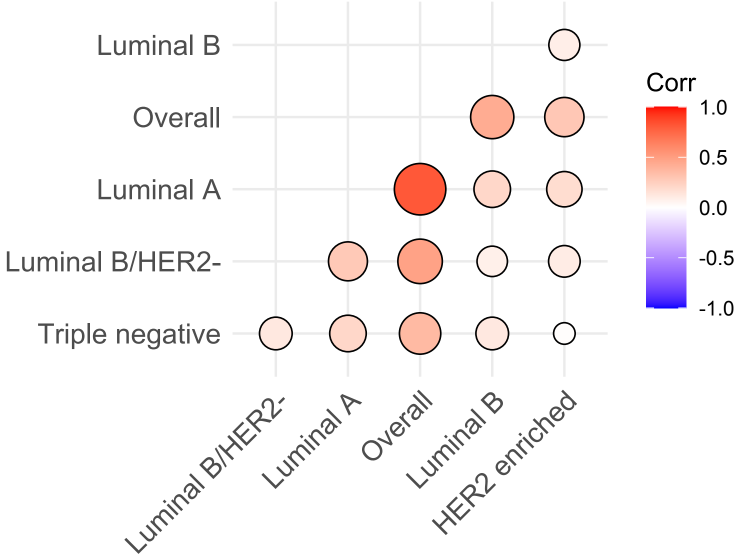

Figure 3.18: Correlation plot of BGW-TWAS Z scores across six breast cancer phenotypes using BCAC GWAS summary statistics.

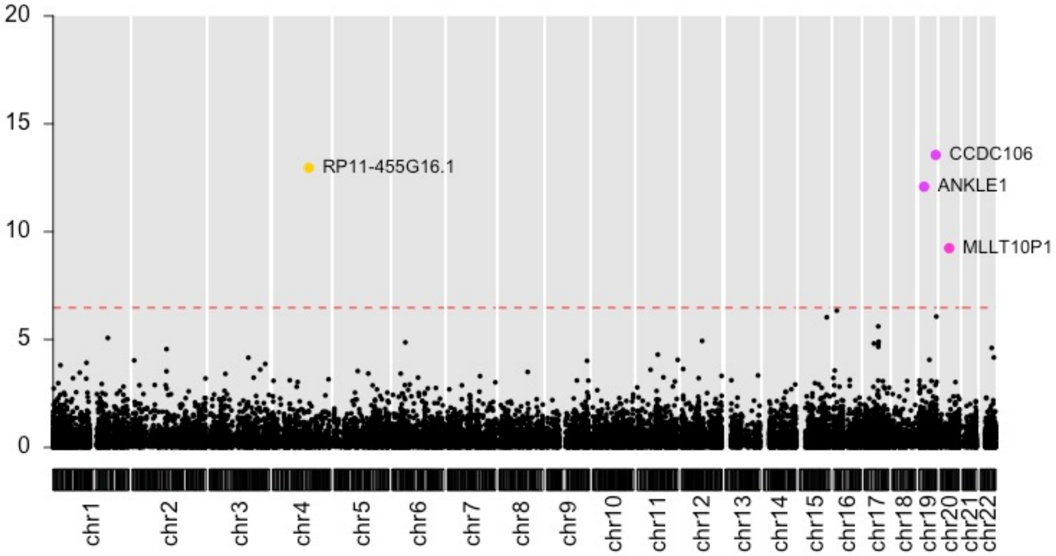

Figure 3.19: Manhattan plot of BGW-TWAS results for non-mucinous ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

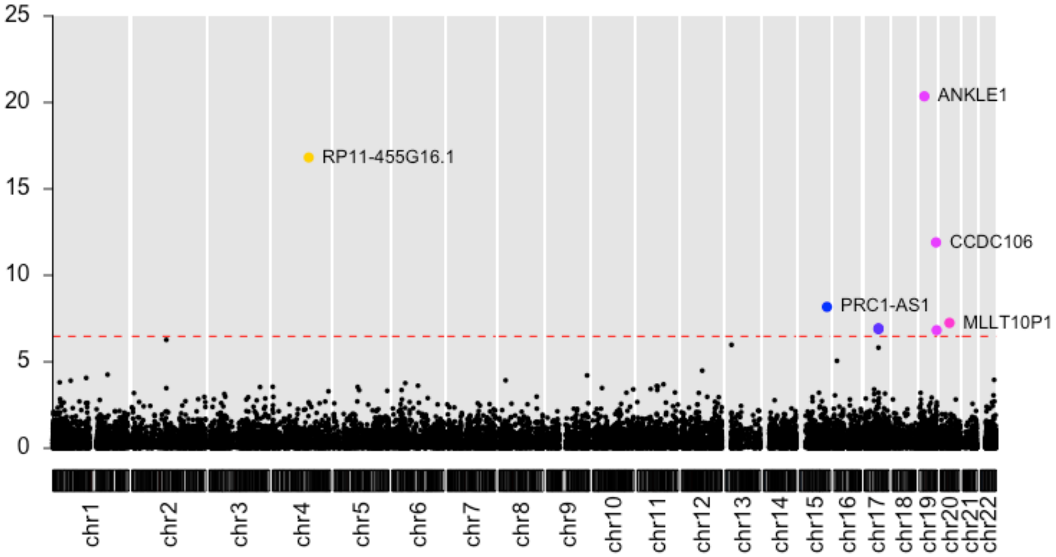

Figure 3.20: Manhattan plot of BGW-TWAS results for high grade serous ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

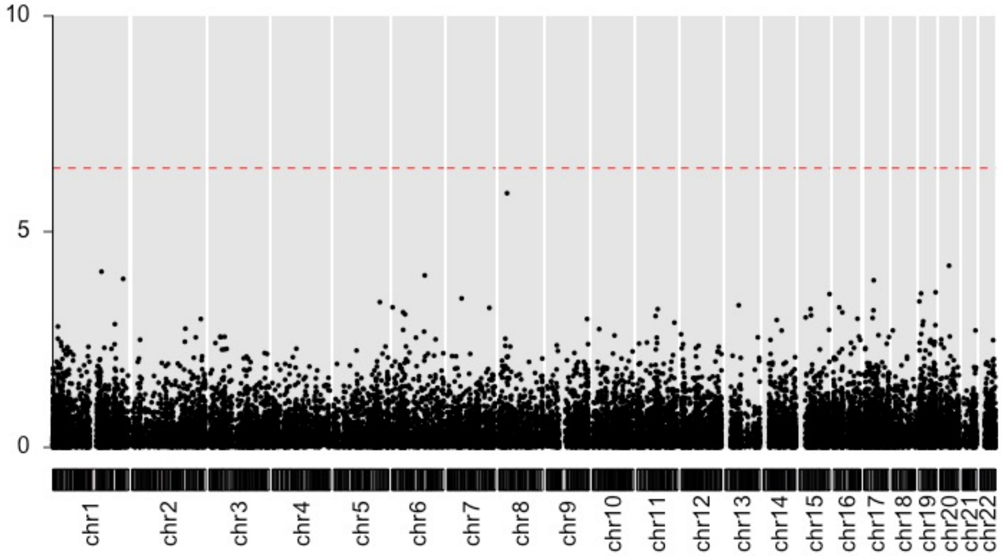

Figure 3.21: Manhattan plot of BGW-TWAS results for low grade serous ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

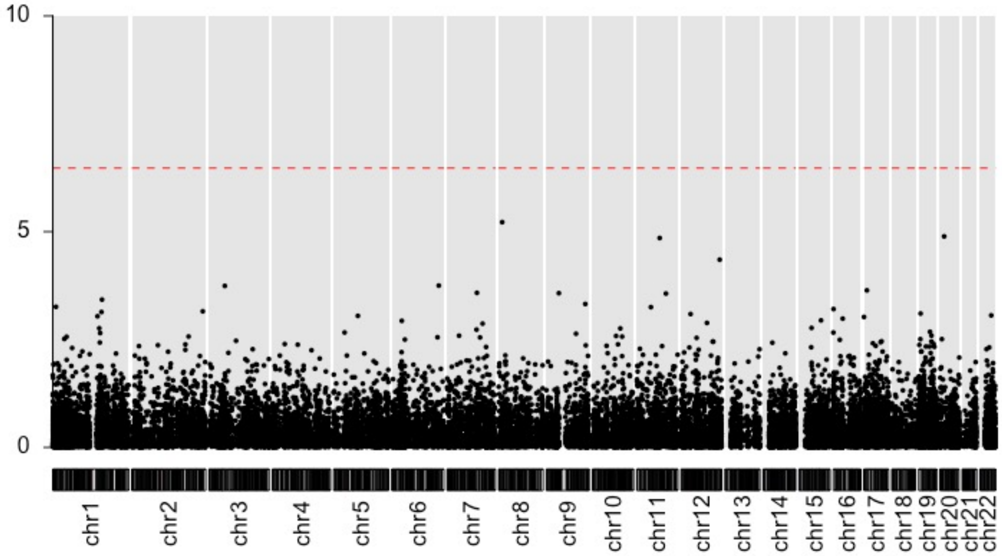

Figure 3.22: Manhattan plot of BGW-TWAS results for mucinous ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

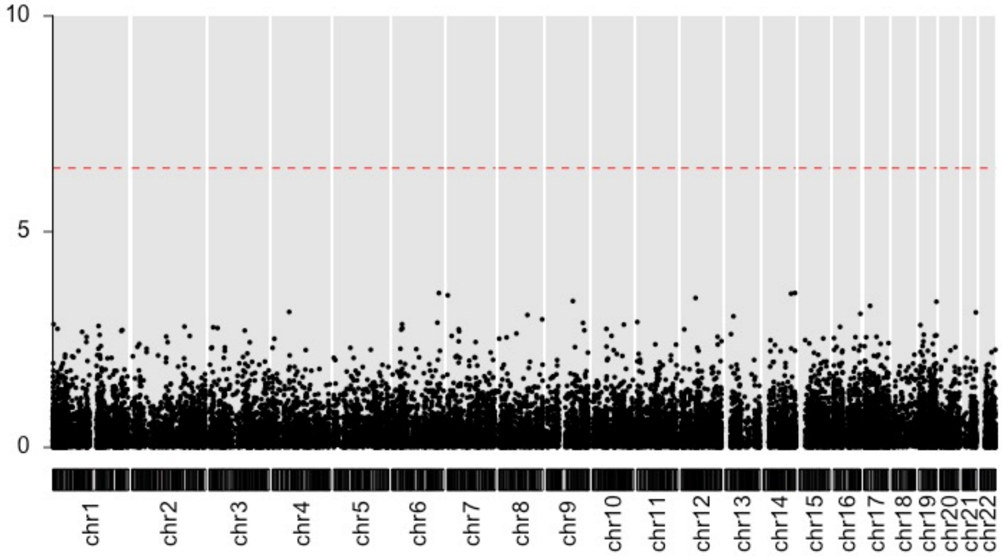

Figure 3.23: Manhattan plot of BGW-TWAS results for endometrioid ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

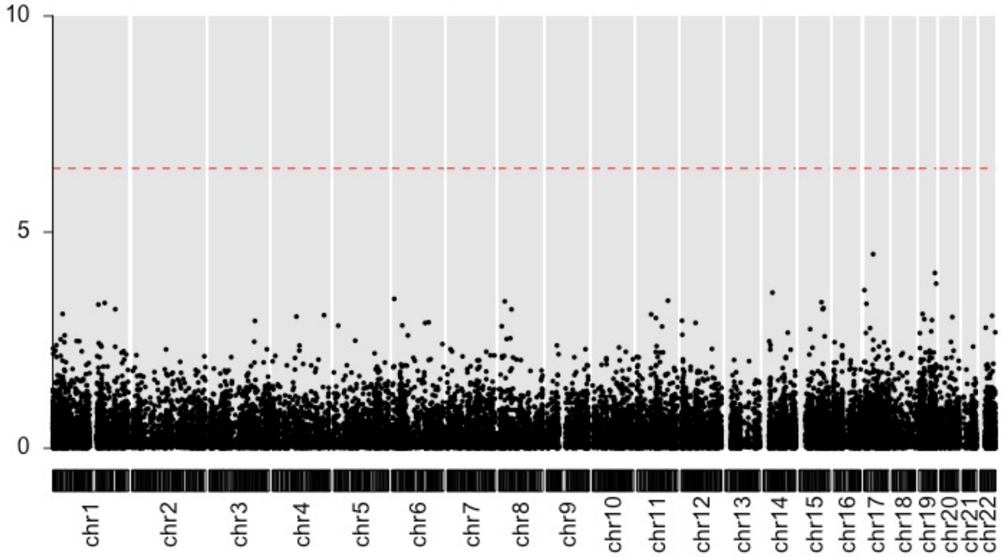

Figure 3.24: Manhattan plot of BGW-TWAS results for clear cell ovarian cancer risk. The dashed line denotes the Bonferroni-adjusted transcriptome-wide significance threshold.

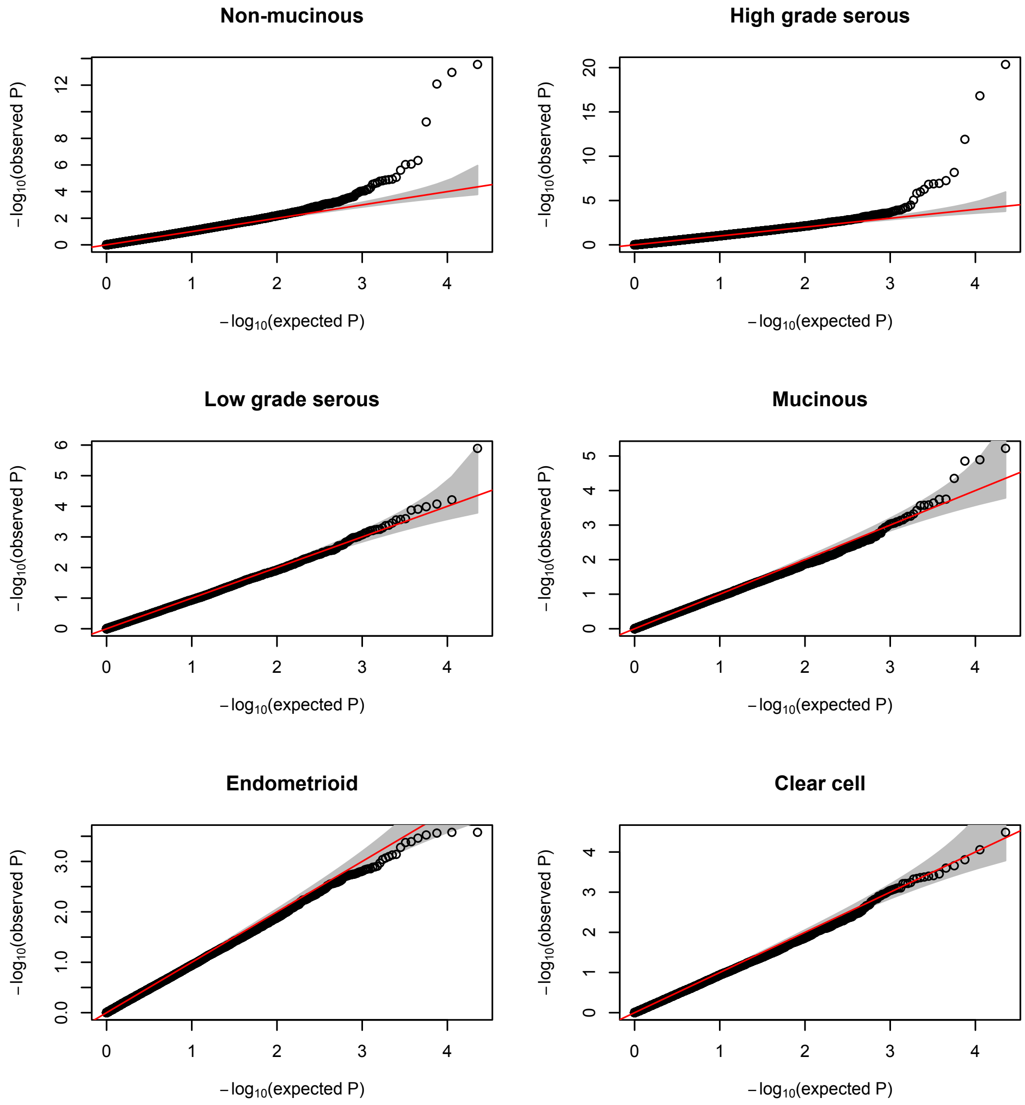

Figure 3.25: Quantile-quantile plots of BGW-TWAS results for six ovarian cancer phenotypes using OCAC GWAS summary statistics.

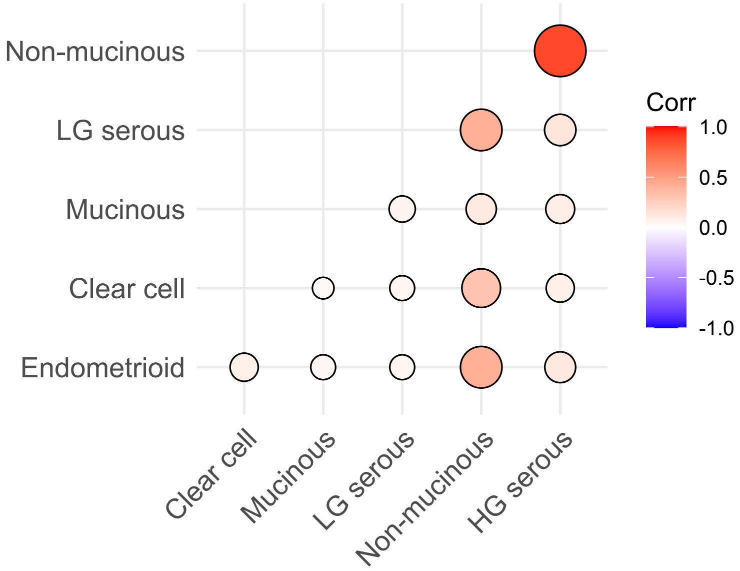

Figure 3.26: Correlation plot of BGW-TWAS Z scores across six ovarian cancer phenotypes using OCAC GWAS summary statistics.

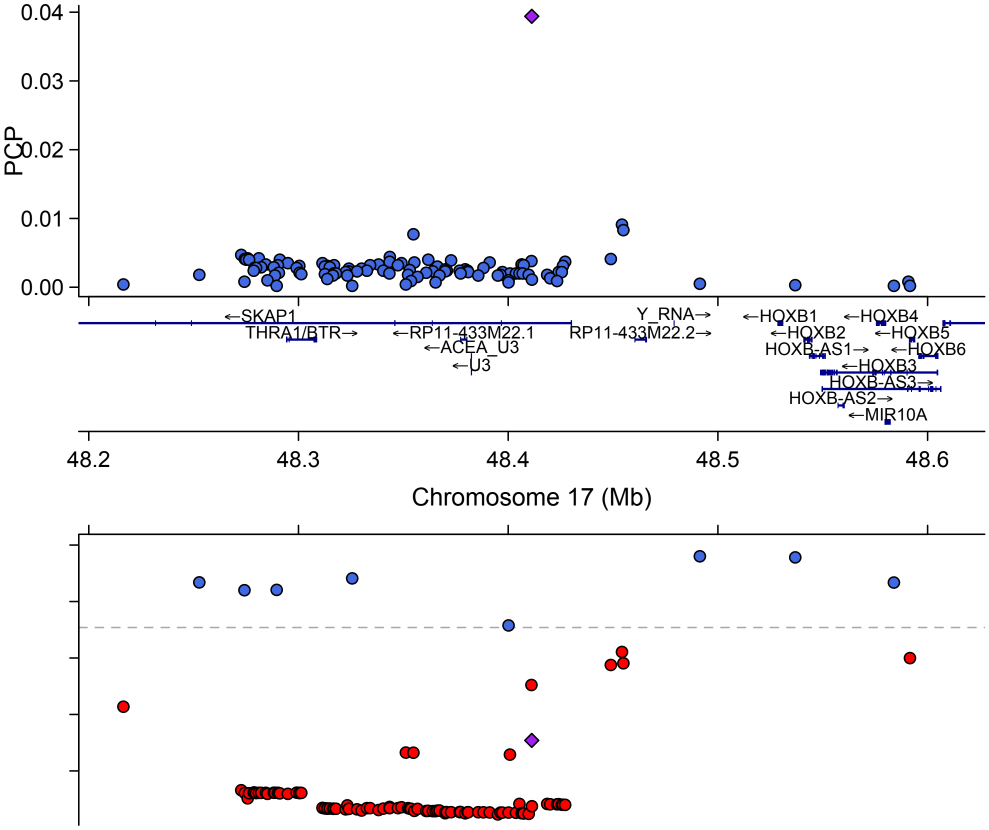

Figure 3.27: The upper figure shows the posterior causal probability of SNPs being eQTLs in the BGW-TWAS GReX imputation model for CCDC106 in ovarian tissue in the region Chr17:48211237-48611237. The bottom figure shows the corresponding -log(p) of GWAS OCAC p-values for these variants in the non-mucinous phenotype analysis.

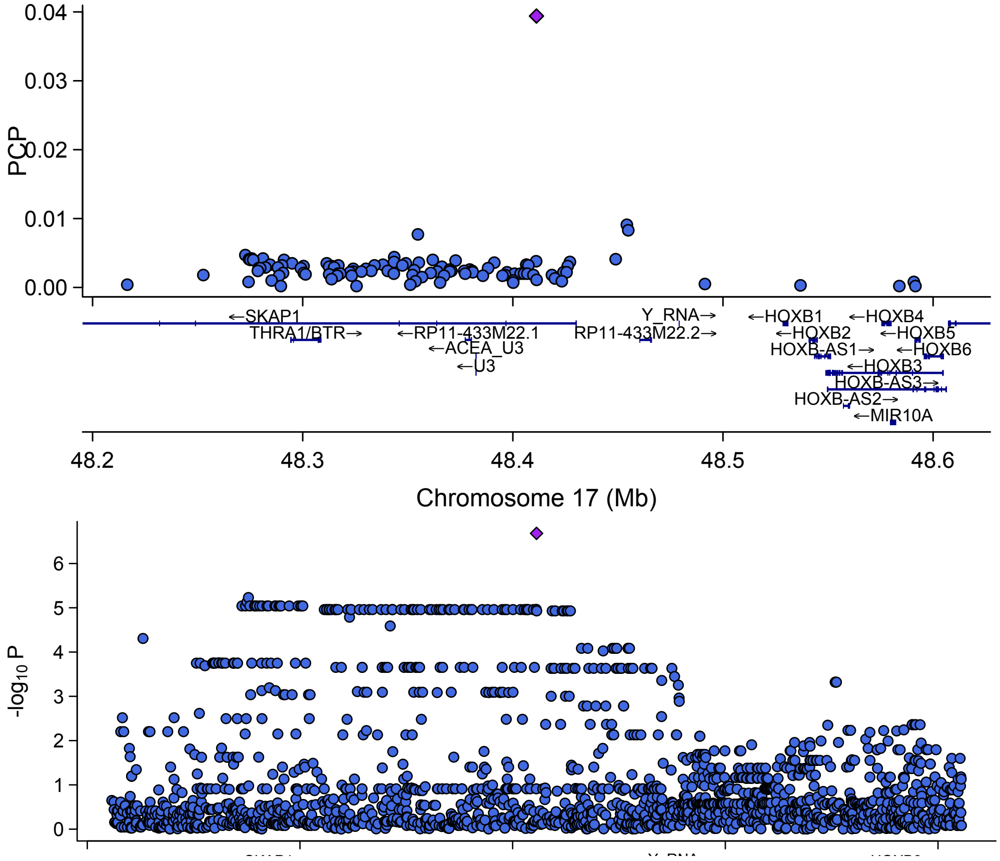

Figure 3.28: The upper figure shows the posterior causal probability of SNPs being eQTLs in the BGW-TWAS GReX imputation model for CCDC106 in ovarian tissue in the region Chr17:48211237-48611237. The bottom figure shows the corresponding -log(p) from single-variant eQTL analysis of all SNPs in this region in GTEx.

Chapter 4

Topic 3. Enhanced transcriptome-wide association analyses in admixed samples using eQTL summary data

4.1 Introduction

While genome-wide association studies (GWAS) have been largely successful in identifying genetic variants associated with a wide range of complex traits and diseases, the majority of the top variants lie in non-protein coding regions of the genome. It has been estimated that up to 90% of GWAS-identified single nucleotide polymorphisms (SNPs) are non-coding variants, and thus the biological mechanisms by which these variants exert their effects on a phenotype remain unclear [131]. Transcriptome-wide association studies (TWAS) have become a field of rapidly growing interest, as these methods seek to elucidate such complex regulatory mechanisms. A typical TWAS assumes two datasets: a training dataset, usually of modest size, that possesses genotype and gene expression data from a tissue related to the outcome of interest, and a testing (GWAS) dataset that possesses genotype and outcome data but lacks expression data. TWAS involves a two-stage process. In Stage I, the process constructs a model on the training dataset to identify variants (expression quantitative trait loci, or eQTLs) associated with expression levels of a target gene and their corresponding effect sizes (weights). In Stage II, the process uses the eQTL weights from Stage I to impute genetically-regulated gene expression (GReX) within subjects from the testing (GWAS) dataset and then tests for association between the imputed expression and the outcome of interest.

Stage I of the majority of established TWAS methods require individual-level training data in order to impute gene expression in Stage II [73–79]. In other words, paired individual-level genotype and transcriptomic datasets from reference databases are typically required to ultimately perform the association test between imputed GReX and the trait of interest. While most individual-level training eQTL datasets from databases like the Genotype-Tissue Expression (GTEx) project [20] are small, there exist much larger publicly available datasets in summary form. Using larger eQTL summary statistic datasets for TWAS can lead to greater accuracy of training models and greater power in Stage II of TWAS. To leverage these potential power gains, Dai et al. recently published a novel TWAS method (OTTERS). This method uses estimated eQTL effect sizes and corresponding single-variant p-values to impute GReX without the need for individual-level transcriptomic data [22]. To do this, OTTERS jointly uses multiple well-established methods developed for modeling polygenic risk scores (PRS) using summary statistic GWAS data to train multiple gene expression imputation models. The specific PRS models used in OTTERS can be selected by the user, and, using these, we can subsequently impute multiple GReX vectors in the Stage II testing dataset. OTTERS then tests for the association between each imputed vector and the trait of interest, and p-values are ultimately combined using the aggregated Cauchy association test (ACAT) [132]. Through simulations, OTTERS demonstrated proper type I error rate control and power gains over a competing TWAS approach that uses individual-level stage I data [74].Additionally, the simulation studies and applied work suggest that the optimal PRS method for imputing GReX is a function of the genetic architecture of the gene under study, e.g., the number of causal eQTLs with non-zero effect sizes and the heritability of expression.

Importantly, OTTERS is intended for TWAS of individuals of only a single, homogeneous ancestral group. Given the increased collection and analysis of multi-ancestry subjects across the entire spectrum of genetic ancestry, there is need to expand the framework to handle genetic and genomic data from diverse groups. However, in order to develop such a method for multi-ancestry groups, we first emphasize that the underlying genetic architecture of gene expression may not be the same across populations and may differ by ancestral group [133]. As a result of these differences in genetic architecture and additional factors that differ between continental ancestries, applied work has shown that gene expression prediction models trained in one ancestral population do not generalize well to other populations [134]. The gene expression prediction models used in such work were similar to the PRS methods used by OTTERS, and such PRS methods are known to have poor transferability across ancestral groups. This has been well-documented across a variety of phenotypic domains. Indeed, research has shown that PRS models trained using GWAS data from one ancestry have significantly decreased prediction performance when applied to individuals of an ancestral background different than that used for training [135–139]. It is hypothesized that this poor portability may be a function of a population-specific effect sizes. These population-specific effects may arise from a multitude of factors that differ across populations, such as gene-environment interaction effects, gene-gene interaction effects, allele frequencies, and linkage disequilibrium (LD) patterns [140–142]. Poor performance of PRS in diverse ancestries is further exacerbated by the consistently lower sample sizes of diverse GWAS cohorts compared to European-ancestry cohorts [135, 143].

The development of sufficiently powered TWAS methods for diverse ancestries is particularly important for admixed individuals, whose genomes are a unique mosaic of multiple continental ancestral groups. Admixed groups account for a large and growing proportion of the United States population, with more than 33.8 million people identifying as multiracial in the 2020 Census [144]. It has been common practice over the past few decades in GWAS to exclude admixed individuals from consideration due to their complex ancestral makeup, and as such, they represent a historically underrepresented group in genetic studies. In response to this realization, researchers have recently considered the utility of including local ancestry information in variant-level genetic association analyses of admixed populations [145–155]. Building on these advancements, we have also seen a recent rise in novel PRS approaches specifically designed for admixed populations [156–158] that we can leverage for related TWAS. For example, Marnetto et al. developed an ancestry-aware approach for PRS that first deconvoluted admixed haplotypes for a test subject and then computed the subject's ancestry-specific components of the PRS using GWAS summary data from the appropriate reference ancestral populations. Authors demonstrated that this approach can not only yield improved phenotype predictability over standard methods but also an unbiased distribution of PRS in recently admixed populations. [156].

Given the existing evidence in the literature for differential eQTL architecture between admixed groups and more ancestrally homogenous subjects [133, 159], we propose an enhanced method for performing TWAS in admixed samples that leverages local ancestry information as well as summary-level eQTL data from multiple reference datasets of differing ancestry. In this method, we first apply the gene-expression training models used by OTTERS separately to each reference dataset. We then apply a local-ancestry deconvolution method to our test admixed sample and then, for a given gene, impute ancestry-specific partial GReX following the method of Marnetto et al. We can then combine the vectors of ancestry-specific GReX into an aggregate ancestry-aware measure of GReX. We can then test whether the aggregate ancestry GReX (as well as ancestry-specific GReX) are associated with outcome. We can then further combine the aggregate and ancestry-specific results together into an omnibus test using a Cauchy combination test similar to ACAT. We evaluate the performance of our method in simulations via expression imputation R^2 and power analyses, and we demonstrate the method is well-calibrated under the null hypothesis of GReX-phenotype independence. We conclude with an application of our method to 29 blood biochemistry phenotypes in two-way African/European admixed individuals in the UK Biobank and compare its performance to expression imputation models that ignore local ancestry.

4.2 Materials and Methods

4.2.1 Overview

The overarching goal of our TWAS is to impute gene expression in GWAS data from admixed individuals using summary-level eQTL data from individuals of the founder (continental) ancestral groups from which the admixed participants are derived. Like most established TWAS, this involves two general stages. Stage I focuses on training models of predicted gene expression in each of the reference summary-eQTL datasets under consideration. Stage II uses estimates from these trained models, along with local ancestry information (discussed later), to estimate ancestry-specific genetically-regulated gene expression and then relate such expression to an outcome of interest.

4.2.2 Modeling Expression in Admixed Individuals

Without loss of generality, we assume our admixed test subjects are two-way admixed, e.g., of African (AFR) and European (EUR) descent. We first define how gene expression is modeled in this population. For the purposes of all subsequent notation, the subscript “1” stands for “AFR”, and “2” stands for “EUR”. Let us assume we have N_{adm} admixed individuals. For a given gene g, we assume there are S cis-eQTLs that are shared between the two ancestral groups and U cis-eQTLs that are unique to only one ancestral group. We define the total number of cis-eQTLs for a given gene in each ancestral group as $V = S + U$.

For $v \in \{1, ...V\}$, let $x^M_{i,v,1}(x^P_{i,v,1}) \in \{0, 1\}$ be the number of minor alleles for vth AFR eQTL on the maternal (paternal) haplotype of subject i. Let $\gamma^M_{i,v,1}(\gamma^P_{i,v,1}) \in \{0, 1\}$ be an indicator variable taking value 1 if the local ancestry of vth AFR eQTL on maternal (paternal) haplotype of subject i is AFR. Similarly, we define $x^M_{i,v,2}(x^P_{i,v,2}) \in \{0, 1\}$ as the number of minor alleles for the vth EUR eQTL on the maternal (paternal) haplotype of subject i. Let $\gamma^M_{i,v,2}(\gamma^P_{i,v,2}) \in \{0, 1\}$ be an indicator variable taking value 1 if the local ancestry of vth EUR eQTL on the maternal (paternal) haplotype of subject i is EUR.

Thus, we can now let $g_{i,v,1} := x^M_{i,v,1}\gamma^M_{i,v,1} + x^P_{i,v,1}\gamma^P_{i,v,1}$ represent the number of AFR-ancestry minor alleles of the vth AFR eQTL and $g_{i,v,2} := x^M_{i,v,2}\gamma^M_{i,v,2} + x^P_{i,v,2}\gamma^P_{i,v,2}$ be the number of EUR-ancestry minor alleles of subject i at vth EUR eQTL. We can arrange these quantities into matrices $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$, each of dimension $N_{adm} \times V$. In these matrices, we further assume each column has been centered to mean zero. We can therefore model the gene expression vector of our testing population as follows:

$$\boldsymbol{E_g} = \boldsymbol{G_1 T_1^{1/2}} \boldsymbol{\beta}_1 + \boldsymbol{G_2 T_2^{1/2}} \boldsymbol{\beta}_2 + \boldsymbol{\epsilon_g}, \;\; \boldsymbol{\epsilon_g} \sim N(\boldsymbol{0}, (1 - h^2_{e,adm})\boldsymbol{I_{N_{adm}}}) \tag{4.1}$$

Here, $\boldsymbol{E_g}$ is the $N_{adm} \times 1$ vector of gene expression in our admixed testing population. $\boldsymbol{T_1}$ and $\boldsymbol{T_2}$ are the $V \times V$ diagonal scaling matrices with diagonal elements $(\boldsymbol{T_l})_{vv} = \tau^2_{vl} = \frac{1}{2f_{vl}(1-f_{vl})}$, where f_{vl} is the MAF of the vth eQTL in reference population l ($l = 1, 2$). $h^2_{e,adm}$ represents the gene expression heritability in our admixed subjects. $\boldsymbol{\beta_l} \in \mathbb{R}^V$ is the vector of causal eQTL effect sizes in population l. $\boldsymbol{\epsilon_g} \in \mathbb{R}^{N_{adm}}$ is the error vector, while $\boldsymbol{I_{N_{adm}}}$ is the $N_{adm} \times N_{adm}$ identity matrix. Using these quantities, we can derive both the joint distribution of all eQTL effect sizes and an estimate for the heritability of gene expression in our admixed subjects (Appendix).

4.2.3 Stage I Reference Expression Model Training via OTTERS

Let us now assume, for our Stage I training datasets, we have summary-level eQTL data derived from N_{ref} individuals in each of the two reference populations that represent the source ancestries among our two-way admixed testing population. In other words, we require eQTL summary data from an AFR cohort and EUR cohort for imputing gene expression in African American individuals. We note that N_{ref} need not be the same for both AFR and EUR groups, but we assume so for ease of presentation.

In each homogeneous ancestral population (AFR, EUR), we can model the genetically-regulated component of gene expression of each gene g as we did for our admixed subjects in Section 4.2.2. As above, expression is a function of the V cis-eQTLs with non-zero effect sizes in each population.

$$\boldsymbol{E_{g1}} = \boldsymbol{X_{g1}}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_{g1}},\ \boldsymbol{\epsilon_{g1}} \sim N(\mathbf{0}, (1 - h^2_{e,1})\boldsymbol{I_{N_{ref}}}) \tag{4.2}$$

$$\boldsymbol{E_{g2}} = \boldsymbol{X_{g2}}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_{g2}},\ \boldsymbol{\epsilon_{g2}} \sim N(\mathbf{0}, (1 - h^2_{e,2})\boldsymbol{I_{N_{ref}}}) \tag{4.3}$$

In the above equations, $\boldsymbol{E_{g1}}$ and $\boldsymbol{E_{g2}}$ are the $N_{ref} \times 1$ vectors of gene expression in our AFR and EUR reference populations. $\boldsymbol{X_{gl}}$ is the $N_{ref} \times V$ matrix of genotypes (0/1/2) with columns centered and standardized by minor allele frequency (MAF) in each population l. $\boldsymbol{\beta_l} \in \mathbb{R}^V$ is the vector of ancestry-specific eQTL effect sizes in population l ($l = 1, 2$). $\boldsymbol{\epsilon_{gl}} \in \mathbb{R}^{N_{ref}}$ are the error vectors for each population l, while $\boldsymbol{I_{N_{ref}}}$ is the $N_{ref} \times N_{ref}$ identity matrix. $h^2_{e,l}$ represents the gene expression heritability in population l.

In reality, we do not know a priori the true V cis-eQTLs present in Equations 4.2 and 4.3. Instead, we attempt to model gene expression using eQTL summary information available from $J >> V$ cis-SNPs found in and around gene g. That is, we have summary-level results (estimated effect sizes and p-values) from the below single-variant linear regression models (Equations 4.4 and 4.5) for each cis-SNP $j \in \{1..., J\}$ within 1MB of the transcription start site and end site of each gene g.

$$\boldsymbol{E_{g1}} = \frac{\boldsymbol{x_{j1}} - 2f_{j1}}{\sqrt{2f_{j1}(1-f_{j1})}}\beta_{j1} + \boldsymbol{\epsilon_{j1}},\ \boldsymbol{\epsilon_{j1}} \sim N(0, \sigma^2_{\epsilon_{j1}}\boldsymbol{I}), j = 1, ...J. \tag{4.4}$$

$$\boldsymbol{E_{g2}} = \frac{\boldsymbol{x_{j2}} - 2f_{j2}}{\sqrt{2f_{j2}(1-f_{j2})}}\beta_{j2} + \boldsymbol{\epsilon_{j2}},\ \boldsymbol{\epsilon_{j2}} \sim N(0, \sigma^2_{\epsilon_{j2}}\boldsymbol{I}), j = 1, ...J. \tag{4.5}$$

These linear regression models are fitted separately in each population l. In the above equations, $\boldsymbol{x_{j1}}$ and $\boldsymbol{x_{j2}}$ are the $N_{ref} \times 1$ vectors of genotypes (0/1/2) for the jth variant, $j \in \{1..., J\}$, in the AFR and EUR reference populations, respectively. Similarly, we have error terms $\boldsymbol{\epsilon_{j1}}$ and $\boldsymbol{\epsilon_{j2}}$. β_{j1} and β_{j2} are the standardized marginal eQTL effect sizes for this variant in AFR and EUR. Derived from these fitted models, for each gene in population l, we have summary statistics in the form of $\widehat{\beta_{jl}}$ (the marginal least squares effect estimate) and corresponding p-value p_{jl}. These summarize the marginal association of variant $j \in \{1, .., J\}$ with the expression of the gene of interest in population l.

OTTERS PRS Models

Using these marginal estimated eQTL effect size vectors $(\widehat{\beta}_{11}, ..., \widehat{\beta_{J1}})$, $(\widehat{\beta}_{12}, ..., \widehat{\beta_{J2}})$ and corresponding marginal p-value vectors from the single-variant eQTL data, we next use OTTERS to train PRS models to impute gene expression in both populations separately. While the OTTERS pipeline includes multiple frequentist and Bayesian PRS methods as options, here we focus on two as illustrative examples. We briefly summarize the methodology of these two PRS models below.

Pruning and Thresholding (P+T): This approach includes two steps: pruning, or clumping, of variants to exclude correlated SNPs, and thresholding to keep only those SNPs significantly associated with gene expression [160]. First, variants are filtered to include only those with p-values $< P_T$. Next, among these, a set of pairwise-independent variants are selected as those with linkage disequilibrium (LD) $R^2 < R_T$, preferentially keeping those with the smallest p-value. These pruning and thresholding steps are performed in the OTTERS pipeline using PLINK 1.9 [161]. For our analysis, we used thresholds $P_T = (0.05, 0.001)$ and left $R_T = 1$. We chose to not first prune SNPs as LD patterns differ considerably between populations, and OTTERS previously indicated that LD pruning did not significantly impact the performance of their method. We used the marginal standardized eQTL effect sizes from SNPs meeting this criteria to predict expression.

lassosum: This method represents a summary-statistics-based version of the least absolute shrinkage and selection operation (lasso) pipeline, a penalized variable selection approach for dimension reduction with a large number of predictors (variants) [162]. Full details on lassosum are provided elsewhere [163]. In contrast to P+T methods, lassosum requires LD blocks from an external reference panel. LD blocks were pre-calculated using `lddetect` [93] and data from 1000 Genomes (AFR and EUR populations, respectively) [91]. Note, for reasons indicated above, we did not perform LD clumping prior to training PRS via lassosum.

Using the two pruning and threshold approaches (P+T0.05, P+T0.001) and the lassosum approach, we have three sets of estimated eQTL effect size vectors for each training population for a given gene g: $\widehat{\boldsymbol{\beta_1}}^{\boldsymbol{\omega}}, \widehat{\boldsymbol{\beta_2}}^{\boldsymbol{\omega}}$, where $\omega \in \{P + T0.05, P +$

$T0.001, \text{lassosum}\}$.

4.2.4 Stage II Imputing Expression in Admixed Individuals

As the true causal eQTLs in each ancestry are unknown, we propose to impute ancestry-specific components of GReX in our admixed testing set using the trained PRS models of gene expression in the two reference eQTL datasets. Let us assume that we have phased genotype information and have performed ancestry deconvolution of our admixed testing dataset haplotypes. For each PRS model considered $\omega \in$ {P+T0.05, P+T0.001, lassosum}, we use the estimated eQTL effect size vectors from Stage I, $\widehat{\boldsymbol{\beta}_1^{\boldsymbol{\omega}}}$ and $\widehat{\boldsymbol{\beta}_2^{\boldsymbol{\omega}}}$, to impute ancestry-specific partial gene expression PRS (aspPS) in the manner of Marnetto et al. [156]. We can then add together the AFR- and EUR-specific partial components of GReX to create a combined PRS (casPS). We provide the details on these local ancestry-aware approaches to expression imputation, as well as a comparative description of the standard PRS approach to expression imputation that ignores local ancestry (PS), in the next few sections.

Local Ancestry-Aware Approaches (aspPS, casPS)

Using the notation of Section 4.2.2 and 4.2.3, recall we have assumed there are J total cis variants in the region of gene g and N_{adm} admixed testing subjects. Let $\boldsymbol{\Gamma}^{\boldsymbol{M}}(\boldsymbol{\Gamma}^{\boldsymbol{P}})$ be a $(N_{adm} \times J)$ matrix where the (i, j) element equals 1 if individual i has an AFR allele at jth SNP on his/her maternal (paternal) haplotype and 0 otherwise. We further define $\boldsymbol{X}^{\boldsymbol{M}}(\boldsymbol{X}^{\boldsymbol{P}})$ as the $N_{adm} \times J$ ancestry-standardized maternal (paternal) haplotype (0/1) matrix with $(\boldsymbol{X}^{\boldsymbol{M}})_{i,j} = \frac{x_{i,j}^M - 2f}{\sqrt{2f(1-f)}}$. Here, $x_{i,j}^M$ is the minor allele count of the ith individual at the jth variant on the maternal haplotype and f is the MAF of that variant in EUR or AFR, depending on the local ancestry of that SNP. Let $\widehat{\boldsymbol{\beta}_1^{\boldsymbol{\omega}}}(\widehat{\boldsymbol{\beta}_2^{\boldsymbol{\omega}}}) \in \mathbb{R}^J$ be the estimated eQTL effect size vector from a given PRS method ω using reference AFR (EUR) eQTL data. We can impute the AFR component of

GReX (aspPS_1^ω) and the EUR component of GReX (aspPS_2^ω) as follows:

$$\text{aspPS}_1^\omega = (\boldsymbol{X}_M \odot \boldsymbol{\Gamma_M} + \boldsymbol{X}_P \odot \boldsymbol{\Gamma_P})\widehat{\boldsymbol{\beta_1^\omega}} \tag{4.6}$$

$$\text{aspPS}_2^\omega = (\boldsymbol{X}_M \odot (1 - \boldsymbol{\Gamma_M}) + \boldsymbol{X}_P \odot (1 - \boldsymbol{\Gamma_P}))\widehat{\boldsymbol{\beta_2^\omega}} \tag{4.7}$$

In the above equations, the symbol $\odot$ indicates element-wise multiplication of matrices. For each PRS model ω, we then impute the total GReX in our admixed testing samples as the sum of these two ancestry-specific components:

$$\text{casPS}^\omega = \text{aspPS}_1^\omega + \text{aspPS}_2^\omega \tag{4.8}$$

Non-Ancestry-Aware Approach (PS)

We can also compare the performance of the ancestry-aware approaches to imputing GReX (aspPS_1, aspPS_2, casPS) to the standard PRS methodologies that ignore local ancestry in admixed individuals. For this, let $\boldsymbol{X}$ be the $N_{adm} \times J$ standardized (columns centered and scaled to unit variance, not ancestry-specific standardization) matrix of minor allele counts (0/1/2). We can impute GReX in our admixed testing data as:

$$\text{PS}_z^\omega = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{z}}^{\boldsymbol{\omega}} \tag{4.9}$$

Here, $\hat{\boldsymbol{\beta}}^{\boldsymbol{\omega}} \in \mathbb{R}^J$ is the estimated eQTL effect sizes vector from a given PRS method ω for a given set of eQTL summary statistics z. This standard approach could use the same reference population eQTL summary data as those used in the LA-aware approaches above ($z =$ AFR, EUR), or it could be trained using summary eQTL data from a sample of admixed individuals that is independent of our testing dataset ($z =$ ADMIX).

4.2.5 Stage II Gene-Trait Association Test

Once we have our all of our vectors of imputed gene expression in our admixed testing dataset $\{\text{aspPS}_1^\omega,\ \text{aspPS}_2^\omega,\ \text{caspPS}^\omega,\ \text{PS}_z^\omega\}$ for $\omega \in$ {P+T0.05, P+T0.001, lassosum} and $z \in$ {AFR, EUR, ADMIX}, we assume that our phenotype vector ($\boldsymbol{y} \in \mathbb{R}^{N_{adm}}$), for whom we want to estimate the association with GReX, has already been adjusted for the effects of important non-genetic confounders. We will perform simple linear or logistic regression of $\boldsymbol{y}$ on each of the aspPS, casPS, and PS imputed GReX vectors individually. For each ω and z, we thus obtain an ordinary least squares p-value. Using the approach employed by OTTERS [22], we can then combine these p-values in multiple ways using the Aggregated Cauchy Association Test (ACAT) [132]. First, we can aggregate across ω (our PRS models) to generate one p-value for each of our LA-aware approaches (aspPS$_1$, aspPS$_2$, casPS) and, similarly, one aggregate p-value for each of the standard PS approaches (PS). We refer to this as Level 1 aggregation. For example, to get the aggregate p-value for casPS, let p_ω be the p-value from the simple regression of $\boldsymbol{y}$ on casPS$^\omega$ for $\omega \in$ {P+T0.05, P+T0.001, lassosum}:

$$T = \sum_{\omega} k_\omega \tan\{(0.5 - p_\omega)\pi\} \tag{4.10}$$

$$p_{\text{casPS}} \approx 0.5 - \{\arctan(T / \sum_{\omega} k_\omega)\}/\pi \tag{4.11}$$

Here T is the ACAT test statistic. Authors assume that T approximately follows a Cauchy distribution and can therefore approximate p_{casPS}, the aggregated p-value. k_ω represent the combination weights, and for our purposes, we assume these weights to be equal across all PRS methods. We can then repeat this process (aggregating over PRS models ω) to also calculate $p_{\text{aspPS}_1}, p_{\text{aspPS}_2}, p_{\text{PS}_z}$.

We also propose to perform a second round of p-value aggregation using ACAT (Level 2). In this second round, we can aggregate over all LA-aware p-values p_{casPS},

p_{aspPS_1}, p_{aspPS_2}, or we may even elect to combine all LA-aware p-values with those from the standard PS approaches (e.g., aggregating over all $p_{\text{casPS}}, p_{\text{aspPS}_1}, p_{\text{aspPS}_2}, p_{\text{PS}_z}$).

4.2.6 Simulations

To evaluate the accuracy of our proposed ancestry-aware gene expression imputation approach, we performed extensive simulations. First, we chose our simulated admixed testing dataset to be two-way African and European recently admixed individuals. We used the 1000 Genomes (1KG) Phase 3 biallelic data in GRCh38 as source genomes to simulate multiple sets of admixed genomes [91]. We specifically chose CEU (Utah Residents with Northern and Western European Ancestry) as the European source population and YRI (Yoruba in Ibadan, Nigeria) as the African source population. In order to ensure we had large enough source populations to sample haplotypes from to generate our admixed testing set, we first expanded the YRI and CEU populations to size 10,000 each using `admix-kit` [164]. This tool generates additional sets of 1KG populations using the `HAPGEN2` framework [165].

We selected one gene, ABCA7, on chromosome 19 (1040101:1065571) as our target gene for simulations, as this gene has been used previously for the simulation stage of TWAS [75]. We assumed a window of 1MB upstream and downstream as the simulation region (66,358 variants). We generated different training datasets in which we first simulate gene expression and then subsequently calculate eQTL summary data. We generated AFR and EUR training datasets by randomly selecting 500 of our expanded sample of YRI and CEU subjects, respectively. We further generated several admixed training datasets of size 500 using the expanded sets of 10,000 YRI and CEU source haplotypes and the tool `haptools` [166]. We simulated three admixed training datasets according to different admixture generation parameters. First, we assumed an initial realistic African American demographic model of one pulse of admixture 9 generations ago with 80% contribution from YRI and 20% from CEU

(ADMIX 80 10g). We also considered two additional admixed training sets; the first assumed 5 admixture generations plus an initial AFR population frequency of 80% (ADMIX 80 5g) while the other assumed 5 admixture generations plus an initial AFR population frequency of 50% (ADMIX 50 5g).

To generate our admixed testing datasets, we again assumed the realistic African American demographic model of 10 admixture generations with 80% contribution from YRI (AFR) and 20% from CEU (EUR). These admixture generation settings for our testing dataset match those used to generate the ADMIX 80 10g training dataset. These simulation settings are similar to those used in previous methodological work in admixed individuals [145]. We simulated a total of 10,000 individuals to serve as our pool of admixed testing samples.

Using the simulated haplotypes and genotypes from the non-admixed reference AFR/EUR populations, we next simulated gene expression. In our training datasets (admixed and reference AFR/EUR of size 500), we first simulated gene expression according to the models described in Sections 4.2.2 and 4.2.3. We varied gene expression heritability in the AFR population ($h^2_{e,1}$) and in the EUR population ($h^2_{e,2}$) among $\{0.1, 0.2\}$, excluding the limited utility scenario where both are 0.1. We also varied the number of eQTLs (V) in both populations among $\{2, 10, 100\}$, the proportion of eQTLs that overlap between AFR and EUR populations (OP $= S/V$) among $\{0.5, 1\}$, and the correlation of effect sizes among shared AFR and EUR eQTLs (ρ) among $\{0.5, 1\}$. We also ensured that all SNPs selected as eQTLs have MAF > 0.05 in each of the training datasets. In total, we have 36 combinations of expression simulation parameters.

Since all of these settings consider overlap in eQTLs across ancestries, we took care when simulating correlated eQTL effect size vectors in each ancestry. Using the notation of Equations 4.2 and 4.3, we first randomly drew a temporary $\boldsymbol{\beta_1^*} \sim N(\boldsymbol{0}, \boldsymbol{I_V})$ and calculated the scale factor $\delta_1 \approx \sqrt{\frac{h^2_{e,1}}{V}}$. To simulate eQTL effect sizes in EUR, note that the first S are correlated with the first S elements of the effect size vector in AFR. Thus, we sampled a temporary $\boldsymbol{\beta_2^*}$ as below:

$$\boldsymbol{\beta_2^*} = \begin{pmatrix} \rho\boldsymbol{\beta^*_{1[1:S]}} + \sqrt{1-\rho^2}(\boldsymbol{z}) \\ N(\boldsymbol{0}, \boldsymbol{I_U}) \end{pmatrix} \text{ where } \boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I_S}) \tag{4.12}$$

Let the scale factor for EUR be $\delta_2 \approx \sqrt{\frac{h^2_{e,2}}{V}}$. Thus, for each simulation, the eQTL effect size vectors achieving the desired $h^2_{e,1}, h^2_{e,2}, \rho$, and OP are simply $\boldsymbol{\beta_1} = \delta_1\boldsymbol{\beta_1^*}$ and $\boldsymbol{\beta_2} = \delta_2\boldsymbol{\beta_2^*}$. We then used these ancestry-specific eQTL effect size vectors to simulate gene expression in our reference AFR/EUR training sets and, also, our admixed training/testing sets according to Equation 4.1.

Next, for each simulated training dataset (reference AFR/EUR and our independent admixed training samples), we calculated single-variant eQTL summary statistics and used the OTTERS pipeline to train PRS models. For each dataset, we retained summary data for only those variants with MAF $>$ 0.05 in the corresponding sample. Using these eQTL summary statistics, we then imputed gene expression in our admixed testing set in the manner described in Section 4.2.3. We compared the imputation R^2 (squared correlation between imputed and true expression in our admixed testing set) between our proposed LA-aware approaches $\{\text{aspPS}^\omega_1,\ \text{aspPS}^\omega_2,\ \text{caspPS}^\omega\}$, $\omega \in \{\text{P+T0.05, P+T0.001, lassosum}\}$, and the standard PRS approaches that ignore local ancestry $\{\text{PS}^\omega_z\}$ for $z \in \{\text{AFR, EUR, ADMIX 80 10g, ADMIX 80 5g, ADMIX 50 5g}\}$. For each round of GReX imputation using ancestry-aware approaches, we also performed another round of imputation where we assumed that 10% of the cis-variants in the region on both haplotypes of each individual in the testing set had incorrect local ancestry information.

Next, we performed power and type I error simulations for the Stage II test of GReX-trait association. We simulated the trait according to following equation where we assume the trait is a function of the total gene expression and is not a function of the ancestry-specific components of gene expression in Equation 4.1.

$$\boldsymbol{y} = \alpha \boldsymbol{E_g} + \boldsymbol{\epsilon_y},\ \boldsymbol{\epsilon_y} \sim N(\boldsymbol{0}, (1 - h_p^2)\boldsymbol{I_{N_{adm}}}) \tag{4.13}$$

For both our power and type I error simulations, we calculated the Level 1 ACAT p-values that aggregate over the three respective PRS models: $p_{\text{casPS}}, p_{\text{aspPS}_1}, p_{\text{aspPS}_2}$, and p_{PS_z}. We also then calculated the Level 2 ACAT p-values that aggregated over various combinations of these Level 1 ACAT p-values. For our power simulations, we considered α such that $h_p^2 \in \{0.025,\ 0.1\}$. For our null simulations to evaluate type I error rate, we assume this value was 0. For each of the 1,000 simulations performed for each of the 36 combinations of expression simulation parameters, we performed 10 trait simulations per h_p^2. Thus, for each expression simulation parameter combination, we performed 10,000 power and 10,000 type I error simulations.

4.2.7 Applied Analysis

UK Biobank Data

To assess the utility of our proposed methods in practice, we obtained individual-level genotype and phenotype data from admixed individuals in the UK Biobank (UKB). The UKB is a large-scale biomedical database housing data collected from approximately 500,000 individuals across the UK. This study allows for widespread investigation of the genetic variation associated with hundreds of lifestyle and health factors. To best mimic the settings of our simulation study, we elected to focus our analysis on two-way admixed individuals of African and European ancestry in the UKB. In order to identify these individuals making up our testing dataset, we first performed preliminary subject filtering. Specifically, we excluded subjects who had subsequently withdrawn from the study, those who were marked as "redacted", and those who were marked as outliers based on pre-calculated metrics of heterozygosity and missing rates. We also excluded subjects with putative sex aneuploidy, those with high pre-calculated estimates of relatedness, and those whose genetically-inferred gender differed from their submitted gender. We then removed individuals falling in the "White British subset". These individuals were previously identified by a combination of both self-reported ancestry and genetic PCs. We also excluded individuals who have missing self-reported ethnicity or whose self-reported ethnicity fell among the following: White, Irish, British, Any other white background. Following this, 27,491 subjects remained.

We then performed principal component analysis (PCA), projecting these filtered UKB individuals onto genetic PCs anchored in 1000 Genomes (1KG) data [91]. For this, we first subset 1KG genotype data to include unrelated individuals from the following populations: African (ACB [African Caribbeans in Barbados] and ASW [Americans of African Ancestry in SW USA] excluded) (503), Admixed American (347), East Asian (504), European (503), and South Asian (489). We restricted variants to non-ambiguous SNPs, those found in the UKB GWAS data, those with MAF > 0.05 in each population, those with HWE $p > 1 \times 10^{-6}$. We then pruned remaining variants (window size 1000 bp, step size 50 variants, R^2 threshold 0.1). Using the loadings for the top 10 PCs trained in 1KG samples, we projected the UKB self-reported non-White individuals (27,491) into this space (Appendix Figure 4.16). Following the approach of Atkinson et al. [145], using the 1KG data, we then trained a random forest classifier to predict continental ancestry (1KG population) from the top 10 PCs. We then applied this random forest model to our UKB sample. We excluded any individuals with $< 50\%$ estimated probability of African ancestry. Using the top 3 PCs in the 9,187 meeting this criteria, we constructed a 95% ellipsoid along the African-European cline (Appendix Figure 4.17). We kept the 8,752 UKB individuals lying within the ellipsoid. Finally, we excluded two additional subjects with self-reported Asian/Asian British and White and Asian ethnicity. The remaining

8,750 individuals made up our final two-way, AFR and EUR admixed testing dataset.

Genotype data on these subjects was generated using either the UK BiLEVE or UK Biobank Axiom arrays. Prior to imputation, we followed the pre-imputation quality control pipeline provided at https://www.well.ox.ac.uk/ wrayner/tools/#Checking, using variant data from TOPMed Freeze3a on GRCh37/hg19. We performed imputation, liftover to GRCh38, and phasing using the TOPMed Imputation Server [167–169]. Next, we prepared our AFR and EUR reference data for local ancestry deconvolution of our UKB genotypes. First, we imputed missing genotypes in the 1KG Phase 3 biallelic phased GRCh38 data of AFR and EUR subjects using `BEAGLE`, again excluding ASW and ACB [170]. Using this as our reference population genotype data, we performed local ancestry inference in our admixed UKB testing set using `FLARE` [171]. We assumed 10 generations since admixture (10 admg).

eQTL Summary Data

For our European (EUR) reference eQTL dataset, we downloaded cis-eQTL summary statistics in whole blood from the GTEx V8 dataset (dbGaP phs000424.v8.p2), where cis-eQTL analysis was performed in 570 European-American subjects. To briefly summarize the analysis performed, authors adjusted RNA sequencing for the effects of top 5 PCs, top 60 PEER factors, sequencing platform (Illumina HiSeq 2000 or HiSeq X), sequencing protocol (PCR-based or PCR-free), and sex. For the eQTL analysis, they then restricted genes to those with > 0.1 TPM and ≥ 6 reads in at least 20% of the data samples, and they normalized expression vectors using TMM [98] and inverse normal transformed. SNPs from WGS data with MAF $\geq 1\%$ were retained. Authors performed single-variant cis-eQTL analysis using `FastQTL` [172] and a 1MB window from the transcription start site of each gene.

For our African (AFR) reference eQTL dataset, we used publicly-available whole-blood cis-eQTL summary statistics from a subset of high-African ancestry admixed individuals [133]. In this study, authors mapped ancestry-specific gene expression signatures in 2,733 individuals (African American, Puerto Rican, and Mexican American) from the Genes-Environments and Admixture in Latino Asthmatics (GALA II) study and the Study of African Americans, Asthma, Genes, and Environments (SAGE). Authors obtained paired RNA sequencing data and WGS data. RNA and WGS data processing are described elsewhere [133]. Authors used CEU and YRI HapMap reference genotypes, as well as Indigenous American ancestry reference genotypes, to estimate global measures of ancestry with `ADMIXTURE` [173]. Authors defined a high global African ancestry subset as those with $> 50\%$ estimated global African ancestry (721). They then performed ancestry-specific eQTL analysis in these subjects using a 1MB cis-window from the transcription start site of each gene and `FastQTL` [172]. Analyses adjusted for expression for age, sex, asthma status, top 5 PCs, and 60 PEER factors.

TWAS

As our TWAS traits in our UKB testing dataset, we considered 29 widely-collected blood biomarkers. We first log-normalized raw trait measurements and then we obtained covariate-adjusted phenotypes by taking the residuals of linear regression models of each log trait on the top 20 PCs, sex, age at recruitment, and smoking status (prefer not to answer, never, previous, current). Using our AFR and EUR reference eQTL summary data described above, we first removed ambiguous SNPs (A/T,T/A,G/C,C/G) and only kept eQTL data from SNPs with MAF > 0.01 in each respective sample. We then performed LD clumping using a R^2 threshold of 0.99. Next, we trained AFR and EUR ancestry-specific PRS models of gene expression using OTTERS and P+T0.001, P+T0.05, and lassosum models. We then imputed gene expression in UKB samples using LA-aware approaches (casPS, aspPS) and standard PS approaches (PS AFR, PS EUR). We concluded our analyses by performing simple linear regression analysis for the association of each imputed gene expression vector with each of the 29 adjusted blood biomarker traits. Multiple p-value aggregation approaches via ACAT were considered.

4.3 Results

4.3.1 Expression Imputation Accuracy

To evaluate the GReX imputation accuracy of proposed local-ancestry aware approaches compared to the standard PRS imputation approaches of OTTERS, we used as our testing set a collection of $N_{adm} = 1000, 5000, 10000$ simulated admixed individuals (10 admixture generations, 80% initial contribution of AFR [1KG YRI] haplotypes, 20% contribution of EUR [1KG CEU] haplotypes). We computed our LA-aware measures of GReX using simulated eQTL summary data from a reference AFR and a reference EUR sample (each $N_{ref} = 500$). For comparison, we also computed the standard LA-unaware GReX using each of these reference samples (AFR PS, EUR PS). We also computed a vector of standard PSs using eQTL summary datasets from independent simulated admixed samples of varying admixture parameters. The "ADMIX 80% 10g" PSs represent the GReX we would have imputed had we had access to eQTL data from an independent admixed sample of exactly the same ancestry as our admixed testing set (10 admg, 80% initial AFR contribution). We use the other training datasets with 5 admg and 50-80% initial AFR contribution to examine the impract of "mismatch" between admixed training and test datasets on imputation accuracy. In Figure 4.1, we present the squared correlation (R^2) between imputed GReX and true gene expression in our admixed testing subjects. This figure corresponds to gene expression heritability in both Africans and Europeans ($h^2_{e,1},\ h^2_{e,2}$) of 0.2 and testing sample size of 10,000. We provide imputation R^2 results for other expression heritability and eQTL architecture settings for $N = 10000$

in the Appendix (Figures 4.6-4.7), as imputation accuracy patterns did not differ dramatically by testing sample size.

From these figures, we see a few general trends. Across all imputation approaches, both LA-aware and unaware, as well as across all PRS methods (P+T0.001, P+T0.05, lassosum), we see higher R^2 values for the sparse eQTL scenario (2) compared to the scenarios where the number of causal eQTLs is 10 or 100. There may exist other PRS approaches (PRScs, for example [174]) that may perform better for the scenario in which we expect larger number of true eQTLs. Next, we see that accuracy of our proposed casPS approach, using pruning and thresholding PRSs, tends to be slightly higher than the optimal standard PS approach using admixed eQTL summary data from a perfect ancestry-matched admixture cohort (ADMIX 80 10g PS) when the eQTLs are not exactly the same between AFR and EUR or when the correlation of shared eQTL effect sizes is less than 1. In other words, these simulations suggest that even if we had access to eQTL data from a cohort exactly matched for ancestry to our admixed testing cohort, we would still achieve as high or higher imputation accuracy using reference population eQTL datsets. Across all simulation settings, lassosum performed inconsistently with no notable performance patterns by imputation approach. In the scenario where eQTL architecture for the given gene is identical between AFR and EUR populations (OP = 1, $\rho = 1$), the imputation R^2 appears quite similar between the caPS LA-aware approach and ADMIX 80 10g PS. Furthermore, across all settings, we see a downward trend for R^2 when using standard PS approaches and admixed training data as the training samples become increasingly different from the testing dataset in terms of admixure generation parameters. As the number admixture generations decreases (10 to 5) and the proportion of initial AFR donors decreases (80% to 50%), we generally see less successful GReX imputation.

In Appendix Figures 4.8-4.10, for our LA-aware approaches, we show the impact of LA misclassification of 10% of SNPs in the region of ABCA7 that were eligible for

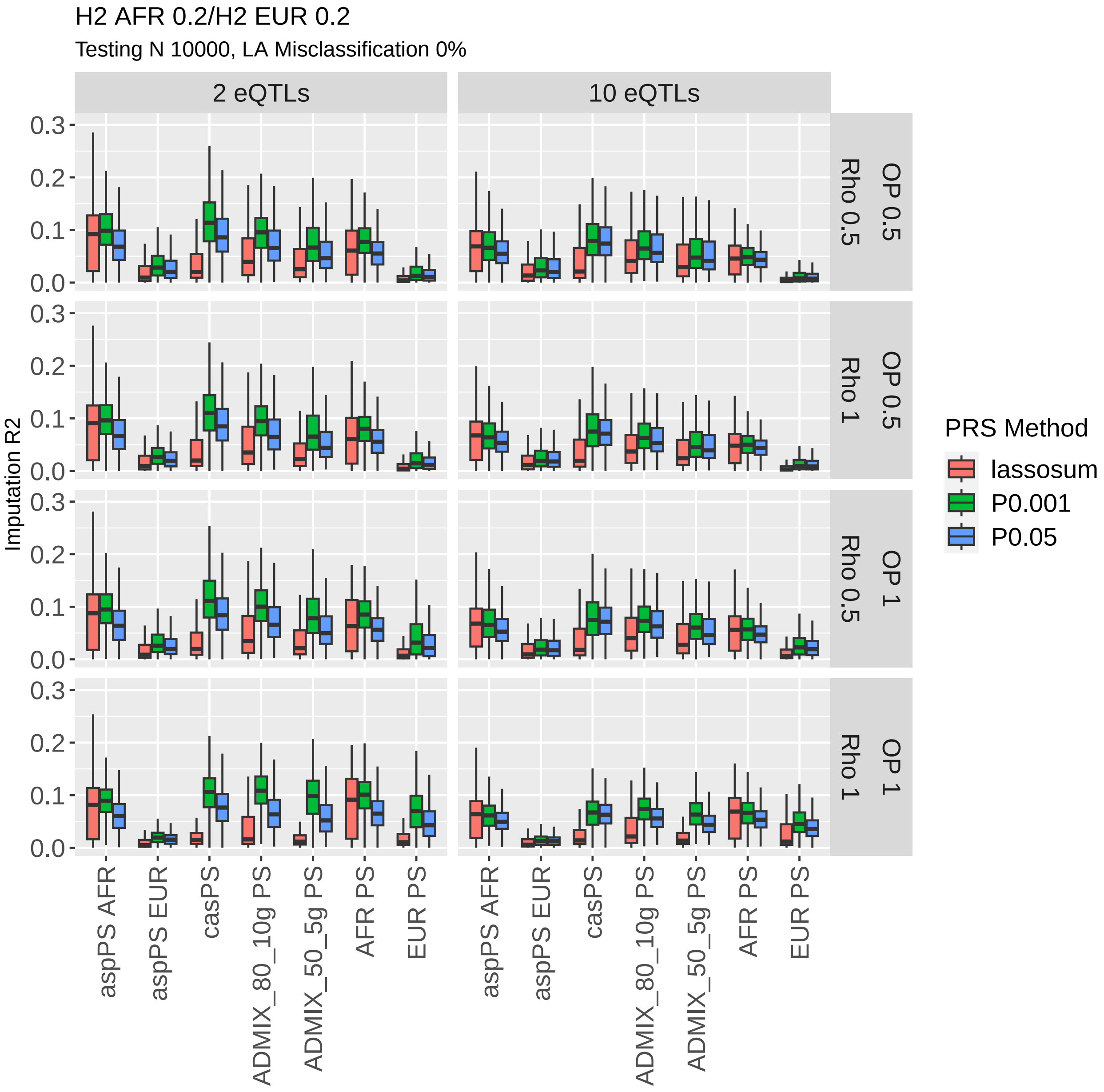

Figure 4.1: Gene expression imputation accuracy in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR, 20% initial contribution from EUR) for expression heritability $h^2_{e,1}$, $h^2_{e,2} = 0.2$. Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indicate the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). The x-axis shows the GReX imputation approach, including our proposed local-ancestry aware methods (aspPS, casPS) and standard PRS imputation approaches (PS). For ancestry-aware methods, we assume no local ancestry misclassification. Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

modeling gene expression. In other words, we assumed 10% of the cis-SNPs of this gene with MAF > 0.05 in the corresponding training datasets (Ref AFR or Ref EUR) had incorrect LA tags (AFR or EUR). We further assumed misclassification of these SNPs on both maternal and paternal haplotypes. We note this likely on the high end for misclassification rates, as popular local ancestry imputation algorithms (e.g., `FLARE`, `MOSAIC`, `RFMix`) have high imputation accuracy for reasonably sized training panels used to infer LA. For example, the squared correlation between inferred and true local ancestry dosage $r^2 \in [0.87,\ 0.96]$ for the three methods mentioned in three-way admixed samples using reference panel sizes of N=400 [171]. Similarly, another benchmarking paper cited `RFMix` as having 89% classification accuracy in the highly complex scenario of a five-way admixed sample [175]. Regardless, we do not see a marked drop in the quantiles of imputation R^2 of our proposed LA-aware GReX imputation approaches (aspPS AFR, aspPS EUR, casPS) across PRS models.

4.3.2 Type I Error Rate

Next, we assessed the type I error rate of the Stage 2 gene-trait association tests. For these simulations, we assumed a null association between true gene expression and each simulated phenotype, i.e., a phenotypic heritability $h_p^2 = 0$. As the type I error rates did not differ dramatically by the gene expression simulation parameters ($h_{e,1}^2,\ h_{e,2}^2,\ \rho$, OP, number of eQTLs), we present quantile-quantile (QQ) plots for p-values across each of the 36 simulation settings, corresponding to a total of 36,000 null simulations per plot.

As we see, the p-values resulting from Level 1 aggregation across PRS models (P+T0.001, P+T0.05, lassosum) for the LA-aware approaches (aspPS AFR, aspPS EUR, casPS) show the expected distribution under the null when we assume no LA misclassification (Figure 4.2) and 10% LA misclassification (Appendix Figure 4.11). Similarly, the Level 1 p-value aggregation for the standard imputation approaches

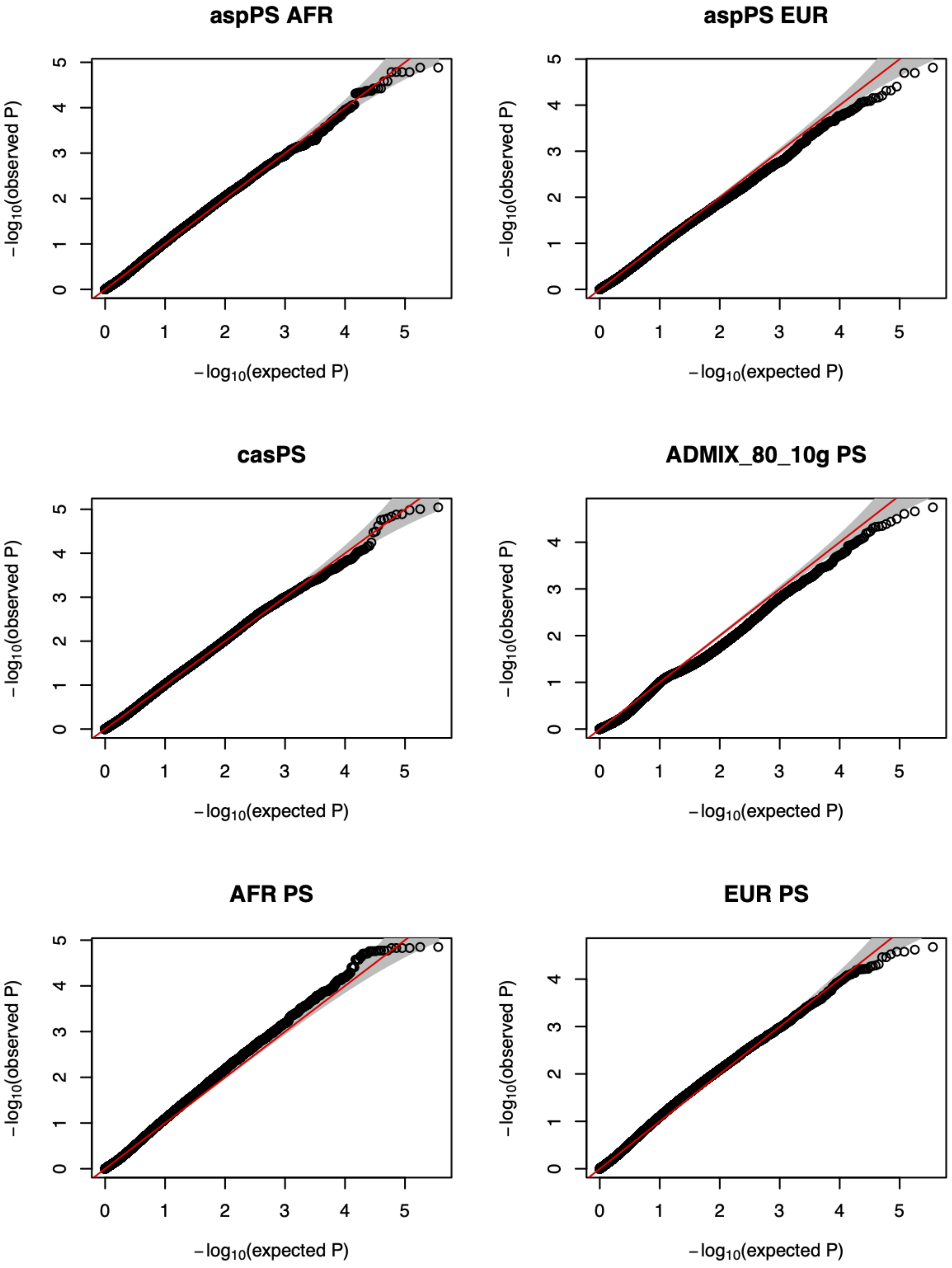

Figure 4.2: QQ plots of p-values from gene-level association tests from both LA-aware and LA-unaware GReX imputation approaches under the null when no association of expression with trait exists. Here, we assume a testing sample size of 10,000. These p-values represent Level 1 p-value aggregation by ACAT, i.e., aggregation of p-values across the three PRS models (P+T0.001, P+T0.05, lassosum). For ancestry-aware methods, we assume no local ancestry misclassification. Each plot shown corresponds to 36,000 total simulations, including all 36 gene expression simulation settings.

using reference eQTL data (PS AFR, PS EUR) also appear to maintain appropriate rates of type I error. We do, however, see a slight deflation of Level 1 aggregated p-values for the standard PS approaches using eQTL data in an independent set of admixed subjects (middle right plot). To assess whether we see any inflation of our gene-trait association p-values at the second level of p-values aggregation, i.e., aggregating Level 1 p-values for our LA-aware approaches and/or standard PS approaches, we constructed a second set of QQ plots in Figure 4.3. Again, we see the expected distribution under the null even when we combine aspPS and casPS p-values with those from the reference-derived PSs (AFR PS, EUR PS) and, further, with the p-values from the independent admixed-derived PSs (middle and right panels, respectively). This applies to both assumptions of 0% (top row) or 10% LA misspecification (bottom row).

4.3.3 Power

Simulation results comparing the performance of our proposed LA-aware approaches to the competing standard GReX imputation approaches under the assumption of a true gene-trait association are summarized in Figure 4.4 (2 eQTLs) and Figure 4.5 (10 eQTLs). These figures reflect normally distributed traits, a testing set sample size of 10,000, significance level $\alpha = 5 \times 10^{-5}$, and no LA misspecification. When comparing the performance of the Level 1 approaches (aggregation across PRS), our proposed LA-aware casPS has higher power than any of the LA-unaware standard PS approaches, including that trained using a perfectly matched admixed sample, when gene expression genetic architecture is different between populations. The power of Level 1 casPS is similar to that of standard PS with a perfectly matched admixed sample (ADMIX 80 10g) when the gene expression genetic architecture is exactly the same between AFR and EUR groups (OP = 1, $\rho = 1$). Across all simulation settings, the Level 2 aggregation approach utilizing LA-aware GReX (casPS, aspPS

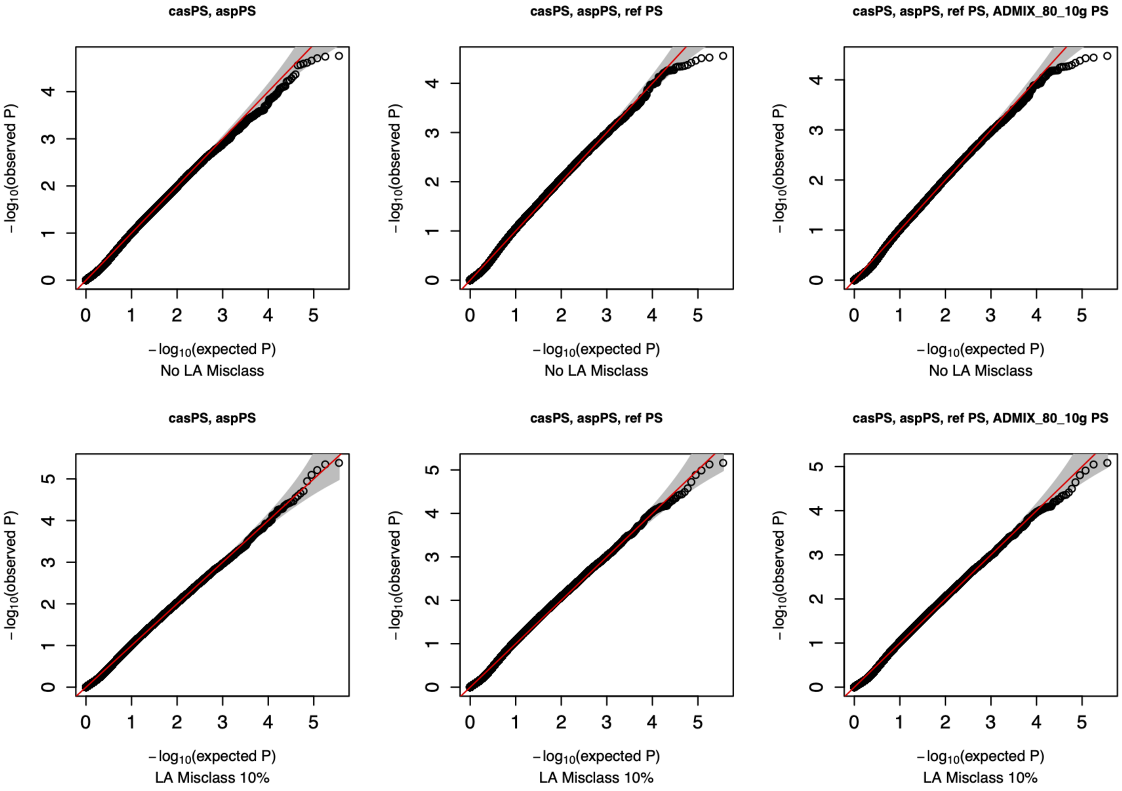

Figure 4.3: QQ plots of p-values from gene-level association tests from LA-aware GReX imputation approaches under the null when no association of expression with trait exists. Here, we assume a testing sample size of 10,000. These p-values represent Level 2 p-value aggregation by ACAT. The p-value aggregation approach is indicated in the plot title. For ancestry-aware methods, we assume either no local ancestry misclassification (No LA Misclass) or misclassification of 10% of SNPs in the gene region (LA Misclass 10%). Each plot shown corresponds to 36,000 total simulations, including all 36 gene expression simulation settings.

AFR, aspPS EUR) and the standard PSs built using reference data (EUR PS, AFR PS) achieved the highest power (dark blue) and is therefore the preferred method going forward.

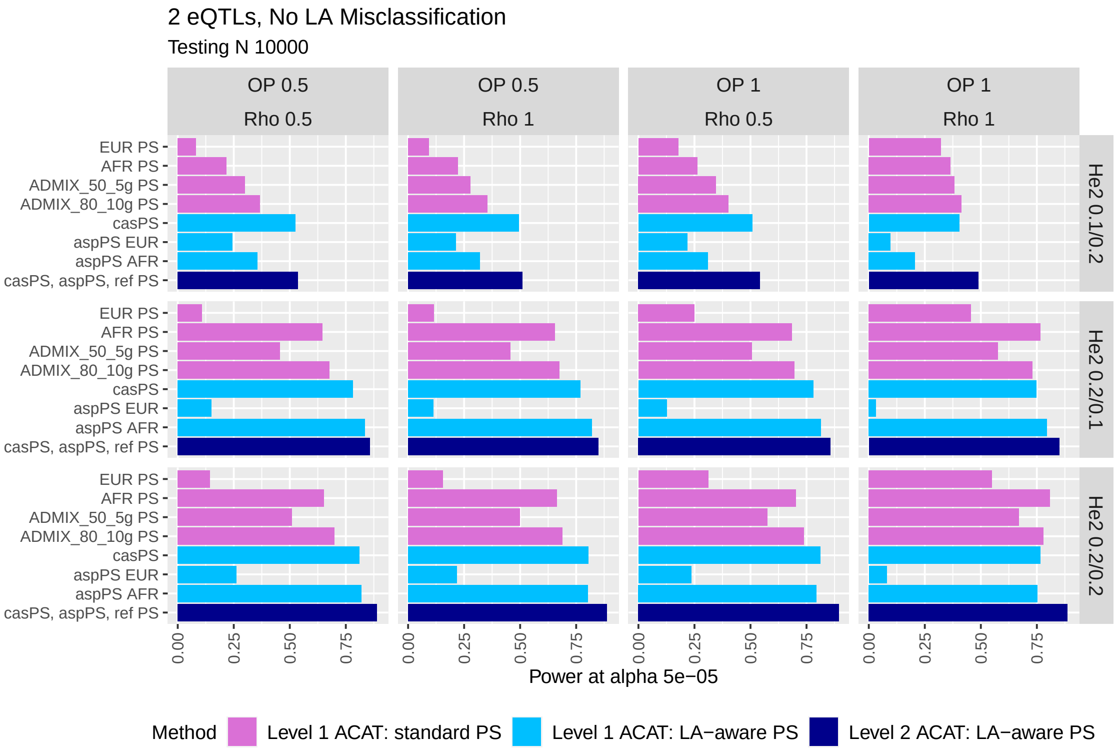

Figure 4.4: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5\times10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 2 eQTLs, no local ancestry (LA) misclassification for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of LA-unaware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

In Appendix Figures 4.12 (2 eQTLs) and 4.13 (10 eQTLs), we illustrate the power of the LA-aware approaches assuming a high rate of LA misclassification (10%) for cis-SNPs in the gene region. With the exception of the scenario in which we have

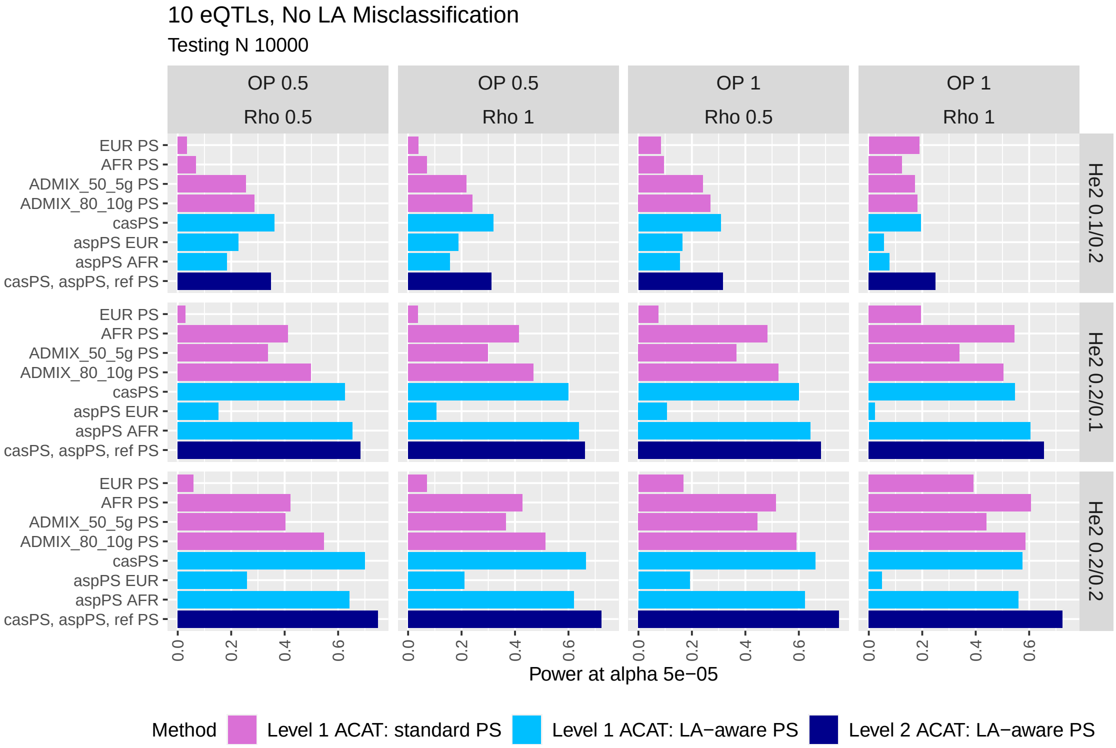

Figure 4.5: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5\times10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 10 eQTLs, no local ancestry (LA) misclassification for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of LA-unaware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

a higher lever of gene expression heritability in Europeans ($h^2_{e,2} = 0.2$) and lower level of expression heritability in Africans ($h^2_{e,1} = 0.1$) and assume 10 eQTLs, we still observe the Level 2 ACAT LA-aware approach achieving highest power. For the anomalous setting mentioned, we note that gene-expression imputation accuracy here is quite low across the board (Appendix Figure 4.7), and thus power is subsequently low for all LA-aware and LA-unaware approaches. This is similar to the scenario where we assume 100 eQTLs for the gene under study. We have comparatively less accurate GReX imputation, and therefore it is unsurprising that we see low power in the downstream association tests across all methods (Appendix Figures 4.14-4.15).

4.3.4 Applied Data Analysis

We applied several approaches to detect genes associated with blood biomarkers by way of genetically-regulated transcriptional activity in a subset of 8,750 two-way African/European admixed subjects in the UK Biobank. We first performed LA deconvolution using `FLARE` and reference genotypes from AFR and EUR cohorts from 1000 Genomes. Through this, we estimated an overall global (genome-wide) proportion of AFR genotypes of 86.6% and an estimated proportion of EUR genotypes of 13.4%. Using our two sets of European-derived and high-African ancestry-derived eQTL summary data in whole blood, we trained our AFR and EUR GReX imputation models for 14,614 genes. Specifically, we used lassosum and pruning and thresholding models, first pruning variants with an $R^2_T = 0.99$ and then filtering by p-value thresholds of $P_T = 0.05, 0.001$. For each PRS model, we then imputed GReX in our admixed UKB testing set in the standard OTTERS approach separately using our AFR-trained models (PS AFR) and EUR-trained models (PS EUR). We also imputed GReX in our proposed LA-aware approach using both sets of PRS models and our inferred LA information. We calculated two ancestry-specific components of gene expresion (aspPS AFR, aspPS EUR) and one vector representing the sum of these two components (casPS).

We tested for the association of each of the vectors of imputed GReX with 29 blood biomarkers: 4 bone and joint traits (alkaline phosphatase, calcium, rheumatoid factor, vitamin D), 8 cardiovascular traits (apolipoprotein A and B, C-reactive protein, cholesterol, HDL cholesterol, LDL cholesterol, lipoprotein A, triglycerides), 2 diabetes-related traits (glucose, HbA1c), 3 hormone traits (insulin-like growth factor 1 [IGF-1], sex hormone binding globulin [SHBG], testosterone), 6 liver-related traits (alanine aminotransferase, albumin, aspartate aminotransferase, direct bilirubin, γ glutamyltransferase, total bilirubin), and 6 traits related to renal function (creatinine, cystatin C, phosphate, total protein, urate, urea). For each approach (PS AFR, PS EUR, aspPS AFR, aspPS EUR, casPS), we calculated a Level 1 ACAT p-value, aggregated across PRS models. We also combined subsets of these Level 1 p-values to calculate two Level 2 ACAT p-values (casPS+aspPS, casPS+aspPS+PS).

We assessed significance at a Bonferroni-adjusted level of $0.05/(29 * 14614) = 1.18 \times 10^{-7}$, adjusting for the total number of phenotypes and number of genes considered. Across all tests and imputation approaches, we identified 265 significant gene-trait associations. We identified associations for 15/29 traits (51.7%), with the most associations observed for SHBG (66), lipoprotein A (31), total bilirubin (29), γ glutamyltransferase (26), and direct bilirubin (24). While 265 significant associations were found, most gene-trait associations were unsurprisingly picked up by more than one imputation approach, and thus we ultimately identified 71 unique gene-trait pairings. In order to evaluate the utility of our LA-aware approaches, we compare and contrast the total number of unique gene-trait associations identified in Figure 4.18 (Appendix). Leveraging our LA-derived ancestry-specific p-values (aspPS AFR/EUR) and the p-values from their combined component (casPS), this Level 2 aggregation successfully identified 68/71 associations (95.8%). In Table 4.1, we present the 15 gene-trait associations identified by this casPS+aspPS approach that were not identified by standard GReX imputation using reference eQTL summary data (PS AFR, PS EUR). We provide the full lists of all gene-trait associations identified by each individual approach in the Appendix (Tables 4.2-4.6).

Of these 15 genes, 14 have consistent evidence in the GWAS literature, with each harboring one or more significant GWAS variant ($p < 5 \times 10^{-8}$) for the relevant traits. However, one association (HNRNPH for lipoprotein A) is not located near known GWAS loci [1]. While HNRNPH is a pseudogene, and thus the biological mechanisms behind this association are thus unclear, previous work has helped elucidate the role that pseudogenes can play in the development of cardiovascular disease [176].

Next, we used the online database TWAS Atlas to examine prior documented TWAS associations for each of these 15 genes [177]. Some of the 15 genes identified exclusively by our approach have been implicated in previous TWAS of relevant traits, while others represent potentially new TWAS findings. For example, we did not find prior TWAS associations of EIF4E2 with traits relevant to total bilirubin, whereas ATG16L1 has been previously implicated in a TWAS of Crohn's disease [178], which is associated with low serum bilirubin [179]. While many of the genes identified by our approach for lipoprotein A have been implicated by previous TWAS of lipid biomarkers of cardiovascular disease, CD36, on the other hand, has only been implicated in a whole blood TWAS of body mass index (BMI) [180]. The four genes associated with SHBG in Table 4.1 were identified in association with endometriosis in a recent TWAS [181]. SHBG plays a role in the availability of sex hormones in the body, and increased levels of SHBG were observed among women with endometriosis compared to controls [182]. Genes CABIN1, LRRC75B, XRCC6, and HNRNPH1P1 did not have any significant prior TWAS associations with biomarker-relevant phenotypes.

Table 4.1: Associations identified in UKB blood biomarker analysis by Level 2 p-value aggregation of LA-aware casPS and aspPSs that were not identified using standard GReX imputation using PS trained in reference AFR and EUR eQTL summary data. The p-values shown below represent ACAT(casPS p, aspPS AFR p, aspPS EUR p).

Phenotype	Group	Gene	Name	Chr	Pos	p
Total bilirubin	Liver	ENSG00000135930	EIF4E2	2	232550593	4.01E-08
Direct bilirubin	Liver	ENSG00000085978	ATG16L1	2	233210051	3.16E-08
Lipoprotein A	Cardiovascular	ENSG00000175048	ZDHHC14	6	157381133	5.98E-08
Lipoprotein A	Cardiovascular	ENSG00000220305	HNRNPH1P1	6	159712801	2.65E-10
Lipoprotein A	Cardiovascular	ENSG00000146457	WTAP	6	159725585	4.23E-08
Lipoprotein A	Cardiovascular	ENSG00000175003	SLC22A1	6	160121789	1.06E-08
Lipoprotein A	Cardiovascular	ENSG00000213071	LPAL2	6	160453428	4.49E-09
Apolipoprotein A	Cardiovascular	ENSG00000135218	CD36	7	80369575	1.08E-07
SHBG	Hormone	ENSG00000170175	CHRNB1	17	7445061	1.98E-08
SHBG	Hormone	ENSG00000209582	SNORA48	17	7574713	5.07E-09
SHBG	Hormone	ENSG00000129244	ATP1B2	17	7646627	1.43E-12
SHBG	Hormone	ENSG00000141499	WRAP53	17	7686372	3.15E-17
Gamma glutamyltransferase	Liver	ENSG00000099991	CABIN1	22	24011192	1.01E-07
Gamma glutamyltransferase	Liver	ENSG00000178026	LRRC75B	22	24585620	3.84E-10
Creatinine	Renal	ENSG00000196419	XRCC6	22	41621119	6.15E-09

4.4 Discussion

In this project, we introduce a novel method for performing transcriptome-wide association analysis in admixed subjects. Genomes of admixed individuals are a mosaic of two or more distinct ancestral groups, and thus local ancestry tract information within each haplotype can be leveraged to improve power in genetic association analyses when causal variant effect sizes differ between populations. This method represents an important contribution to the small but growing catalog of statistical methods dedicated to admixed individuals as it does not require individual-level genotype and gene expression data, as is typically needed in Stage 1 GReX model training in TWAS. Here, we build upon a recently published TWAS approach, OTTERS, that uses well-established PRS models designed for GWAS summary data and applies them to eQTL summary data to impute gene expression in a tissue of interest [22]. In our proposed method, we use multiple sets of eQTL summary statistics, namely those derived from the distinct parent ancestral groups of our admixed testing sample. The framework is flexible enough to also incorporate eQTL summary data from admixed cohorts independent of the testing set. We assessed the performance of our method using both real and simulated admixed data and compare to standard TWAS approaches designed for homogeneous populations that ignore LA information. Our proposed approach achieved higher power compared to standard GReX imputation approaches for the majority of simulation settings when LA imputation is accurate, and in our applied analyses, we were able to identify 15 additional genes that were not found by LA-unaware approaches.

Through our simulations, we first demonstrate that p-values of all variants of our proposed approach yield the expected distribution under the null assumption of no gene-trait association. These variants include two approaches of p-value aggregation: Level (1), wherein we combine p-values from the three PRS models, and Level (2), in which we aggregate some subset of p-values from Level (1). Next, through our power simulations, we make several important observations. First, under all scenarios of 10 or fewer eQTLs , both our Level (1) casPS and Level (2) (casPS+aspPS+PS) aggregation methods achieve superior power to any standard LA-unaware approach when the genetic architecture of gene expression differs between AFR and EUR populations. In fact, there is growing evidence for such ancestry-specific genetic architecture patterns, and estimated gene expression heritability has also been shown to differ significantly by local ancestry at the transcription start site of genes among admixed subjects [133]. Second, even when genetic architecture patterns are identical between ancestries, our LA-aware Level (1) casPS and Level (2) (casPS+aspPS+PS) generally achieve greater power or power comparable to if we employed standard GReX imputation using eQTL summary data from a perfectly ancestrally-matched (number of admixture generations, initial AFR/EUR contribution proportions) independent admixed sample. Third, we also see that our Level (2) approach still performs competitively in these settings when we assume a large number (10%) of local ancestries are misclassified in the gene region.

Finally, in our applied analysis, we use real-world eQTL summary data from a European sample and a high-African ancestry sample of African Americans to perform TWAS of 29 blood biomarker traits. Here, our testing set is two-way African and European admixed subjects from the UK Biobank. We successfully identified 15 significant gene-trait associations using our Level 2 p-value aggregation approach (casPS + aspPSs) that were not picked up using standard GReX PS imputation methods that ignore local ancestry (PS AFR, PS EUR). 14 of these genes are consistent with GWAS loci previously identified for the corresponding traits. However, one gene, HNRNPH on chromosome 6, represents a potentially novel locus for lipoprotein A and was not identified by standard GReX imputation.

We observe several limitations of our present work. We first note that while our method demonstrates desirable power levels for modest trait heritability, the gene expression imputation accuracies are lower across all simulation settings than the assigned true indicated gene expression heritability levels. We argue, however, that our imputation models are trained using PRS approaches to leverage the widespread availability of eQTL summary data, and R^2 has been shown to fall below true trait heritability for common PRS approaches [22, 174], and we expect imputation accuracy to be even lower when ancestry-specific effects are at play. We argue that we may further improve imputation accuracy by including more PRS approaches to our method, and, importantly, those designed precisely for admixed samples (e.g., GAUDI [158]). Additionally, in this project, we only consider two-way admixed individuals for both our simulated analyses and applied work in the UK Biobank. We believe that we can easily extend this approach to allow for three-way or higher levels of admixture, and that the approach is computationally efficient enough to implement this in practice. Finally, we designed our method to utilize eQTL data from ancestrally homogeneous parent ancestral groups. However, when seeking to apply our method to admixed individuals of African ancestry, we note there is a marked paucity of eQTL studies performed in non-admixed African cohorts. The eQTL summary data from African

American individuals with $>$ 50% African ancestry-alleles currently represents the best surrogate dataset of reasonable sample size for our analysis, and this substitution has similarly been employed in another recent methodological paper [158].

4.5 Appendix

4.5.1 Tables

Table 4.2: Associations identified in UKB blood biomarker analysis by Level 2 p-value aggregation of LA-aware casPSs and aspPSs. The p-values shown below represent ACAT(casPS p, aspPS AFR p, aspPS EUR p).

Phenotype	Gene	Name	Chr	Pos	p
Apolipoprotein B	ENSG00000134222	PSRC1	1	109279556	2.86E-11
C-reactive protein	ENSG00000158716	DUSP23	1	159780932	3.05E-10
C-reactive protein	ENSG00000272668	RP11-190A12.8	1	159866954	5.28E-16
C-reactive protein	ENSG00000279430	RP11-190A12.9	1	159910094	5.76E-09
Total bilirubin	ENSG00000135930	EIF4E2	2	232550593	4.01E-08
Direct bilirubin	ENSG00000085978	ATG16L1	2	233210051	3.16E-08
Total bilirubin	ENSG00000085978	ATG16L1	2	233210051	3.55E-22
Direct bilirubin	ENSG00000251791	SCARNA6	2	233288676	1.23E-24
Total bilirubin	ENSG00000251791	SCARNA6	2	233288676	9.79E-50
Direct bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	8.51E-21
Total bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	1.12E-43
Direct bilirubin	ENSG00000077044	DGKD	2	233354507	9.35E-53
Total bilirubin	ENSG00000077044	DGKD	2	233354507	5.98E-77
Direct bilirubin	ENSG00000085982	USP40	2	233475520	6.44E-27
Total bilirubin	ENSG00000085982	USP40	2	233475520	4.31E-58
Direct bilirubin	ENSG00000123485	HJURP	2	233833424	3.99E-28
Total bilirubin	ENSG00000123485	HJURP	2	233833424	1.30E-49
Urate	ENSG00000261490	RP11-448G15.3	4	10068089	3.44E-13
Urate	ENSG00000178163	ZNF518B	4	10439874	4.96E-12
Alkaline phosphatase	ENSG00000112293	GPLD1	6	24428177	1.08E-26
Lipoprotein A	ENSG00000175048	ZDHHC14	6	157381133	5.98E-08
Lipoprotein A	ENSG00000122335	SERAC1	6	158109515	3.90E-09
Lipoprotein A	ENSG00000218226	TATDN2P2	6	158609706	9.25E-09
Lipoprotein A	ENSG00000164691	TAGAP	6	159034468	2.24E-10
Lipoprotein A	ENSG00000220305	HNRNPH1P1	6	159712801	2.65E-10
Lipoprotein A	ENSG00000146457	WTAP	6	159725585	4.23E-08
Lipoprotein A	ENSG00000175003	SLC22A1	6	160121789	1.06E-08
Lipoprotein A	ENSG00000213071	LPAL2	6	160453428	4.49E-09
Lipoprotein A	ENSG00000026652	AGPAT4	6	161129979	6.58E-17
Alkaline phosphatase	ENSG00000135218	CD36	7	80369575	5.28E-16
Apolipoprotein A	ENSG00000135218	CD36	7	80369575	1.08E-07
SHBG	ENSG00000148572	NRBF2	10	63133247	6.99E-10
SHBG	ENSG00000165476	REEP3	10	63521363	6.63E-11
Apolipoprotein A	ENSG00000118137	APOA1	11	116835751	1.32E-08
Phosphate	ENSG00000047621	C12orf4	12	4487728	1.60E-11
Apolipoprotein B	ENSG00000182149	IST1	16	71885233	1.08E-07
SHBG	ENSG00000169992	NLGN2	17	7404874	2.09E-11
SHBG	ENSG00000181284	TMEM102	17	7435443	7.35E-13
SHBG	ENSG00000170175	CHRNB1	17	7445061	1.98E-08
SHBG	ENSG00000239697	TNFSF12	17	7548891	1.70E-18
SHBG	ENSG00000161955	TNFSF13	17	7558292	5.99E-30
SHBG	ENSG00000161960	EIF4A1	17	7572706	1.38E-10
SHBG	ENSG00000209582	SNORA48	17	7574713	5.07E-09
SHBG	ENSG00000129226	CD68	17	7579467	5.74E-32
SHBG	ENSG00000129255	MPDU1	17	7583529	1.73E-29
SHBG	ENSG00000141504	SAT2	17	7626234	2.68E-33
SHBG	ENSG00000129244	ATP1B2	17	7646627	1.43E-12
SHBG	ENSG00000141510	TP53	17	7661779	5.41E-21
SHBG	ENSG00000141499	WRAP53	17	7686372	3.15E-17
SHBG	ENSG00000167874	TMEM88	17	7855065	3.71E-09
SHBG	ENSG00000132518	GUCY2D	17	8002594	4.09E-14
Alkaline phosphatase	ENSG00000171119	NRTN	19	5823802	7.43E-09
Apolipoprotein B	ENSG00000186019	AC084219.4	19	44105463	6.33E-21
Cholesterol	ENSG00000186019	AC084219.4	19	44105463	6.63E-14
LDL direct	ENSG00000186019	AC084219.4	19	44105463	4.88E-18
Apolipoprotein B	ENSG00000130204	TOMM40	19	44890569	8.66E-24
Cholesterol	ENSG00000130204	TOMM40	19	44890569	3.03E-09
LDL direct	ENSG00000130204	TOMM40	19	44890569	1.44E-15
Alkaline phosphatase	ENSG00000142233	NTN5	19	48661407	5.71E-08
Cystatin C	ENSG00000101439	CST3	20	23626706	9.65E-27
Gamma glutamyltransferase	ENSG00000099991	CABIN1	22	24011192	1.01E-07
Gamma glutamyltransferase	ENSG00000099998	GGT5	22	24219654	1.09E-09
Gamma glutamyltransferase	ENSG00000100024	UPB1	22	24494107	1.77E-13
Gamma glutamyltransferase	ENSG00000178026	LRRC75B	22	24585620	3.84E-10
Gamma glutamyltransferase	ENSG00000100031	GGT1	22	24594811	5.82E-16
Gamma glutamyltransferase	ENSG00000284128	BCRP3	22	24644791	3.01E-23
Gamma glutamyltransferase	ENSG00000167037	SGSM1	22	24806169	2.50E-12
Creatinine	ENSG00000196419	XRCC6	22	41621119	6.15E-09

Table 4.3: Associations identified in UKB blood biomarker analysis by Level 2 p-value aggregation of LA-aware casPSs, aspPSs, and standard GReX imputation using PS trained in reference AFR and EUR eQTL summary data. The p-values shown below represent ACAT(casPS p, aspPS AFR p, aspPS EUR p, PS AFR, PS EUR).

Phenotype	Gene	Name	Chr	Pos	p
Apolipoprotein B	ENSG00000134222	PSRC1	1	109279556	7.04E-12
C-reactive protein	ENSG00000158716	DUSP23	1	159780932	4.65E-12
C-reactive protein	ENSG00000272668	RP11-190A12.8	1	159866954	8.80E-16
C-reactive protein	ENSG00000279430	RP11-190A12.9	1	159910094	4.51E-09
C-reactive protein	ENSG00000171786	NHLH1	1	160367067	1.72E-08
Total bilirubin	ENSG00000135930	EIF4E2	2	232550593	6.68E-08
Direct bilirubin	ENSG00000085978	ATG16L1	2	233210051	5.27E-08
Total bilirubin	ENSG00000085978	ATG16L1	2	233210051	5.92E-22
Direct bilirubin	ENSG00000251791	SCARNA6	2	233288676	1.11E-24
Total bilirubin	ENSG00000251791	SCARNA6	2	233288676	7.81E-50
Direct bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	1.32E-20
Total bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	1.83E-43
Direct bilirubin	ENSG00000077044	DGKD	2	233354507	5.55E-77
Total bilirubin	ENSG00000077044	DGKD	2	233354507	2.73E-111
Direct bilirubin	ENSG00000085982	USP40	2	233475520	2.66E-29
Total bilirubin	ENSG00000085982	USP40	2	233475520	7.19E-58
Direct bilirubin	ENSG00000123485	HJURP	2	233833424	4.56E-29
Total bilirubin	ENSG00000123485	HJURP	2	233833424	6.68E-50
Urate	ENSG00000261490	RP11-448G15.3	4	10068089	1.08E-39
Urate	ENSG00000071127	WDR1	4	10074339	3.38E-09
Urate	ENSG00000178163	ZNF518B	4	10439874	5.61E-22
Alkaline phosphatase	ENSG00000112293	GPLD1	6	24428177	1.43E-26
Lipoprotein A	ENSG00000175048	ZDHHC14	6	157381133	9.96E-08
Lipoprotein A	ENSG00000122335	SERAC1	6	158109515	6.38E-09
Lipoprotein A	ENSG00000218226	TATDN2P2	6	158609706	1.46E-08
Lipoprotein A	ENSG00000164691	TAGAP	6	159034468	3.57E-10
Lipoprotein A	ENSG00000220305	HNRNPH1P1	6	159712801	4.42E-10
Lipoprotein A	ENSG00000146457	WTAP	6	159725585	7.05E-08
Lipoprotein A	ENSG00000175003	SLC22A1	6	160121789	1.77E-08
Lipoprotein A	ENSG00000213071	LPAL2	6	160453428	7.49E-09
Lipoprotein A	ENSG00000026652	AGPAT4	6	161129979	1.10E-16
Alkaline phosphatase	ENSG00000135218	CD36	7	80369575	7.73E-16
SHBG	ENSG00000148572	NRBF2	10	63133247	5.89E-10
SHBG	ENSG00000165476	REEP3	10	63521363	1.10E-10
Apolipoprotein A	ENSG00000118137	APOA1	11	116835751	1.75E-08
Phosphate	ENSG00000047621	C12orf4	12	4487728	1.19E-11
SHBG	ENSG00000169992	NLGN2	17	7404874	3.42E-11
SHBG	ENSG00000181284	TMEM102	17	7435443	1.15E-12
SHBG	ENSG00000170175	CHRNB1	17	7445061	3.25E-08
SHBG	ENSG00000239697	TNFSF12	17	7548891	2.83E-18
SHBG	ENSG00000161955	TNFSF13	17	7558292	9.98E-30
SHBG	ENSG00000161960	EIF4A1	17	7572706	5.75E-13
SHBG	ENSG00000209582	SNORA48	17	7574713	8.45E-09
SHBG	ENSG00000238917	SNORD10	17	7576811	4.09E-13
SHBG	ENSG00000129226	CD68	17	7579467	6.67E-32
SHBG	ENSG00000129255	MPDU1	17	7583529	1.56E-29
SHBG	ENSG00000141504	SAT2	17	7626234	4.25E-33
SHBG	ENSG00000129244	ATP1B2	17	7646627	2.38E-12
SHBG	ENSG00000141510	TP53	17	7661779	9.02E-21
SHBG	ENSG00000141499	WRAP53	17	7686372	5.25E-17
SHBG	ENSG00000167874	TMEM88	17	7855065	1.36E-09
SHBG	ENSG00000132518	GUCY2D	17	8002594	6.81E-14
Alkaline phosphatase	ENSG00000171119	NRTN	19	5823802	7.65E-09
Apolipoprotein B	ENSG00000186019	AC084219.4	19	44105463	6.83E-21
Cholesterol	ENSG00000186019	AC084219.4	19	44105463	1.03E-13
LDL direct	ENSG00000186019	AC084219.4	19	44105463	6.86E-18
Apolipoprotein B	ENSG00000130204	TOMM40	19	44890569	1.43E-23
Cholesterol	ENSG00000130204	TOMM40	19	44890569	4.84E-09
LDL direct	ENSG00000130204	TOMM40	19	44890569	2.37E-15
Alkaline phosphatase	ENSG00000142233	NTN5	19	48661407	2.60E-08
Cystatin C	ENSG00000101439	CST3	20	23626706	5.07E-31
Gamma glutamyltransferase	ENSG00000099998	GGT5	22	24219654	5.10E-10
Gamma glutamyltransferase	ENSG00000100024	UPB1	22	24494107	2.95E-13
Gamma glutamyltransferase	ENSG00000178026	LRRC75B	22	24585620	6.40E-10
Gamma glutamyltransferase	ENSG00000100031	GGT1	22	24594811	9.69E-16
Gamma glutamyltransferase	ENSG00000284128	BCRP3	22	24644791	2.79E-24
Gamma glutamyltransferase	ENSG00000167037	SGSM1	22	24806169	1.74E-12
Creatinine	ENSG00000196419	XRCC6	22	41621119	1.03E-08

Table 4.4: Associations identified in UKB blood biomarker analysis by Level 2 p-value aggregation of LA-aware aspPSs. The p-values shown below represent ACAT(aspPS AFR p, aspPS EUR p).

Phenotype	Gene	Name	Chr	Pos	p
Apolipoprotein B	ENSG00000134222	PSRC1	1	109279556	1.90E-11
C-reactive protein	ENSG00000158716	DUSP23	1	159780932	2.04E-10
C-reactive protein	ENSG00000272668	RP11-190A12.8	1	159866954	7.72E-16
C-reactive protein	ENSG00000279430	RP11-190A12.9	1	159910094	3.92E-09
Total bilirubin	ENSG00000135930	EIF4E2	2	232550593	2.67E-08
Direct bilirubin	ENSG00000085978	ATG16L1	2	233210051	2.11E-08
Total bilirubin	ENSG00000085978	ATG16L1	2	233210051	2.37E-22
Direct bilirubin	ENSG00000251791	SCARNA6	2	233288676	8.22E-25
Total bilirubin	ENSG00000251791	SCARNA6	2	233288676	6.53E-50
Direct bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	2.07E-20
Total bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	4.46E-43
Direct bilirubin	ENSG00000077044	DGKD	2	233354507	6.23E-53
Total bilirubin	ENSG00000077044	DGKD	2	233354507	3.98E-77
Direct bilirubin	ENSG00000085982	USP40	2	233475520	5.54E-25
Total bilirubin	ENSG00000085982	USP40	2	233475520	1.75E-44
Direct bilirubin	ENSG00000123485	HJURP	2	233833424	2.66E-28
Total bilirubin	ENSG00000123485	HJURP	2	233833424	8.68E-50
Urate	ENSG00000261490	RP11-448G15.3	4	10068089	2.29E-13
Urate	ENSG00000178163	ZNF518B	4	10439874	3.31E-12
Alkaline phosphatase	ENSG00000112293	GPLD1	6	24428177	7.18E-27
Lipoprotein A	ENSG00000175048	ZDHHC14	6	157381133	8.58E-08
Lipoprotein A	ENSG00000122335	SERAC1	6	158109515	2.60E-09
Lipoprotein A	ENSG00000218226	TATDN2P2	6	158609706	1.18E-08
Lipoprotein A	ENSG00000164691	TAGAP	6	159034468	2.49E-10
Lipoprotein A	ENSG00000220305	HNRNPH1P1	6	159712801	2.16E-10
Lipoprotein A	ENSG00000146457	WTAP	6	159725585	2.83E-08
Lipoprotein A	ENSG00000175003	SLC22A1	6	160121789	7.14E-09
Lipoprotein A	ENSG00000213071	LPAL2	6	160453428	2.99E-09
Lipoprotein A	ENSG00000026652	AGPAT4	6	161129979	4.39E-17
Alkaline phosphatase	ENSG00000135218	CD36	7	80369575	5.80E-16
Apolipoprotein A	ENSG00000135218	CD36	7	80369575	1.10E-07
SHBG	ENSG00000148572	NRBF2	10	63133247	7.03E-10
SHBG	ENSG00000165476	REEP3	10	63521363	4.42E-11
Apolipoprotein A	ENSG00000118137	APOA1	11	116835751	1.71E-08
Phosphate	ENSG00000047621	C12orf4	12	4487728	2.30E-11
Apolipoprotein B	ENSG00000182149	IST1	16	71885233	7.19E-08
SHBG	ENSG00000169992	NLGN2	17	7404874	1.39E-11
SHBG	ENSG00000181284	TMEM102	17	7435443	1.00E-11
SHBG	ENSG00000239697	TNFSF12	17	7548891	1.13E-18
SHBG	ENSG00000161955	TNFSF13	17	7558292	3.99E-30
SHBG	ENSG00000161960	EIF4A1	17	7572706	1.95E-10
SHBG	ENSG00000129226	CD68	17	7579467	3.83E-32
SHBG	ENSG00000129255	MPDU1	17	7583529	1.15E-29
SHBG	ENSG00000141504	SAT2	17	7626234	1.79E-33
SHBG	ENSG00000129244	ATP1B2	17	7646627	1.09E-12
SHBG	ENSG00000141510	TP53	17	7661779	8.61E-21
SHBG	ENSG00000141499	WRAP53	17	7686372	2.10E-17
SHBG	ENSG00000167874	TMEM88	17	7855065	2.91E-09
SHBG	ENSG00000132518	GUCY2D	17	8002594	2.73E-14
Alkaline phosphatase	ENSG00000171119	NRTN	19	5823802	9.98E-09
Apolipoprotein B	ENSG00000186019	AC084219.4	19	44105463	4.22E-21
Cholesterol	ENSG00000186019	AC084219.4	19	44105463	4.42E-14
LDL direct	ENSG00000186019	AC084219.4	19	44105463	3.25E-18
Apolipoprotein B	ENSG00000130204	TOMM40	19	44890569	8.41E-20
LDL direct	ENSG00000130204	TOMM40	19	44890569	5.99E-13
Alkaline phosphatase	ENSG00000142233	NTN5	19	48661407	3.81E-08
Cystatin C	ENSG00000101439	CST3	20	23626706	6.50E-27
Gamma glutamyltransferase	ENSG00000099991	CABIN1	22	24011192	7.65E-08
Gamma glutamyltransferase	ENSG00000099998	GGT5	22	24219654	7.79E-10
Gamma glutamyltransferase	ENSG00000100024	UPB1	22	24494107	1.18E-13
Gamma glutamyltransferase	ENSG00000178026	LRRC75B	22	24585620	2.15E-09
Gamma glutamyltransferase	ENSG00000100031	GGT1	22	24594811	3.88E-16
Gamma glutamyltransferase	ENSG00000284128	BCRP3	22	24644791	4.05E-23
Gamma glutamyltransferase	ENSG00000167037	SGSM1	22	24806169	2.90E-12
Creatinine	ENSG00000196419	XRCC6	22	41621119	4.10E-09

Table 4.5: Associations identified in UKB blood biomarker analysis by Level 1 p-value aggregation of standard GReX imputation using PSs trained in reference AFR eQTL summary data. The p-values shown below represent p-value aggregation of the three PRS models.

Phenotype	Gene	Name	Chr	Pos	p
Apolipoprotein B	ENSG00000134222	PSRC1	1	109279556	1.65E-12
C-reactive protein	ENSG00000158716	DUSP23	1	159780932	9.39E-13
C-reactive protein	ENSG00000272668	RP11-190A12.8	1	159866954	2.83E-09
C-reactive protein	ENSG00000279430	RP11-190A12.9	1	159910094	1.70E-09
Total bilirubin	ENSG00000085978	ATG16L1	2	233210051	4.48E-11
Total bilirubin	ENSG00000251791	SCARNA6	2	233288676	1.07E-07
Direct bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	4.04E-20
Total bilirubin	ENSG00000259793	RP11-400N9.1	2	233351132	1.59E-42
Direct bilirubin	ENSG00000085982	USP40	2	233475520	2.09E-20
Total bilirubin	ENSG00000085982	USP40	2	233475520	3.82E-48
Urate	ENSG00000178163	ZNF518B	4	10439874	1.12E-22
Alkaline phosphatase	ENSG00000112293	GPLD1	6	24428177	1.41E-26
Lipoprotein A	ENSG00000122335	SERAC1	6	158109515	6.96E-08
Lipoprotein A	ENSG00000218226	TATDN2P2	6	158609706	5.91E-08
Lipoprotein A	ENSG00000164691	TAGAP	6	159034468	1.61E-09
Lipoprotein A	ENSG00000026652	AGPAT4	6	161129979	2.40E-13
Alkaline phosphatase	ENSG00000135218	CD36	7	80369575	1.27E-15
SHBG	ENSG00000148572	NRBF2	10	63133247	2.38E-10
Apolipoprotein A	ENSG00000118137	APOA1	11	116835751	1.71E-08
Phosphate	ENSG00000047621	C12orf4	12	4487728	4.32E-12
SHBG	ENSG00000169992	NLGN2	17	7404874	4.29E-10
SHBG	ENSG00000181284	TMEM102	17	7435443	4.07E-12
SHBG	ENSG00000239697	TNFSF12	17	7548891	2.30E-08
SHBG	ENSG00000161955	TNFSF13	17	7558292	6.24E-21
SHBG	ENSG00000161960	EIF4A1	17	7572706	1.15E-13
SHBG	ENSG00000129226	CD68	17	7579467	4.41E-32
SHBG	ENSG00000129255	MPDU1	17	7583529	6.83E-30
SHBG	ENSG00000141504	SAT2	17	7626234	1.79E-32
SHBG	ENSG00000141510	TP53	17	7661779	1.58E-17
SHBG	ENSG00000167874	TMEM88	17	7855065	3.49E-10
SHBG	ENSG00000132518	GUCY2D	17	8002594	2.50E-11
Alkaline phosphatase	ENSG00000171119	NRTN	19	5823802	4.00E-09
Apolipoprotein B	ENSG00000130204	TOMM40	19	44890569	3.68E-22
Cholesterol	ENSG00000130204	TOMM40	19	44890569	2.31E-08
LDL direct	ENSG00000130204	TOMM40	19	44890569	4.37E-14
Cystatin C	ENSG00000101439	CST3	20	23626706	1.01E-31
Gamma glutamyltransferase	ENSG00000099998	GGT5	22	24219654	1.42E-10
Gamma glutamyltransferase	ENSG00000284128	BCRP3	22	24644791	5.90E-25
Gamma glutamyltransferase	ENSG00000167037	SGSM1	22	24806169	5.99E-13

Table 4.6: Associations identified in UKB blood biomarker analysis by Level 1 p-value aggregation of standard GReX imputation using PSs trained in reference EUR eQTL summary data. The p-values shown below represent p-value aggregation of the three PRS models.

Phenotype	Gene	Name	Chr	Pos	p
C-reactive protein	ENSG00000171786	NHLH1	1	160367067	3.48E-09
Direct bilirubin	ENSG00000251791	SCARNA6	2	233288676	4.87E-25
Total bilirubin	ENSG00000251791	SCARNA6	2	233288676	3.00E-50
Direct bilirubin	ENSG00000077044	DGKD	2	233354507	1.11E-77
Total bilirubin	ENSG00000077044	DGKD	2	233354507	5.46E-112
Direct bilirubin	ENSG00000085982	USP40	2	233475520	5.33E-30
Total bilirubin	ENSG00000085982	USP40	2	233475520	1.48E-49
Direct bilirubin	ENSG00000123485	HJURP	2	233833424	9.79E-30
Total bilirubin	ENSG00000123485	HJURP	2	233833424	1.93E-50
Urate	ENSG00000261490	RP11-448G15.3	4	10068089	2.17E-40
Urate	ENSG00000071127	WDR1	4	10074339	6.77E-10
Alkaline phosphatase	ENSG00000112293	GPLD1	6	24428177	3.72E-13
SHBG	ENSG00000165476	REEP3	10	63521363	3.78E-09
Apolipoprotein B	ENSG00000182149	IST1	16	71885233	1.08E-07
SHBG	ENSG00000239697	TNFSF12	17	7548891	1.42E-08
SHBG	ENSG00000238917	SNORD10	17	7576811	8.19E-14
SHBG	ENSG00000141504	SAT2	17	7626234	8.63E-08
Apolipoprotein B	ENSG00000186019	AC084219.4	19	44105463	3.88E-21
Cholesterol	ENSG00000186019	AC084219.4	19	44105463	3.23E-13
LDL direct	ENSG00000186019	AC084219.4	19	44105463	8.81E-18
Alkaline phosphatase	ENSG00000142233	NTN5	19	48661407	7.17E-09
Cystatin C	ENSG00000101439	CST3	20	23626706	1.26E-15
Gamma glutamyltransferase	ENSG00000100024	UPB1	22	24494107	2.06E-10
Gamma glutamyltransferase	ENSG00000100031	GGT1	22	24594811	7.11E-08
Gamma glutamyltransferase	ENSG00000284128	BCRP3	22	24644791	1.43E-12

4.5.2 Figures

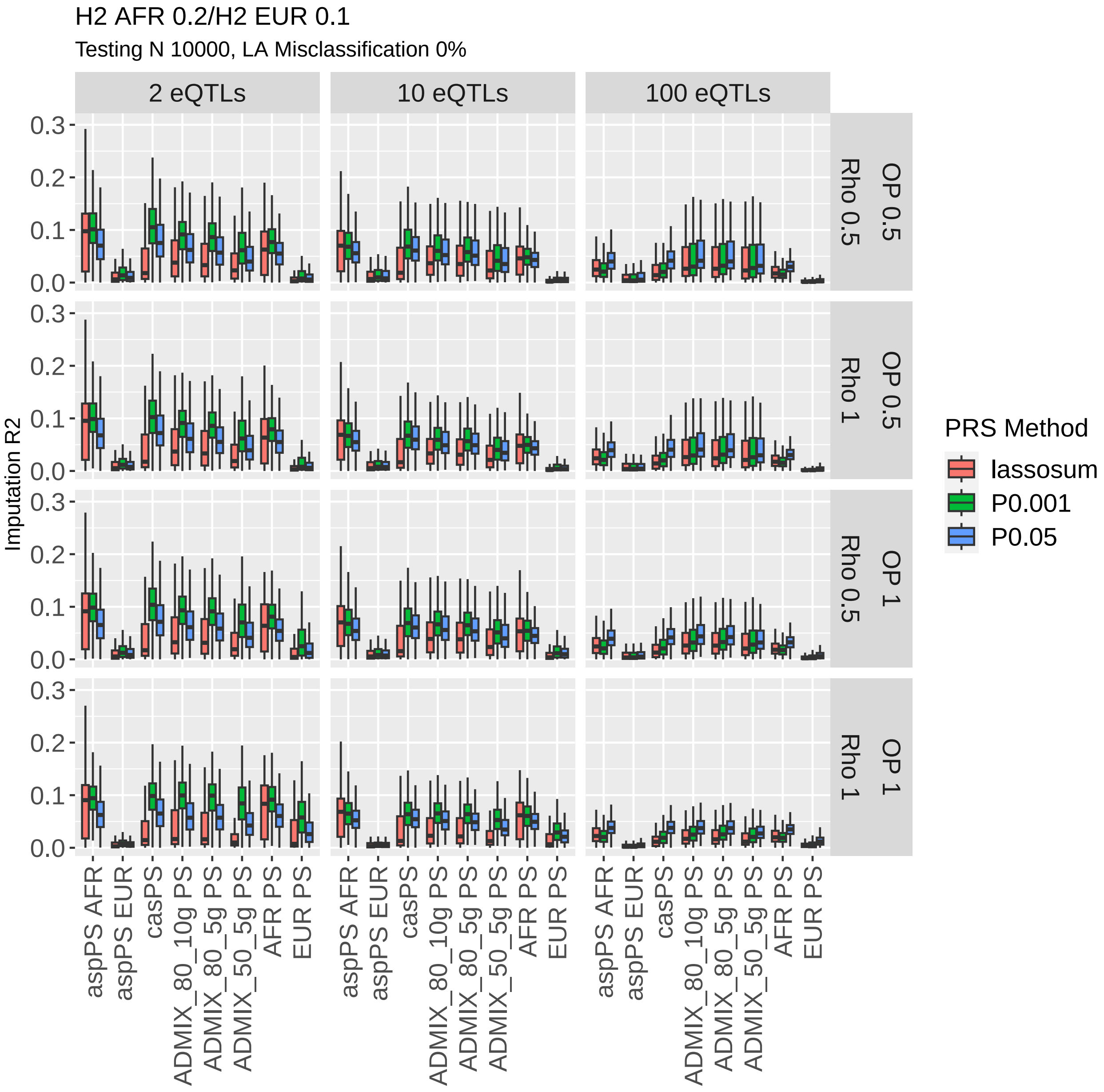

Figure 4.6: Gene expression imputation accuracy in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR) for expression heritability $h^2_{e,1} = 0.2$, $h^2_{e,2} = 0.1$. Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indication the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). The x-axis shows the GReX imputation approach, including our proposed local-ancestry aware methods (aspPS, casPS) and standard PRS imputation approaches (PS). For ancestry-aware methods, we assume no local ancestry misclassification. Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

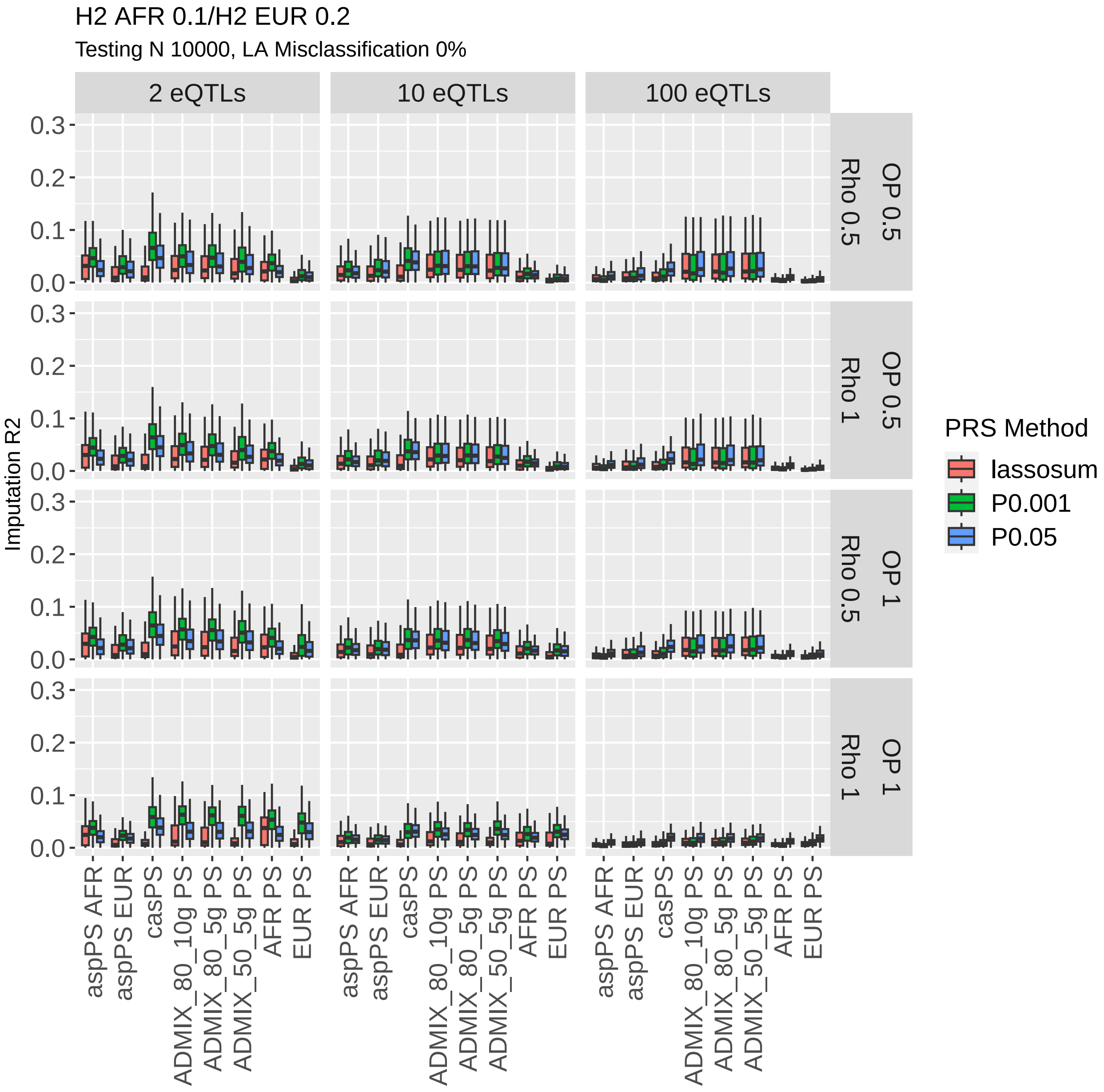

Figure 4.7: Gene expression imputation accuracy in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR) for expression heritability $h^2_{e,1} = 0.1$, $h^2_{e,2} = 0.2$. Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indication the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). The x-axis shows the GReX imputation approach, including our proposed local-ancestry aware methods (aspPS, casPS) and standard PRS imputation approaches (PS). For ancestry-aware methods, we assume no local ancestry misclassification. Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

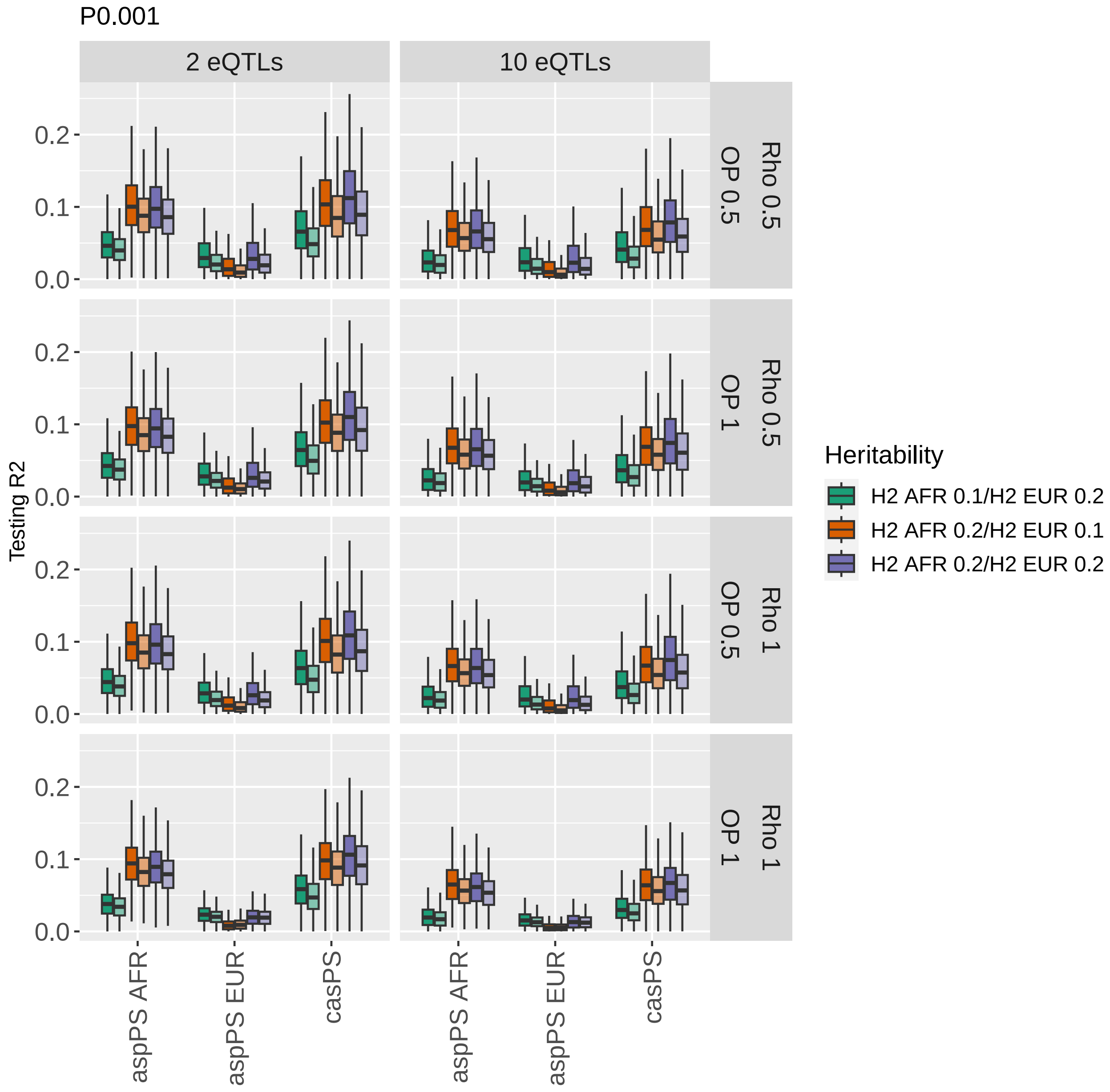

Figure 4.8: Gene expression imputation accuracy of pruning and thresholding ($P_T = 0.001$) for LA-aware approaches in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR). Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indicate the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). For these ancestry-aware methods, we assume either no LA misclassification (darker boxes) or 10% misclassification of SNPs in the gene region (lighter boxes). Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

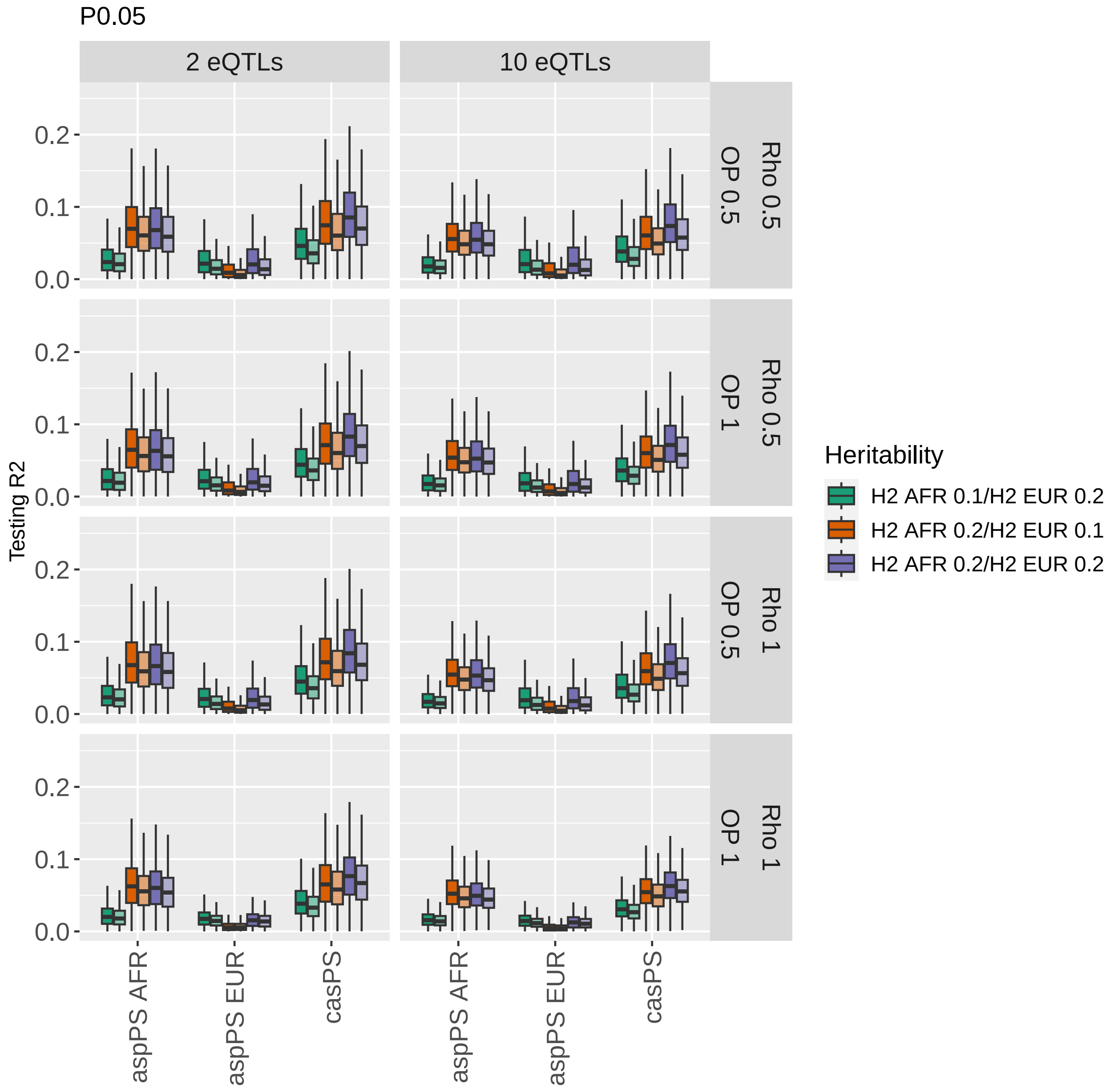

Figure 4.9: Gene expression imputation accuracy of pruning and thresholding ($P_T = 0.05$) for LA-aware approaches in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR). Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indicate the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). For these ancestry-aware methods, we assume either no LA misclassification (darker boxes) or 10% misclassification of SNPs in the gene region (lighter boxes). Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

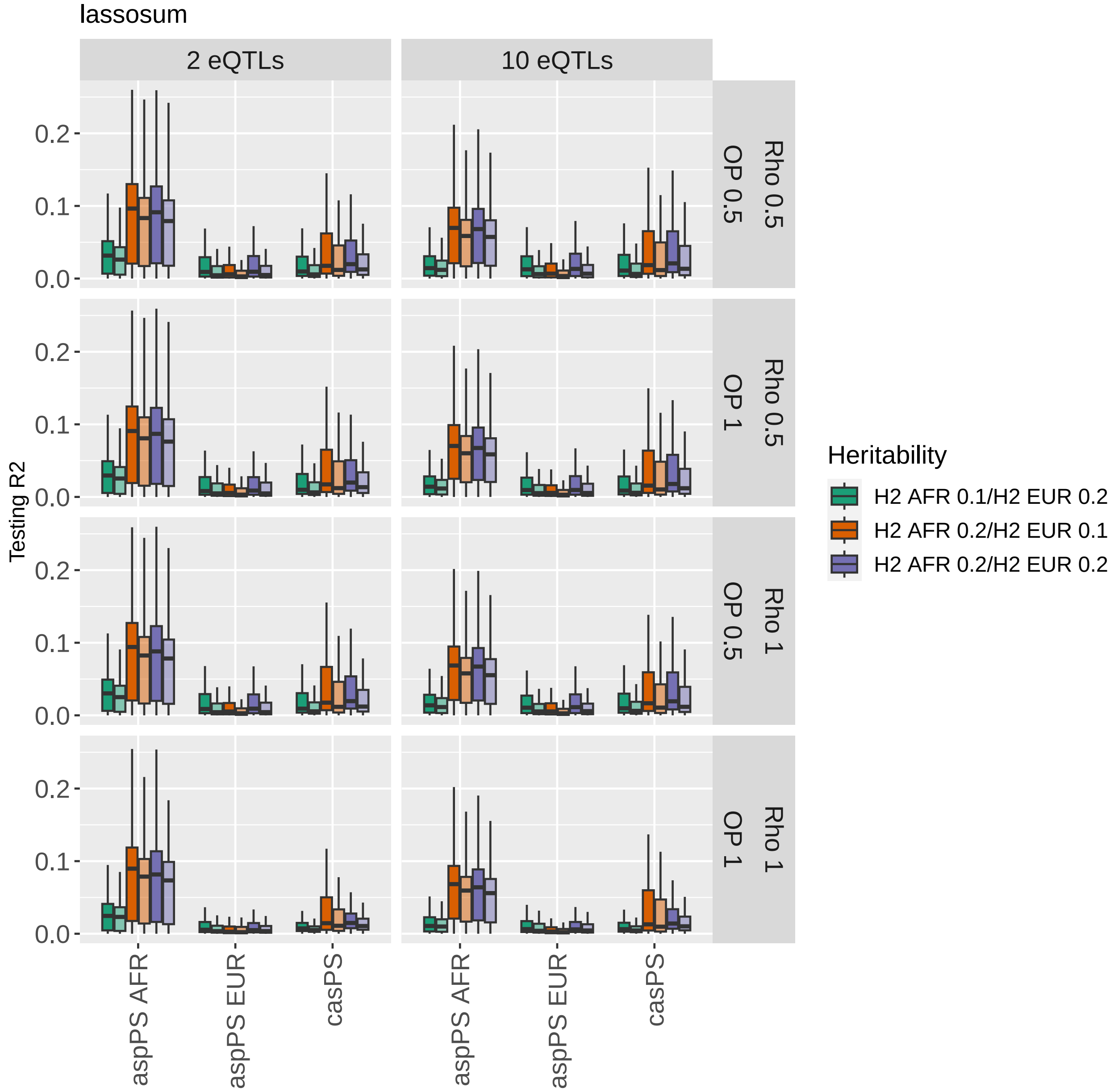

Figure 4.10: Gene expression imputation accuracy of lassosum for LA-aware approaches in 10,000 admixed testing samples (10 admixture generations, 80% initial contribution from AFR). Vertical panels indicate the true number of causal SNPs for gene expression (eQTLs). Horizontal panels indicate the proportion of eQTLs that overlap (OP) between AFR and EUR ancestries, as well as the correlation in eQTL effect sizes for shared eQTLs between the two ancestral groups (ρ). For these ancestry-aware methods, we assume either no LA misclassification (darker boxes) or 10% misclassification of SNPs in the gene region (lighter boxes). Whiskers of boxplot extend to maximum/minimum point that is less than 1.5*IQR from the third/first quartiles.

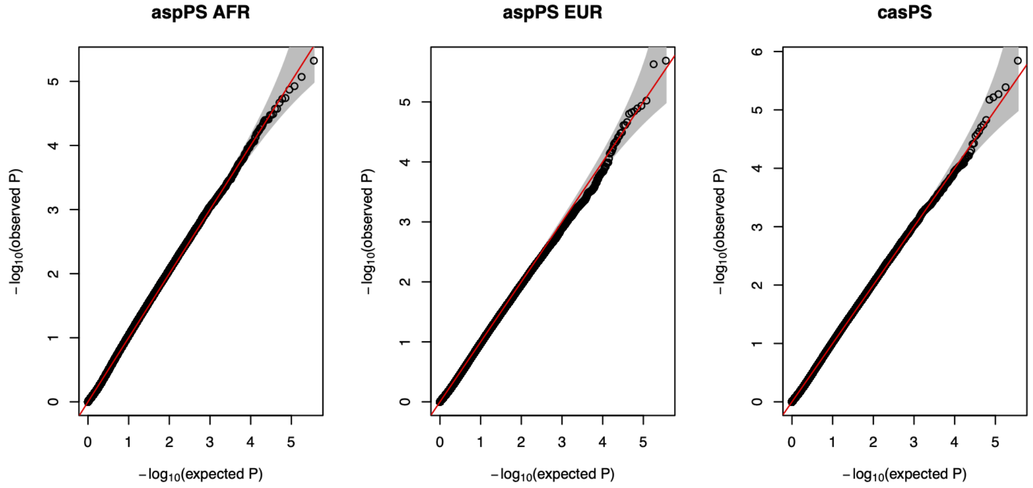

Figure 4.11: QQ plots of p-values from gene-level association tests from LA-aware GReX imputation approaches under the null when no association of expression with trait exists. For these ancestry-aware methods, we assume 10% local ancestry misclassification for SNPs in the gene region. Here, we assume a testing sample size of 10,000. These p-values represent Level 1 p-value aggregation by ACAT, i.e., aggregation of p-values across the three PRS models (P+T0.001, P+T0.05, lassosum). Each plot shown corresponds to 36,000 total simulations, including all 36 gene expression simulation settings.

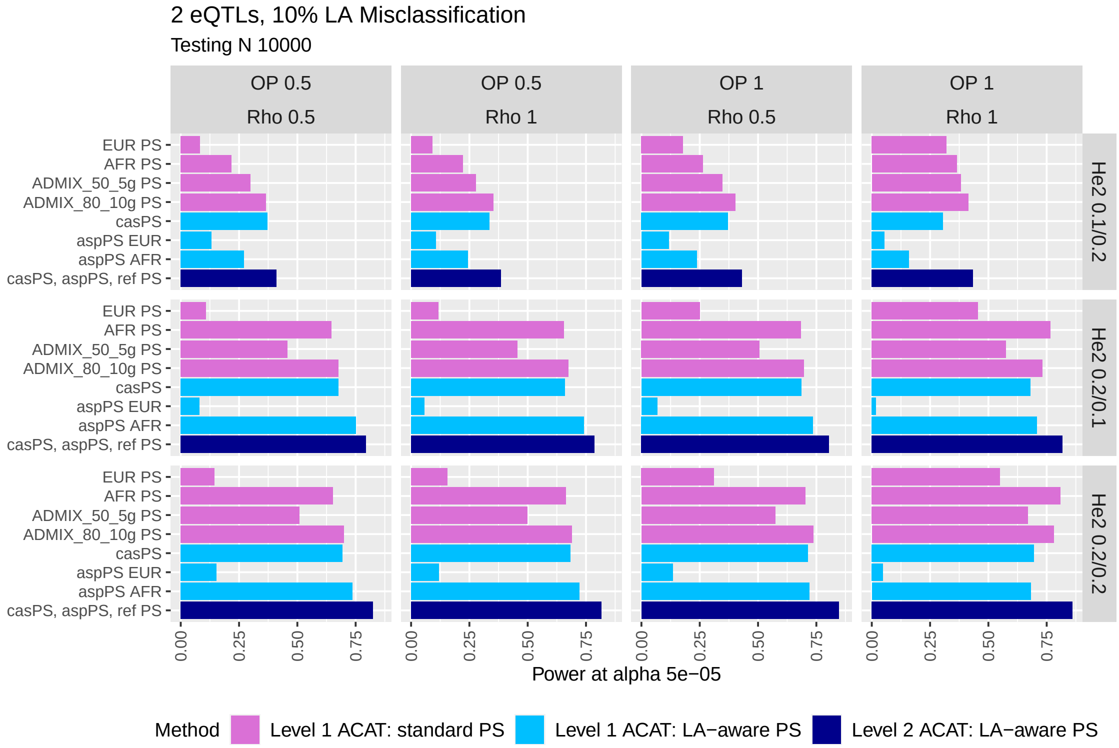

Figure 4.12: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5\times10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 2 eQTLs, 10% local ancestry (LA) misclassification of cis-SNPs for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of LA-unaware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

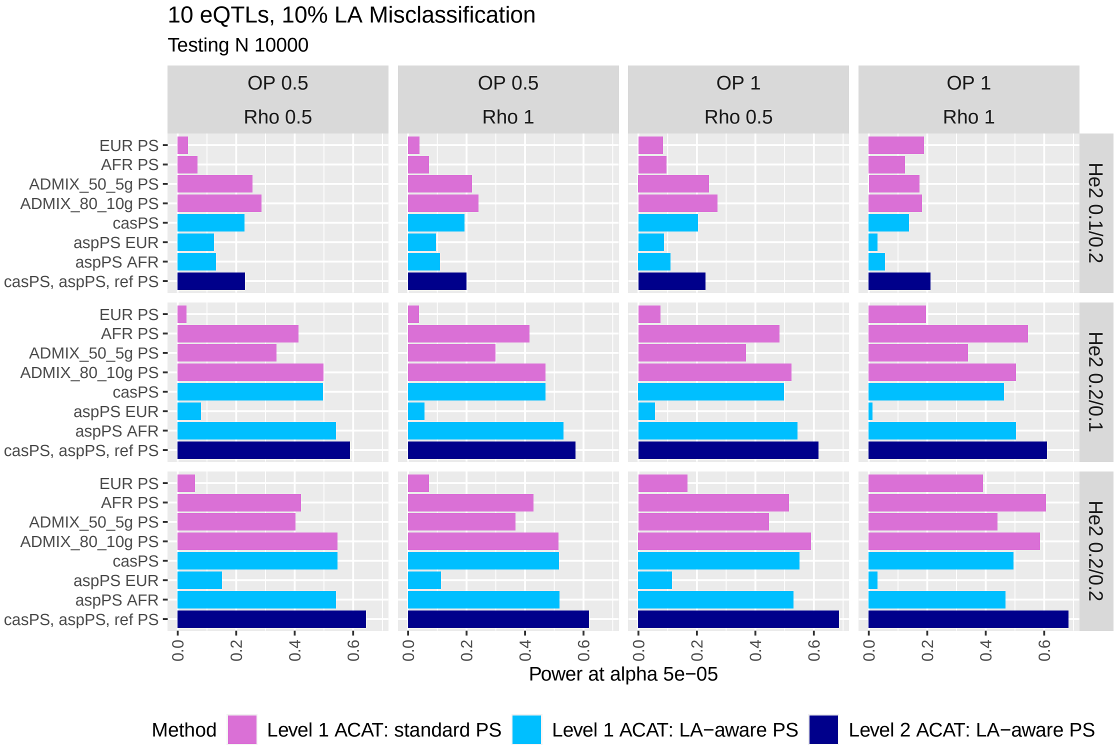

Figure 4.13: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5{\times}10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 10 eQTLs, 10% local ancestry (LA) misclassification of cis-SNPs for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of LA-unaware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

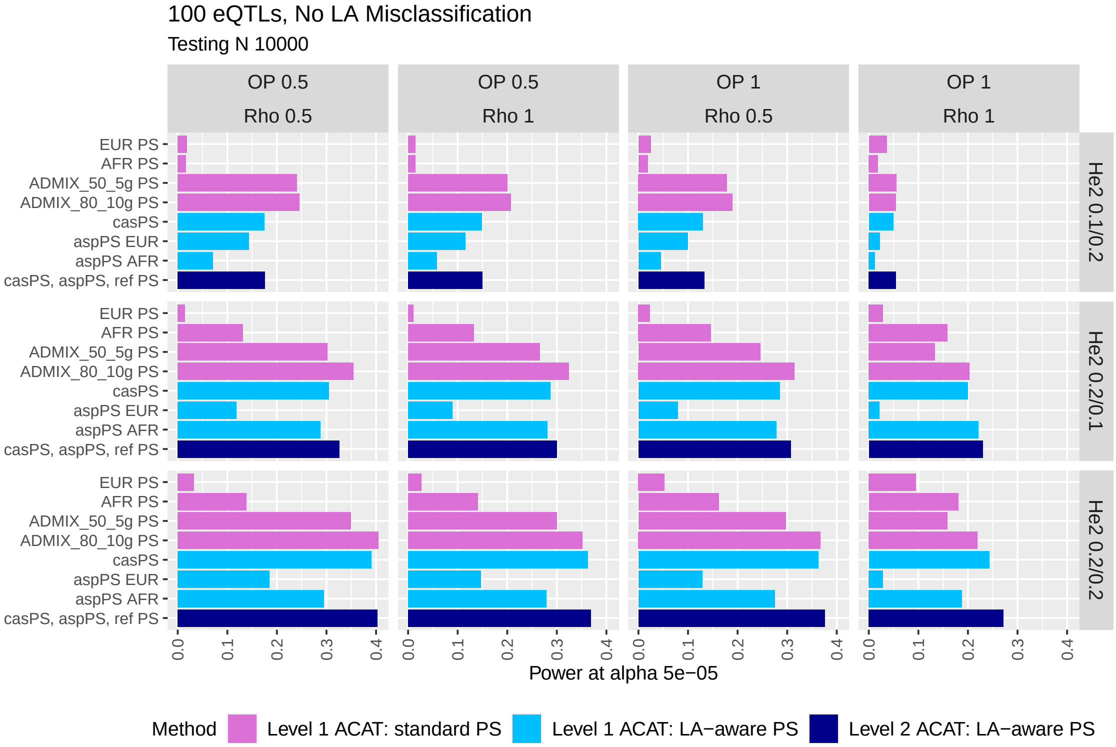

Figure 4.14: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5 \times 10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 100 eQTLs, no local ancestry (LA) misclassification for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of non-LA-aware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

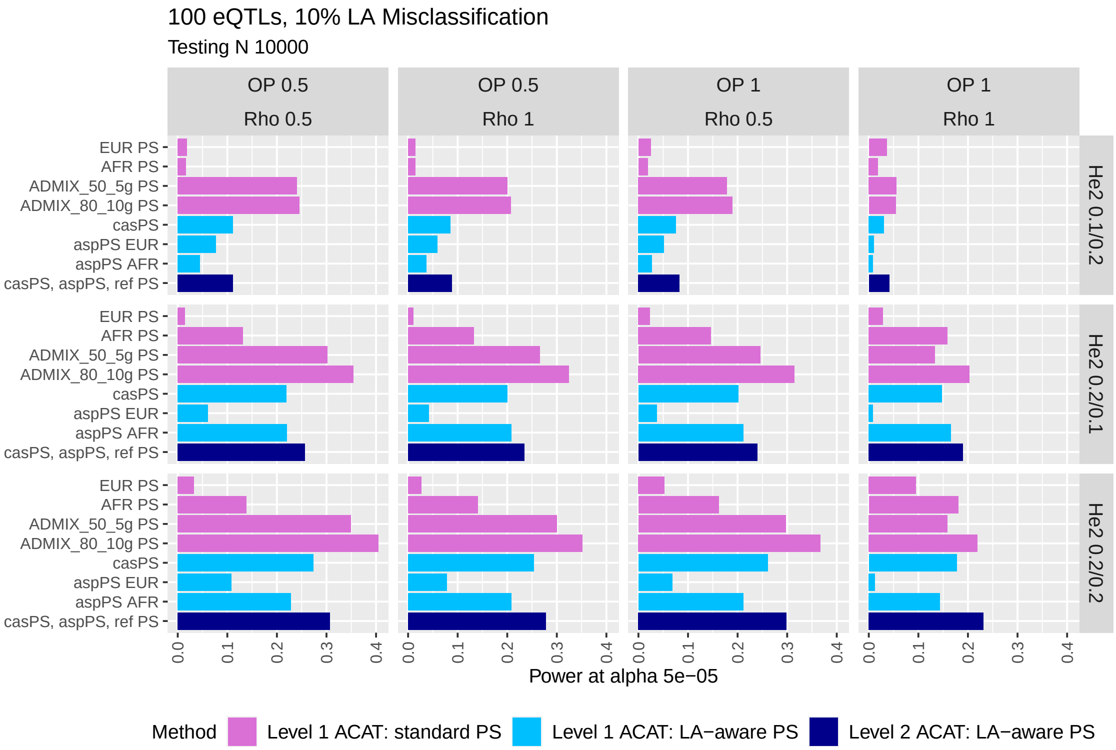

Figure 4.15: Power of gene-level association tests of imputed GReX vectors and simulated trait at significance level $\alpha = 5 \times 10^{-5}$. Here, we assume a phenotypic heritability of $h_p^2 = 0.025$, 100 eQTLs, 10% local ancestry (LA) misclassification of cis-SNPs for LA-aware approaches, and a testing dataset sample size of 10,000. Vertical panels indicate the proportion of eQTLs that are shared between AFR and EUR ancestries (OP) and the correlation of eQTL effect sizes for shared eQTLs (ρ). Horizontal panels indicate the gene expression heritability in AFR and EUR ancestries (h_e^2 AFR/EUR). Pink bars indicate the power of non-LA-aware GReX imputation approaches, with p-values aggregated across the three PRS models (ACAT Level 1). Light blue bars indicate LA-aware approaches with Level 1 p-value aggregation by ACAT. Dark blue bars indicate the power of LA-aware approaches, aggregating both PRS p-values and the resulting p-values of casPS, aspPSs (AFR and EUR), and standard PSs trained in the two AFR/EUR reference populations (ACAT Level 2).

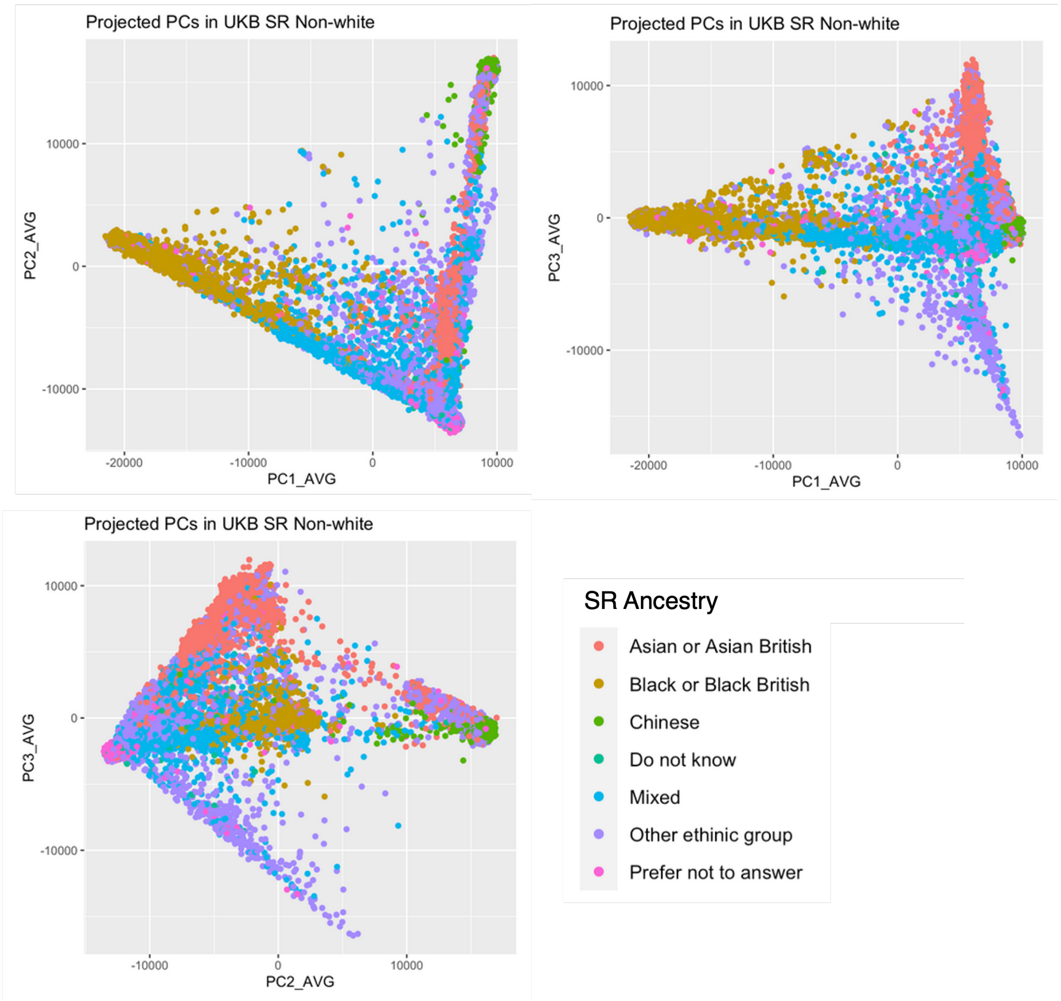

Figure 4.16: Projection of UKB self-reported non-White individuals (N=27,491) onto three-dimensional principal component space calculated using 1000 Genomes samples from the following superpopulations: African, American, East Asian, European and South Asian. Coloring of samples indicates self-reported (SR) ethnicity of UKB subjects.

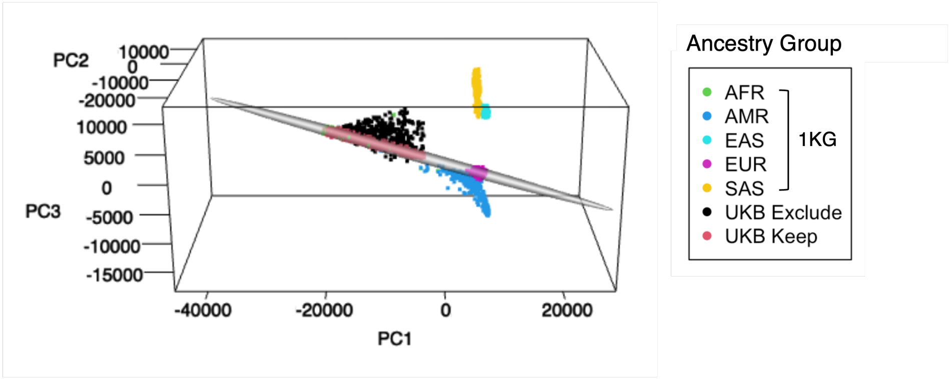

Figure 4.17: Projection onto 1000 Genomes principal component space of N=9,187 UKB self-reported non-white subjects with > 50% probability of AFR ancestry by the random forest ancestry classification model. These axes were calculated using reference samples from the following superpopulations: African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). These subjects are also included on the plot to provide orientation to UKB subjects. UKB subjects (red, black) are colored by whether they fall within the 95% ellipsoid along the AFR-EUR cline.

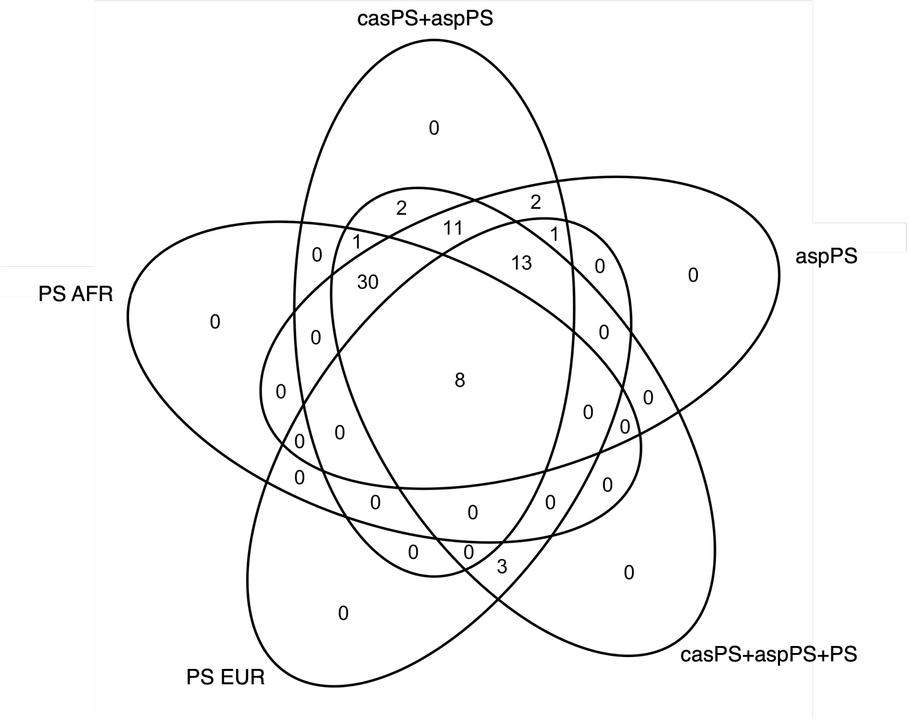

Figure 4.18: Number of significant gene-trait associations found across all 29 blood biomarker traits in UKB analysis. Counts are grouped by GReX imputation and p-value aggregation approach.

4.5.3 Proofs

Expression heritability in admixed subjects

Let there be N admixed individuals. Assume there are V total eQTLs in each ancestry. In general, a subscript 1 corresponds to the African (AFR) reference population, and subscript 2 corresponds to the European (EUR) reference population. Consider the following definitions for admixed subject i, where MAC stands for minor allele count and LA stands for local ancestry:

- $x_{ivm1} \in \{0,1\}$: MAC for vth AFR eQTL on maternal haplotype
- $x_{ivp1} \in \{0,1\}$: MAC for vth AFR eQTL on paternal haplotype
- $x_{ivm2} \in \{0,1\}$: MAC for vth EUR eQTL on maternal haplotype
- $x_{ivp2} \in \{0,1\}$: MAC for vth EUR eQTL on paternal haplotype
- $\gamma_{ivm1} \in \{0,1\}$: 1 if AFR LA of vth AFR eQTL on maternal haplotype
- $\gamma_{ivp1} \in \{0,1\}$: 1 if AFR LA of vth AFR eQTL on paternal haplotype
- $\gamma_{ivm2} \in \{0,1\}$: 1 if EUR LA of vth EUR eQTL on maternal haplotype
- $\gamma_{ivp2} \in \{0,1\}$: 1 if EUR LA of vth EUR eQTL on paternal haplotype

Let g_{iv1} represent the number of AFR-ancestry minor alleles of the vth AFR eQTL, and g_{iv2} be the number of EUR-ancestry minor alleles of subject i at vth EUR eQTL. We can formally define this as:

$g_{iv1} := x_{ivm1}\gamma_{ivm1} + x_{ivp1}\gamma_{ivp1}$

$g_{iv2} := x_{ivm2}\gamma_{ivm2} + x_{ivp2}\gamma_{ivp2}$

We can also arrange these quantitites into the following matrices, where each column has been centered:

$$G_1 = \begin{pmatrix} \underset{\sim}{g}_{11} & \cdots & \underset{\sim}{g}_{V1} \end{pmatrix}_{N\times V} \text{ where } \underset{\sim}{g}_{v1} = \begin{pmatrix} g_{1v1} - \bar{g}_{v1} \\ \vdots \\ g_{Nv1} - \bar{g}_{v1} \end{pmatrix}_{N\times 1} \quad \text{and } \bar{g}_{v1} = \frac{1}{N}\sum_{i=1}^{N} g_{iv1}$$

$$G_2 = \begin{pmatrix} \underset{\sim}{g}_{12} & \cdots & \underset{\sim}{g}_{V2} \end{pmatrix}_{N\times V} \text{ where } \underset{\sim}{g}_{v2} = \begin{pmatrix} g_{1v2} - \bar{g}_{v2} \\ \vdots \\ g_{Nv2} - \bar{g}_{v2} \end{pmatrix}_{N\times 1} \quad \text{and } \bar{g}_{v2} = \frac{1}{N}\sum_{i=1}^{N} g_{iv2}$$

The following quantities will also be needed to derive heritability:

$$f_{v1} = \frac{\sum_{i=1}^{N} g_{iv1}}{\sum_{i=1}^{N}\gamma_{ivm1} + \sum_{i=1}^{N}\gamma_{ivp1}} = \text{AFR-specific MAF at } v\text{th AFR eQTL}$$

$$f_{v2} = \frac{\sum_{i=1}^{N} g_{iv2}}{\sum_{i=1}^{N}\gamma_{ivm2} + \sum_{i=1}^{N}\gamma_{ivp2}} = \text{EUR-specific MAF at } v\text{th EUR eQTL}$$

$$\theta_{v1} = \frac{\sum_{i=1}^{N}\gamma_{ivm1} + \sum_{i=1}^{N}\gamma_{ivp1}}{2N} = \text{proportion alleles at } v\text{th AFR eQTL that are AFR ancestry}$$

$$\theta_{v2} = \frac{\sum_{i=1}^{N}\gamma_{ivm2} + \sum_{i=1}^{N}\gamma_{ivp2}}{2N} = \text{proportion alleles at } v\text{th EUR eQTL that are EUR ancestry}$$

These definitions imply the following:

$$\sum_{i=1}^{N} g_{iv1} = 2N\theta_{v1}f_{v1}$$

$$\sum_{i=1}^{N} g_{iv2} = 2N\theta_{v2}f_{v2}$$

$$\bar{g}_{v1} = 2\theta_{v1}f_{v1}$$

$$\bar{g}_{v2} = 2\theta_{v2}f_{v2}$$

Let us assume that we want to standardize genotypes by local ancestry. We also assume that, of the V eQTLs in each ancestry, the first S are shared between AFR and EUR, and the remaining U are unique to each ancestry ($S + U = V$). We can therefore model the $N \times 1$ phenotype outcome vector (gene expression) $\underset{\sim}{y}$ as:

$$\underset{\sim}{y} = G_1 T_1^{1/2}\dot{\beta}_1 + G_2 T_2^{1/2}\dot{\beta}_2 + \underset{\sim}{\epsilon}$$

Here, $\underset{\sim}{\epsilon} \sim N(0, (1 - h^2_{adm})I_N)$ and $\dot{\beta}_1, \dot{\beta}_2$ represent the $V \times 1$ vectors of ancestry-specific effects per genotype standard deviation. The phenotypic heritability is h^2_{adm}. Let T_1 be a $V \times V$ diagonal matrix with $(T_1)_{vv} = \tau^2_{v1} = \frac{1}{2f_{v1}(1-f_{v1})}$ and T_2 be a $V \times V$ diagonal matrix with $(T_2)_{vv} = \tau^2_{v2} = \frac{1}{2f_{v2}(1-f_{v2})}$. Also, let h^2_1 be the heritability of expression in AFR ancestry and h^2_2 be heritability of expression in EUR ancestry. Let $\rho = \text{Corr}(\dot{\beta}_{1v}, \dot{\beta}_{2v}), v \in \{1, ..., S\}$, i.e., the correlation of ancestry-specific effects for causal eQTLs that are common to both ancestries. We can model these effects as follows:

$$\begin{pmatrix}\dot{\beta}_1\\ \dot{\beta}_2\end{pmatrix}_{2V\times 1} = \begin{pmatrix}\dot{\beta}_{1S}\\ \dot{\beta}_{1U}\\ \dot{\beta}_{2S}\\ \dot{\beta}_{2U}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ \vdots\\ \vdots\\ 0\end{pmatrix}, \begin{pmatrix}\frac{h_1^2}{V}I_S & 0_{S\times U} & \frac{\rho}{V}\sqrt{h_1^2h_2^2}I_S & 0_{S\times U}\\ & \frac{h_1^2}{V}I_U & 0_{U\times S} & 0_{U\times U}\\ & & \frac{h_2^2}{V}I_S & 0_{S\times U}\\ & & & \frac{h_2^2}{V}I_U\end{pmatrix}\right)$$

To derive the heritability in admixed subjects, let's rewrite $G_1T_1^{1/2}\dot{\beta}_1$ as $G_1\beta_1$ and $G_2T_2^{1/2}\dot{\beta}_2$ as $G_2\beta_2$, where:

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{2V\times 1} = \begin{pmatrix}\beta_{1S}\\ \beta_{1U}\\ \beta_{2S}\\ \beta_{2U}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ \vdots\\ \vdots\\ 0\end{pmatrix}, \begin{pmatrix}A & 0_{S\times U} & B & 0_{S\times U}\\ & C & 0_{U\times S} & 0_{U\times U}\\ & & D & 0_{S\times U}\\ & & & E\end{pmatrix}\right)$$

$A_{S\times S} = \mathrm{Cov}(\beta_{1S}, \beta_{1S}) = \frac{h_1^2}{V}diag(\tau_{v1}^2), v = 1, ..., S$

$B_{S\times S} = \mathrm{Cov}(\beta_{1S}, \beta_{2S}) = \frac{\rho}{V}\sqrt{h_1^2h_2^2}diag(\tau_{v1}\tau_{v2}), v = 1, ..., S$

$C_{U\times U} = \mathrm{Cov}(\beta_{1U}, \beta_{1U}) = \frac{h_1^2}{V}diag(\tau_{v1}^2), v = S+1, ..., V$

$D_{S\times S} = \mathrm{Cov}(\beta_{2S}, \beta_{2S}) = \frac{h_2^2}{V}diag(\tau_{v2}^2), v = 1, ..., S$

$E_{U\times U} = \mathrm{Cov}(\beta_{2U}, \beta_{2U}) = \frac{h_2^2}{V}diag(\tau_{v2}^2), v = S+1, ..., V$

Since the variance of the phenotype (expression) is assumed to be 1, we can define $h_{adm}^2 = \mathrm{Var}(G_1\beta_1 + G_2\beta_2)$.

$\mathrm{Var}(G_1\beta_1 + G_2\beta_2) = \frac{1}{N}tr(E[(G_1\beta_1 + G_2\beta_2)(G_1\beta_1 + G_2\beta_2)'])$

$= \frac{1}{N}tr(E[G_1\beta_1\beta_1'G_1' + G_2\beta_2\beta_1'G_1' + G_1\beta_1\beta_2'G_2' + G_2\beta_2\beta_2'G_2'])$

$= \frac{1}{N}tr(E[G_1\beta_1\beta_1'G_1'] + E[G_2\beta_2\beta_1'G_1'] + E[G_1\beta_1\beta_2'G_2'] + E[G_2\beta_2\beta_2'G_2'])$

$= \frac{1}{N}\{tr(E[G_1\beta_1\beta_1'G_1']) + tr(E[G_2\beta_2\beta_1'G_1']) + tr(E[G_1\beta_1\beta_2'G_2']) + tr(E[G_2\beta_2\beta_2'G_2'])\}$

We can rewrite each component as follows:

$tr(E[G_1\beta_1\beta_1'G_1']) = E(tr[G_1\beta_1\beta_1'G_1'])$

$= E(tr[\beta_1\beta_1'G_1'G_1]$

$= tr(E[\beta_1\beta_1'G_1'G_1])$

$= tr[E(\beta_1\beta_1')G_1'G_1]$

$$= tr(\text{Cov}[\beta_1, \beta_1]G_1'G_1)$$

Since $\text{Cov}[\beta_1, \beta_1]$ is a diagonal matrix, the diagonal elements of the matrix product are the product of diagonal elements of each matrix. Now, let us derive some quantities needed to find the diagonal elements of $G'_{.}G_{.}$

$$\sum_{i=1}^{N} g_{iv1}^2 = \sum_{i=1}^{N}(x_{ivm1}\gamma_{ivm1} + x_{ivp1}\gamma_{ivp1})^2$$
$$= \sum_{i=1}^{N}(x_{ivm1}^2\gamma_{ivm1}^2 + 2x_{ivm1}x_{ivp1}\gamma_{ivm1}\gamma_{ivp1} + x_{ivp1}^2\gamma_{ivp1}^2)$$
$$= \sum_{i=1}^{N} x_{ivm1}\gamma_{ivm1} + 2\sum_{i=1}^{N} x_{ivm1}x_{ivp1}\gamma_{ivm1}\gamma_{ivp1} + \sum_{i=1}^{N} x_{ivp1}\gamma_{ivp1}$$
$$= N\bar{g}_{v1} + 2N\theta_{v1}^2 f_{v1}^2$$
$$= 2N\theta_{v1}f_{v1} + 2N\theta_{v1}^2 f_{v1}^2$$
$$= 2N\theta_{v1}f_{v1}(1 + \theta_{v1}f_{v1})$$

Similarly,

$$\sum_{i=1}^{N} g_{iv2}^2 = 2N\theta_{v2}f_{v2}(1 + \theta_{v2}f_{v2})$$

$$\sum_{i=1}^{N} g_{iv1}g_{iv2} = \sum_{i=1}^{N}(x_{ivm1}\gamma_{ivm1} + x_{ivp1}\gamma_{ivp1})(x_{ivm2}\gamma_{ivm2} + x_{ivp2}\gamma_{ivp2})$$
$$= \sum_{i=1}^{N} x_{ivm1}\gamma_{ivm1}x_{ivm2}\gamma_{ivm2} + \sum_{i=1}^{N} x_{ivp1}\gamma_{ivp1}x_{ivm2}\gamma_{ivm2} + \sum_{i=1}^{N} x_{ivm1}\gamma_{ivm1}x_{ivp2}\gamma_{ivp2} +$$
$$\sum_{i=1}^{N} x_{ivp1}\gamma_{ivp1}x_{ivp2}\gamma_{ivp2}$$

If $v \leq S$, $\gamma_{ivm1} = 1 - \gamma_{ivm2}$ and $\theta_{v1} = 1 - \theta_{v2}$:

$$\sum_{i=1}^{N} g_{iv1}g_{iv2} = \sum_{i=1}^{N} x_{ivp1}\gamma_{ivp1}x_{ivm2}\gamma_{ivm2} + \sum_{i=1}^{N} x_{ivm1}\gamma_{ivm1}x_{ivp2}\gamma_{ivp2}$$
$$= 2Nf_{v1}f_{v2}\theta_{v1}\theta_{v2}$$

If $v > S$, the probability that the vth AFR eQTL is of African ancestry and the probability that the vth EUR eQTL is of European ancestry on the same haplotype are not independent, thus:

$$\sum_{i=1}^{N} g_{iv1}g_{iv2} = 2Nf_{v1}f_{v2}\theta_{v1}\theta_{v2} + 2N\theta_{12}f_{v1}f_{v2}$$

Now, to calculate the diagonal elements:

$$(G_1'G_1)_{vv} = \underset{\sim}{g}'_{v1}\underset{\sim}{g}_{v1}$$
$$= \sum_{i=1}^{N}(g_{iv1} - \bar{g}_{v1})^2$$

$$\begin{aligned}
&= \sum_{i=1}^{N} g_{iv1}^2 - N\bar{g}_{v1}^2 \\
&= 2N\theta_{v1}f_{v1}(1+\theta_{v1}f_{v1}) - N(2\theta_{v1}f_{v1})^2 \\
&= 2N\theta_{v1}f_{v1}(1-\theta_{v1}f_{v1})
\end{aligned}$$

If $v \leq S$:

$$(G_1'G_2)_{vv} = \underset{\sim}{g}_{v1}'\underset{\sim}{g}_{v2} = -2Nf_{v1}f_{v2}\theta_{v1}\theta_{v2}$$

If $v > S$:

$$(G_1'G_2)_{vv} = \underset{\sim}{g}_{v1}'\underset{\sim}{g}_{v2} = 2Nf_{v1}f_{v2}(\theta_{v12} - \theta_{v1}\theta_{v2})$$

Now, to sum the diagonal elements:

$$tr(\text{Cov}[\beta_1,\beta_1]G_1'G_1) = \sum_{v=1}^{V} \tfrac{h_1^2}{V}\tau_{v1}^2 2N\theta_{v1}f_{v1}(1-\theta_{v1}f_{v1}) = \tfrac{Nh_1^2}{V}\sum_{v=1}^{V} \tfrac{\theta_{v1}(1-\theta_{v1}f_{v1})}{1-f_{v1}}$$

$$tr(\text{Cov}[\beta_2,\beta_2]G_2'G_2) = \sum_{v=1}^{V} \tfrac{h_2^2}{V}\tau_{v2}^2 2N\theta_{v2}f_{v2}(1-\theta_{v2}f_{v2}) = \tfrac{Nh_2^2}{V}\sum_{v=1}^{V} \tfrac{\theta_{v2}(1-\theta_{v2}f_{v2})}{1-f_{v2}}$$

$$tr(\text{Cov}[\beta_1,\beta_2]G_1'G_2) = \sum_{v=1}^{S} \tfrac{\rho}{V}\sqrt{h_1^2h_2^2}\tau_{v1}\tau_{v2}(-2Nf_{v1}f_{v2}\theta_{v1}\theta_{v2}) = \tfrac{-N\rho\sqrt{h_1^2h_2^2}}{V}\sum_{v=1}^{S} \tfrac{\theta_{v1}\theta_{v2}\sqrt{f_{v1}f_{v2}}}{\sqrt{(1-f_{v1})(1-f_{v2})}}$$

Thus, we have for heritability in admixed subjects:

$$\begin{aligned}
h_{adm}^2 &= \text{Var}(G_1\beta_1 + G_2\beta_2) \\
&= \tfrac{h_1^2}{V}\sum_{v=1}^{V} \tfrac{\theta_{v1}(1-\theta_{v1}f_{v1})}{1-f_{v1}} + \tfrac{h_2^2}{V}\sum_{v=1}^{V} \tfrac{\theta_{v2}(1-\theta_{v2}f_{v2})}{1-f_{v2}} - \tfrac{2\rho\sqrt{h_1^2h_2^2}}{V}\sum_{v=1}^{S} \tfrac{\theta_{v1}\theta_{v2}\sqrt{f_{v1}f_{v2}}}{\sqrt{(1-f_{v1})(1-f_{v2})}}
\end{aligned}$$

Chapter 5

Future Work

There are several avenues we are interested in pursuing to extend the work presented in this dissertation. In our first project (Chapter 2), we presented a powerful genome-wide test for detecting parent-of-origin effects (POEs) in multiple continuous phenotypes. Rather than testing all genome-wide variants initially for POEs, we can alternatively implement a two-stage screening procedure that may mitigate the multiple testing burden. In the first stage, we propose to perform a standard GWAS for marginal (not parent-of-origin dependent) variant associations that considers multiple traits jointly. We restrict consideration to marginal association tests that are orthogonal to POIROT and thus provide complementary information. We can then efficiently test a smaller subset of top SNPs identified from the first stage for POEs. Another limitation we acknowledge is the requirement of continuous phenotypes. We are also interested in the possible extension of our approach to accommodate dichotomous multivariate traits in addition to continuous traits. One potential solution would be to use liability-threshold models [183] that can effectively transform a binary outcome into a continuous-valued posterior mean genetic liability. We could then use these estimated posterior mean genetic liability scores as phenotypes with current the POIROT architecture.

In our second project (Chapter 3), we presented our transcriptome-wide association study (TWAS) of breast and ovarian cancer that leveraged common cancer subtypes and the regulatory effects of both proximal (cis) and distal (trans) genetic variants on gene expression. While we were successfully able to identify over 100 genes associated with one or more cancer phenotypes, we saw limited validation of these genes using available independent GWAS and RNA sequencing datasets. In particular, we are primarily interested in validating the genes that have not been identified by previous GWAS or TWAS of these cancers that are largely driven by strong trans-eQTL effects. Therefore, innovative approaches are needed to validate the high probability trans-eQTLs of such putatively novel risk genes. One possible extension would be analysis of Hi-C or other chromatin conformation capture data to quantify the rate at which our novel target genes and their corresponding regions of high-evidence trans-eQTLs interact physically in the nucleus. The three-dimensional interaction of these regions in both normal breast tissue samples and established breast cancer cell lines would provide valuable insight into the complex genome-wide genetic regulation that we observe. Further, while we trained a second set of genome-wide expression imputation models for the 101 genes identified as significant in our original breast cancer analyses in breast tumor tissue, we would like to extend our work by training transcriptome-wide BGW-TWAS models of all genes in breast and ovarian tumor tissue to better understand how germline control of somatic gene expression changes during oncogenic transition from normal to tumor tissue.

In the third project (Chapter 4), we introduced a method for performing TWAS in admixed subjects that involves deconvolution of gene expression into ancestry-specific components using local ancestry information and summary eQTL data from multiple reference populations. We note that while our proposed method circumvents the need for individual-level Stage 1 training data (genotype and gene expression), we do require individual-level Stage 2 data (genotype and phenotype information) in our admixed testing samples. Several TWAS methods, including OTTERS, have convenient extensions of their approaches that allow for the substitution of summary-level GWAS information in Stage II [21, 22, 74, 76]. We posit an extension of our presently proposed method that would combine summary-level eQTL data in reference populations with summary-level GWAS data from admixed groups to perform TWAS, as well. This extension would likely require the use of advanced techniques to estimate the local ancestry proportions at all loci in the GWAS study, as well as build on established TWAS test statistics for summary GWAS data (e.g., sPrediXcan) [76]. We are also interested in a potential extension of our method that can leverage summary eQTL data from multiple tissue types jointly.

Bibliography

1. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51,** D977–D985. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkac1010 (2024) (Jan. 2023).
2. Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews. Genetics* **14,** 609–617. ISSN: 1471-0056. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3926806/ (2021) (Sept. 2013).
3. Connolly, S. & Heron, E. A. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. eng. *Briefings in Bioinformatics* **16,** 429–448. ISSN: 1477-4054 (May 2015).
4. Weinberg, C. R., Wilcox, A. J. & Lie, R. T. A Log-Linear Approach to Case-Parent–Triad Data: Assessing Effects of Disease Genes That Act Either Directly or through Maternal Effects and That May Be Subject to Parental Imprinting. en. *The American Journal of Human Genetics* **62,** 969–978. ISSN: 0002-9297. https://www.sciencedirect.com/science/article/pii/S0002929707609902 (2022) (Apr. 1998).
5. Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. en. *Genetic Epidemiology* **26,** 167–185. ISSN: 1098-2272. http://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.10307 (2022) (2004).
6. Howey, R. & Cordell, H. J. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics* **13,** 149. ISSN: 1471-2105. https://doi.org/10.1186/1471-2105-13-149 (2022) (June 2012).
7. Ainsworth, H. F., Unwin, J., Jamison, D. L. & Cordell, H. J. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. en. *Genetic Epidemiology* **35.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.20547, 19–45. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20547 (2022) (2011).

8. Sinsheimer, J. S., Palmer, C. G. S. & Woodward, J. A. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. eng. *Genetic Epidemiology* **24,** 1–13. ISSN: 0741-0395 (Jan. 2003).

9. Howey, R. *et al.* Increased Power for Detection of Parent-of-Origin Effects via the Use of Haplotype Estimation. eng. *American Journal of Human Genetics* **97,** 419–434. ISSN: 1537-6605 (Sept. 2015).

10. Becker, T., Baur, M. P. & Knapp, M. Detection of parent-of-origin effects in nuclear families using haplotype analysis. eng. *Human Heredity* **62,** 64–76. ISSN: 0001-5652 (2006).

11. Zhou, J.-Y. *et al.* A powerful parent-of-origin effects test for qualitative traits incorporating control children in nuclear families. en. *Journal of Human Genetics* **57.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Imprinting;Statistical methods Subject_term_id: imprinting;statistical-methods, 500–507. ISSN: 1435-232X. `https://www.nature.com/articles/jhg201258` (2021) (Aug. 2012).

12. Weinberg, C. R. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. eng. *American Journal of Human Genetics* **65,** 229–235. ISSN: 0002-9297 (July 1999).

13. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. en. *Nature* **462.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7275 Primary_atype: Research Publisher: Nature Publishing Group, 868–874. ISSN: 1476-4687. `https://www.nature.com/articles/nature08625` (2021) (Dec. 2009).

14. Hoggart, C. J. *et al.* Novel Approach Identifies SNPs in SLC2A10 and KCNK9 with Evidence for Parent-of-Origin Effect on Body Mass Index. en. *PLOS Genetics* **10.** Publisher: Public Library of Science, e1004508. ISSN: 1553-7404. `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004508` (2021) (July 2014).

15. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337.** Publisher: American Association for the Advancement of Science, 1190–1195. `https://www.science.org/doi/10.1126/science.1222794` (2022) (Sept. 2012).

16. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. English. *The American Journal of Human Genetics* **100.** Publisher: Elsevier, 228–237. ISSN: 0002-9297, 1537-6605. `https://www.cell.com/ajhg/abstract/S0002-9297(16)30532-8` (2022) (Feb. 2017).

17. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. en. *Nature Genetics* **53.** Number: 9 Publisher: Nature Publishing Group, 1300–1310. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-021-00913-z` (2022) (Sept. 2021).

18. Head, S. T., Leslie, E. J., Cutler, D. J. & Epstein, M. P. POIROT: a powerful test for parent-of-origin effects in unrelated samples leveraging multiple phenotypes. *Bioinformatics* **39,** btad199. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btad199 (2024) (Apr. 2023).

19. Head, S. T. *et al. Cis- and trans-eQTL TWAS of breast and ovarian cancer identify more than 100 risk associated genes in the BCAC and OCAC consortia* en. Pages: 2023.11.09.566218 Section: New Results. Nov. 2023. https://www.biorxiv.org/content/10.1101/2023.11.09.566218v1 (2024).

20. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369.** Publisher: American Association for the Advancement of Science, 1318–1330. https://www.science.org/doi/full/10.1126/science.aaz1776 (2022) (Sept. 2020).

21. Luningham, J. M. *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. English. *The American Journal of Human Genetics* **107.** Publisher: Elsevier, 714–726. ISSN: 0002-9297, 1537-6605. https://www.cell.com/ajhg/abstract/S0002-9297(20)30291-3 (2021) (Oct. 2020).

22. Dai, Q. *et al.* OTTERS: a powerful TWAS framework leveraging summary-level reference data. en. *Nature Communications* **14.** Number: 1 Publisher: Nature Publishing Group, 1271. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-36862-w (2023) (Mar. 2023).

23. Guilmatre, A & Sharp, A. Parent of origin effects. en. *Clinical Genetics* **81,** 201–209. ISSN: 1399-0004. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-0004.2011.01790.x (2021) (2012).

24. Rampersaud, E., Mitchell, B. D., Naj, A. C. & Pollin, T. I. Investigating parent of origin effects in studies of Type 2 Diabetes and Obesity. *Current diabetes reviews* **4,** 329–339. ISSN: 1573-3998. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896493/ (2021) (Nov. 2008).

25. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. en. *Nature Reviews Genetics* **2.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group, 21–32. ISSN: 1471-0064. https://www.nature.com/articles/35047554 (2021) (Jan. 2001).

26. Barlow, D. P. Genomic imprinting: a mammalian epigenetic discovery model. eng. *Annual Review of Genetics* **45,** 379–403. ISSN: 1545-2948 (2011).

27. Peters, J. The role of genomic imprinting in biology and disease: an expanding view. en. *Nature Reviews Genetics* **15.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Development;Disease genetics;Imprinting;Metabolism Subject_term_id: development;disease-genetics;imprinting;metabolism, 517–530. ISSN: 1471-0064. https://www.nature.com/articles/nrg3766 (2021) (Aug. 2014).

28. Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C. G. & Polychronakos, C. Parental genomic imprinting of the human IGF2 gene. en. *Nature Genetics* **4.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group, 98–101. ISSN: 1546-1718. https://www.nature.com/articles/ng0593-98 (2022) (May 1993).

29. Temple, I. K. *et al.* An imprinted gene(s) for diabetes? en. *Nature Genetics* **9.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Correspondence Publisher: Nature Publishing Group, 110–112. ISSN: 1546-1718. https://www.nature.com/articles/ng0295-110 (2022) (Feb. 1995).

30. Huxtable, S. J. *et al.* Analysis of parent-offspring trios provides evidence for linkage and association between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class III variable number tandem repeat alleles. eng. *Diabetes* **49,** 126–130. ISSN: 0012-1797 (Jan. 2000).

31. Polychronakos, C. & Kukuvitis, A. Parental genomic imprinting in endocrinopathies. eng. *European Journal of Endocrinology* **147,** 561–569. ISSN: 0804-4643 (Nov. 2002).

32. Dong, C. *et al.* Possible Genomic Imprinting of Three Human Obesity–Related Genetic Loci. *American Journal of Human Genetics* **76,** 427–437. ISSN: 0002-9297. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1196395/ (2021) (Mar. 2005).

33. Wang, K.-S., Liu, X., Zhang, Q., Aragam, N. & Pan, Y. Parent-of-origin effects of FAS and PDLIM1 in attention-deficit/hyperactivity disorder. en. *Journal of Psychiatry and Neuroscience* **37.** Publisher: Journal of Psychiatry and Neuroscience Section: Research Paper, 46–52. ISSN: 1180-4882. https://www.jpn.ca/content/37/1/46 (2022) (Jan. 2012).

34. Palmer, C. G. S. *et al.* HLA-B Maternal-Fetal Genotype Matching Increases Risk of Schizophrenia. en. *The American Journal of Human Genetics* **79,** 710–715. ISSN: 0002-9297. https://www.sciencedirect.com/science/article/pii/S000292970763081X (2022) (Oct. 2006).

35. Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. en. *Human Genetics* **137,** 39–44. ISSN: 1432-1203. https://doi.org/10.1007/s00439-017-1854-z (2022) (Jan. 2018).

36. He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. eng. *Genetics* **173,** 1885–1891. ISSN: 0016-6731 (Aug. 2006).

37. Kocarnik, J. M. & Fullerton, S. M. Returning pleiotropic results from genetic testing to patients and research participants. eng. *JAMA* **311,** 795–796. ISSN: 1538-3598 (Feb. 2014).

38. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. eng. *Nature Reviews. Genetics* **14,** 483–495. ISSN: 1471-0064 (July 2013).

39. O'Reilly, P. F. *et al.* MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. en. *PLOS ONE* **7.** Publisher: Public Library of Science, e34861. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034861 (2022) (May 2012).

40. O'Brien, P. C. Robust Procedures for Testing Equality of Covariance Matrices. *Biometrics* **48.** Publisher: [Wiley, International Biometric Society], 819–827. ISSN: 0006-341X. https://www.jstor.org/stable/2532347 (2021) (1992).

41. Box, G. E. P. A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika* **36.** Publisher: [Oxford University Press, Biometrika Trust], 317–346. ISSN: 0006-3444. https://www.jstor.org/stable/2332671 (2021) (1949).

42. Tiku, M. & Balakrishnan, N. Testing equality of population variances the robust way. *Communications in Statistics - Theory and Methods* **13.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610928408828818, 2143–2159. ISSN: 0361-0926. https://doi.org/10.1080/03610928408828818 (2022) (Jan. 1984).

43. Hotelling, H. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics* **2.** Publisher: Institute of Mathematical Statistics, 360–378. ISSN: 0003-4851. https://www.jstor.org/stable/2957535 (2021) (1931).

44. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the Use of Variance per Genotype as a Tool to Identify Quantitative Trait Interaction Effects: A Report from the Women's Genome Health Study. en. *PLOS Genetics* **6.** Publisher: Public Library of Science, e1000981. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000981 (2023) (June 2010).

45. Vale, C. D. & Maurelli, V. A. Simulating multivariate nonnormal distributions. en. *Psychometrika* **48,** 465–471. ISSN: 1860-0980. https://doi.org/10.1007/BF02293687 (2022) (Sept. 1983).

46. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. en. *Genetic Epidemiology* **32,** 361–369. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20310 (2021) (2008).

47. Broadaway, K. A. *et al.* A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. eng. *American Journal of Human Genetics* **98,** 525–540. ISSN: 1537-6605 (Mar. 2016).

48. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102.** Publisher: Proceedings of the National Academy of Sciences, 15545–15550. https://www.pnas.org/doi/10.1073/pnas.0506580102 (2023) (Oct. 2005).

49. Mootha, V. K. *et al.* PGC-1a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. en. *Nature Genetics* **34.** Number: 3 Publisher: Nature Publishing Group, 267–273. ISSN: 1546-1718. https://www.nature.com/articles/ng1180 (2023) (July 2003).

50. Ray, D., Pankow, J. S. & Basu, S. USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis. en. *Genetic Epidemiology* **40.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.21937, 20–34. ISSN: 1098-2272. http://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21937 (2022) (2016).

51. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. eng. *Nature genetics* **53,** 1260–1269. ISSN: 1546-1718. https://europepmc.org/articles/PMC8349845 (2023) (Aug. 2021).

52. Klimentidis, Y. C. *et al.* Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. eng. *Diabetes* **69,** 2194–2205. ISSN: 1939-327X. https://europepmc.org/articles/PMC7506834 (2023) (Oct. 2020).

53. Cj, W. *et al.* Discovery and refinement of loci associated with lipid levels. English. *Nature Genetics* **45,** 1274–1283. ISSN: 1061-4036, 1546-1718. https://europepmc.org/article/med/24097068 (2023) (Oct. 2013).

54. Luedi, P. P. *et al.* Computational and experimental identification of novel human imprinted genes. eng. *Genome Research* **17,** 1723–1730. ISSN: 1088-9051 (Dec. 2007).

55. Hochner, H. *et al.* Parent-of-Origin Effects of the APOB Gene on Adiposity in Young Adults. *PLoS Genetics* **11,** e1005573. ISSN: 1553-7390. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4599806/ (2023) (Oct. 2015).

56. Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C. & Aulchenko, Y. S. Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genetics* **11,** 92. ISSN: 1471-2156. https://doi.org/10.1186/1471-2156-11-92 (2022) (Oct. 2010).

57. *American Cancer Society* en. http://cancerstatisticscenter.cancer.org/ (2022).

58. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. en. *Nature* **551.** Number: 7678 Publisher: Nature Publishing Group, 92–94. ISSN: 1476-4687. https://www.nature.com/articles/nature24284 (2022) (Nov. 2017).

59. Adedokun, B. *et al.* Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women. en. *Nature Communications* **12.** Number: 1 Publisher: Nature Publishing Group, 4198. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-24327-x (2022) (July 2021).

60. Shu, X. *et al.* Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 1217. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-020-15046-w` (2022) (Mar. 2020).

61. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. en. *Nature Genetics* **52.** Number: 6 Publisher: Nature Publishing Group, 572–581. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-020-0609-2` (2022) (June 2020).

62. Ahearn, T. U. *et al.* Common variants in breast cancer risk loci predispose to distinct tumor subtypes. *Breast Cancer Research* **24,** 2. ISSN: 1465-542X. `https://doi.org/10.1186/s13058-021-01484-x` (2022) (Jan. 2022).

63. Lawrenson, K. *et al.* Genome-wide association studies identify susceptibility loci for epithelial ovarian cancer in east Asian women. en. *Gynecologic Oncology* **153,** 343–355. ISSN: 0090-8258. `https://www.sciencedirect.com/science/article/pii/S0090825819301337` (2022) (May 2019).

64. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. eng. *Nature Genetics* **49,** 680–691. ISSN: 1546-1718 (May 2017).

65. Kar, S. P. *et al.* Common Genetic Variation and Susceptibility to Ovarian Cancer: Current Insights and Future Directions. *Cancer Epidemiology, Biomarkers & Prevention* **27,** 395–404. ISSN: 1055-9965. `https://doi.org/10.1158/1055-9965.EPI-17-0315` (2022) (Apr. 2018).

66. Milne, R. L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. en. *Nature Genetics* **49.** Number: 12 Publisher: Nature Publishing Group, 1767–1778. ISSN: 1546-1718. `https://www.nature.com/articles/ng.3785` (2023) (Dec. 2017).

67. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. eng. *Nature Genetics* **45,** 392–398, 398e1–2. ISSN: 1546-1718 (Apr. 2013).

68. Kar, S. P. *et al.* Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer. *Human Genetics and Genomics Advances* **2,** 100042. ISSN: 2666-2477. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8312632/` (2022) (June 2021).

69. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Frontiers in Genetics* **12.** ISSN: 1664-8021. `https://www.frontiersin.org/article/10.3389/fgene.2021.713230` (2022) (2021).

70. Feng, H. *et al.* Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. en. *Genetic Epidemiology* **44,** 442–468. ISSN: 1098-2272. `http://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22288` (2022) (2020).

71. Gusev, A. *et al.* A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. en. *Nature Genetics* **51.** Number: 5 Publisher: Nature Publishing Group, 815–823. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-019-0395-x` (2022) (May 2019).

72. Ferreira, M. A. *et al.* Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. en. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 1741. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-018-08053-5` (2022) (Apr. 2019).

73. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. en. *Nature Genetics* **47.** Number: 9 Publisher: Nature Publishing Group, 1091–1098. ISSN: 1546-1718. `https://www.nature.com/articles/ng.3367` (2022) (Sept. 2015).

74. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. en. *Nature Genetics* **48.** Number: 3 Publisher: Nature Publishing Group, 245–252. ISSN: 1546-1718. `https://www.nature.com/articles/ng.3506` (2022) (Mar. 2016).

75. Nagpal, S. *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. en. *The American Journal of Human Genetics* **105,** 258–266. ISSN: 0002-9297. `https://www.sciencedirect.com/science/article/pii/S0002929719302058` (2022) (Aug. 2019).

76. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. en. *Nature Communications* **9.** Number: 1 Publisher: Nature Publishing Group, 1825. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-018-03621-1` (2022) (May 2018).

77. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nature genetics* **51,** 675–682. ISSN: 1061-4036. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6619422/` (2022) (Apr. 2019).

78. Parrish, R. L., Gibson, G. C., Epstein, M. P. & Yang, J. TIGAR-V2: Efficient TWAS tool with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. en. *Human Genetics and Genomics Advances* **3,** 100068. ISSN: 2666-2477. `https://www.sciencedirect.com/science/article/pii/S266624772100049X` (2022) (Jan. 2022).

79. Yuan, Z. *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 3861. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-020-17668-6` (2022) (July 2020).

80. Miki, Y. *et al.* A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science* **266.** Publisher: American Association for the Advancement of Science, 66–71. `https://www.science.org/doi/10.1126/science.7545954` (2022) (Oct. 1994).

81. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. eng. *Nature* **378,** 789–792. ISSN: 0028-0836 (Dec. 1995).

82. Jiang, X. *et al.* Shared heritability and functional enrichment across six solid cancers. en. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 431. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-018-08054-4` (2022) (Jan. 2019).

83. Lawrenson, K. *et al.* Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. en. *Nature Communications* **7.** Number: 1 Publisher: Nature Publishing Group, 12675. ISSN: 2041-1723. `https://www.nature.com/articles/ncomms12675` (2022) (Sept. 2016).

84. Kar, S. P. *et al.* Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. eng. *Cancer Discovery* **6,** 1052–1067. ISSN: 2159-8290 (Sept. 2016).

85. Carroll, J. S. Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer. *European Journal of Endocrinology* **175,** R41–R49. ISSN: 0804-4643. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065078/` (2023) (July 2016).

86. Chatterjee, N. A Two-Stage Regression Model for Epidemiological Studies With Multivariate Disease Classification Data. *Journal of the American Statistical Association* **99,** 127–138. ISSN: 0162-1459. `https://doi.org/10.1198/016214504000000124` (2023) (Mar. 2004).

87. Zhang, H. *et al.* A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *Biostatistics (Oxford, England)* **22,** 772–788. ISSN: 1465-4644. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8511944/` (2022) (Feb. 2020).

88. Mavaddat, N. *et al.* Pathology of Breast and Ovarian Cancers among BRCA1 and BRCA2 Mutation Carriers: Results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiology, Biomarkers & Prevention* **21,** 134–147. ISSN: 1055-9965. `https://doi.org/10.1158/1055-9965.EPI-11-0775` (2022) (Jan. 2012).

89. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. eng. *Nature Genetics* **45,** 353–361, 361e1–2. ISSN: 1546-1718 (Apr. 2013).

90. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. eng. *Nature Genetics* **47,** 373–380. ISSN: 1546-1718 (Apr. 2015).

91. Auton, A. *et al.* A global reference for human genetic variation. en. *Nature* **526.** Number: 7571 Publisher: Nature Publishing Group, 68–74. ISSN: 1476-4687. https://www.nature.com/articles/nature15393 (2022) (Oct. 2015).

92. Coetzee, S. *et al.* Integrative multi-omics analyses to identify the genetic and functional mechanisms underlying ovarian cancer risk regions.

93. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. eng. *Bioinformatics (Oxford, England)* **32,** 283–285. ISSN: 1367-4811 (Jan. 2016).

94. Barbeira, A. N. *et al.* Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. en. *Genetic Epidemiology* **44.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.22346, 854–867. ISSN: 1098-2272. http://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22346 (2022) (2020).

95. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 4423. ISSN: 2041-1723. http://www.nature.com/articles/s41467-020-18246-6 (2022) (Sept. 2020).

96. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. English. *Cancer Cell* **37.** Publisher: Elsevier, 639–654.e6. ISSN: 1535-6108, 1878-3686. https://www.cell.com/cancer-cell/abstract/S1535-6108(20)30211-7 (2022) (May 2020).

97. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. eng. *Bioinformatics (Oxford, England)* **26,** 2867–2873. ISSN: 1367-4811 (Nov. 2010).

98. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** R25. ISSN: 1474-760X. https://doi.org/10.1186/gb-2010-11-3-r25 (2022) (Mar. 2010).

99. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. en. *Nature Protocols* **7.** Number: 3 Publisher: Nature Publishing Group, 500–507. ISSN: 1750-2799. https://www.nature.com/articles/nprot.2011.457 (2022) (Mar. 2012).

100. Wang, L., Babushkin, N., Liu, Z. & Liu, X. *Trans-eQTL mapping in gene sets identifies network effects of genetic variants* en. Pages: 2022.11.11.516189 Section: New Results. Nov. 2022. https://www.biorxiv.org/content/10.1101/2022.11.11.516189v1 (2024).

101. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. en. *PLOS Genetics* **10.** Publisher: Public Library of Science, e1004383. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004383 (2024) (May 2014).

102. Wu, L. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. en. *Nature Genetics* **50.** Number: 7 Publisher: Nature Publishing Group, 968–978. ISSN: 1546-1718. https://www.nature.com/articles/s41588-018-0132-x (2022) (July 2018).

103. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. en. *Nature Genetics* **36.** Number: 4 Publisher: Nature Publishing Group, 388–393. ISSN: 1546-1718. https://www.nature.com/articles/ng1333 (2022) (Apr. 2004).

104. Devlin, B. & Roeder, K. Genomic control for association studies. eng. *Biometrics* **55,** 997–1004. ISSN: 0006-341X (Dec. 1999).

105. Li, X. *et al.* Association of multiple genetic variants with breast cancer susceptibility in the Han Chinese population. eng. *Oncotarget* **7,** 85483–85491. ISSN: 1949-2553 (Dec. 2016).

106. Palmer, J. R. *et al.* Genetic susceptibility loci for subtypes of breast cancer in an African American population. eng. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* **22,** 127–134. ISSN: 1538-7755 (Jan. 2013).

107. Bose, M. *et al.* A catalog of curated breast cancer genes. *Breast Cancer Research and Treatment* **191,** 431–441. ISSN: 0167-6806. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8763822/ (2023) (2022).

108. Hu, X., Liu, Y., Du, Y., Cheng, T. & Xia, W. Long non-coding RNA BLACAT1 promotes breast cancer cell proliferation and metastasis by miR-150-5p/CCR2. *Cell & Bioscience* **9,** 14. ISSN: 2045-3701. https://doi.org/10.1186/s13578-019-0274-2 (2022) (Jan. 2019).

109. Hoffman, J. D. *et al.* Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. en. *PLOS Genetics* **13.** Publisher: Public Library of Science, e1006690. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006690 (2022) (Mar. 2017).

110. Couch, F. J. *et al.* Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. eng. *Nature Communications* **7,** 11375. ISSN: 2041-1723 (Apr. 2016).

111. Stevens, K. N. *et al.* Common Breast Cancer Susceptibility Loci Are Associated with Triple-Negative Breast Cancer. *Cancer Research* **71,** 6240–6249. ISSN: 0008-5472. `https://doi.org/10.1158/0008-5472.CAN-11-1266` (2023) (Sept. 2011).

112. Kuchenbaecker, K. B. *et al.* Identification of six new susceptibility loci for invasive epithelial ovarian cancer. en. *Nature Genetics* **47.** Number: 2 Publisher: Nature Publishing Group, 164–171. ISSN: 1546-1718. `https://www.nature.com/articles/ng.3185` (2023) (Feb. 2015).

113. Pharoah, P. D. P. *et al.* GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. eng. *Nature Genetics* **45,** 362–370, 370e1–2. ISSN: 1546-1718 (Apr. 2013).

114. Cai, Q. *et al.* Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. eng. *Nature Genetics* **46,** 886–890. ISSN: 1546-1718 (Aug. 2014).

115. Zhao, Z. *et al.* Association of genetic susceptibility variants for type 2 diabetes with breast cancer risk in women of European ancestry. eng. *Cancer causes & control: CCC* **27,** 679–693. ISSN: 1573-7225 (May 2016).

116. H, S. *et al.* A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. English. *Nature Genetics* **41,** 996–1000. ISSN: 1061-4036, 1546-1718. `https://europepmc.org/article/MED/19648919` (2023) (Aug. 2009).

117. Cesaratto, L. *et al.* BNC2 is a putative tumor suppressor gene in high-grade serous ovarian carcinoma and impacts cell survival after oxidative stress. *Cell Death & Disease* **7,** e2374. ISSN: 2041-4889. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059877/` (2023) (Sept. 2016).

118. Buckley, M. A. *et al.* Functional Analysis and Fine Mapping of the 9p22.2 Ovarian Cancer Susceptibility Locus. *Cancer Research* **79,** 467–481. ISSN: 0008-5472. `https://doi.org/10.1158/0008-5472.CAN-17-3864` (2023) (Feb. 2019).

119. Shan, N., Wang, Z. & Hou, L. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics* **20,** 126. ISSN: 1471-2105. `https://doi.org/10.1186/s12859-019-2651-6` (2023) (Mar. 2019).

120. Dutta, D. *et al.* Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. en. *Nature Communications* **13.** Number: 1 Publisher: Nature Publishing Group, 4323. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-022-31845-9` (2023) (July 2022).

121. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. English. *The American Journal of Human Genetics* **100.** Publisher: Elsevier, 473–487. ISSN: 0002-9297, 1537-6605. https://www.cell.com/ajhg/abstract/S0002-9297(17)30032-0 (2024) (Mar. 2017).

122. Adhikari, H. *et al.* Oncogenic KRAS is dependent upon an EFR3A-PI4KA signaling axis for potent tumorigenic activity. en. *Nature Communications* **12.** Number: 1 Publisher: Nature Publishing Group, 5248. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-25523-5 (2023) (Sept. 2021).

123. Obtulowicz, T. *et al.* Oxidative stress and 8-oxoguanine repair are enhanced in colon adenoma and carcinoma patients. eng. *Mutagenesis* **25,** 463–471. ISSN: 1464-3804 (Sept. 2010).

124. Coskun, E. *et al.* Addiction to MTH1 protein results in intense expression in human breast cancer tissue as measured by liquid chromatography-isotope-dilution tandem mass spectrometry. eng. *DNA repair* **33,** 101–110. ISSN: 1568-7856 (Sept. 2015).

125. SUN, P. *et al.* Introduction to DOK2 and its Potential Role in Cancer. *Physiological Research* **70,** 671–685. ISSN: 0862-8408. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8820521/ (2023) (Sept. 2021).

126. Zhao, N. *et al.* CCDC106 promotes the proliferation and invasion of ovarian cancer cells by suppressing p21 transcription through a p53-independent pathway. eng. *Bioengineered* **13,** 10956–10972. ISSN: 2165-5987 (Apr. 2022).

127. Hu, X., Yang, L. & Mo, Y.-Y. Role of Pseudogenes in Tumorigenesis. *Cancers* **10,** 256. ISSN: 2072-6694. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6115995/ (2023) (Aug. 2018).

128. Martini, R. *et al.* African Ancestry-Associated Gene Expression Profiles in Triple-Negative Breast Cancer Underlie Altered Tumor Biology and Clinical Outcome in Women of African Descent. eng. *Cancer Discovery* **12,** 2530–2551. ISSN: 2159-8290 (Nov. 2022).

129. Aran, D. *et al.* Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications* **8,** 1077. ISSN: 2041-1723. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5651823/ (2023) (Oct. 2017).

130. Ongen, H. *et al.* Putative cis-regulatory drivers in colorectal cancer. en. *Nature* **512.** Number: 7512 Publisher: Nature Publishing Group, 87–90. ISSN: 1476-4687. https://www.nature.com/articles/nature13602 (2023) (Aug. 2014).

131. Giral, H., Landmesser, U. & Kratzer, A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Frontiers in Cardiovascular Medicine* **5,** 181. ISSN: 2297-055X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6304420/ (2021) (Dec. 2018).

132. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. English. *The American Journal of Human Genetics* **104.** Publisher: Elsevier, 410–421. ISSN: 0002-9297, 1537-6605. `https://www.cell.com/ajhg/abstract/S0002-9297(19)30002-3` (2024) (Mar. 2019).

133. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. en. *Nature Genetics* **55.** Number: 6 Publisher: Nature Publishing Group, 952–963. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-023-01377-z` (2023) (June 2023).

134. Keys, K. L. *et al.* On the cross-population generalizability of gene expression prediction models. en. *PLOS Genetics* **16.** Publisher: Public Library of Science, e1008927. ISSN: 1553-7404. `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008927` (2024) (Aug. 2020).

135. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. en. *Nature Genetics* **51.** Number: 4 Publisher: Nature Publishing Group, 584–591. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-019-0379-x` (2024) (Apr. 2019).

136. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine* **375.** Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsa1507092, 655–665. ISSN: 0028-4793. `https://doi.org/10.1056/NEJMsa1507092` (2024) (Aug. 2016).

137. Reisberg, S., Iljasenko, T., Läll, K., Fischer, K. & Vilo, J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. en. *PLOS ONE* **12.** Publisher: Public Library of Science, e0179238. ISSN: 1932-6203. `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179238` (2024) (July 2017).

138. De La Vega, F. M. & Bustamante, C. D. Polygenic risk scores: a biased prediction? *Genome Medicine* **10,** 100. ISSN: 1756-994X. `https://doi.org/10.1186/s13073-018-0610-x` (2024) (Dec. 2018).

139. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biology* **19,** 179. ISSN: 1474-760X. `https://doi.org/10.1186/s13059-018-1561-7` (2024) (Nov. 2018).

140. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. English. *The American Journal of Human Genetics* **101.** Publisher: Elsevier, 5–22. ISSN: 0002-9297, 1537-6605. `https://www.cell.com/ajhg/abstract/S0002-9297(17)30240-9` (2024) (July 2017).

141. Scutari, M., Mackay, I. & Balding, D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. en. *PLOS Genetics* **12.** Publisher: Public Library of Science, e1006288. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006288 (2024) (Sept. 2016).

142. Skotte, L., Jørsboe, E., Korneliussen, T. S., Moltke, I. & Albrechtsen, A. Ancestry-specific association mapping in admixed populations. en. *Genetic Epidemiology* **43,** 506–521. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22200 (2024) (2019).

143. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. English. *Cell* **177.** Publisher: Elsevier, 26–31. ISSN: 0092-8674, 1097-4172. https://www.cell.com/cell/abstract/S0092-8674(19)30231-4 (2024) (Mar. 2019).

144. Bureau, U. C. *2020 Census Illuminates Racial and Ethnic Composition of the Country* tech. rep. Section: Government (Aug. 2021). https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html (2024).

145. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. en. *Nature Genetics* **53.** Number: 2 Publisher: Nature Publishing Group, 195–204. ISSN: 1546-1718. https://www.nature.com/articles/s41588-020-00766-y (2023) (Feb. 2021).

146. Zhang, J. & Stram, D. O. The Role of Local Ancestry Adjustment in Association Studies Using Admixed Populations. en. *Genetic Epidemiology* **38.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.21835, 502–515. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21835 (2024) (2014).

147. Lachance, J. *et al.* Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. English. *Cell* **150.** Publisher: Elsevier, 457–469. ISSN: 0092-8674, 1097-4172. https://www.cell.com/cell/abstract/S0092-8674(12)00831-8 (2024) (Aug. 2012).

148. Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I. & London, S. J. Joint testing of genotype and ancestry association in admixed families. en. *Genetic Epidemiology* **34,** 783–791. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20520 (2024) (2010).

149. Coram, M. A. *et al.* Genome-wide Characterization of Shared and Distinct Genetic Components that Influence Blood Lipid Levels in Ethnically Diverse Human Populations. English. *The American Journal of Human Genetics* **92.** Publisher: Elsevier, 904–916. ISSN: 0002-9297, 1537-6605. https://www.cell.com/ajhg/abstract/S0002-9297(13)00212-7 (2024) (June 2013).

150. Aschard, H., Gusev, A., Brown, R. & Pasaniuc, B. Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genetics* **16,** 124. ISSN: 1471-2156. `https://doi.org/10.1186/s12863-015-0283-z` (2024) (Oct. 2015).

151. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. eng. *American Journal of Human Genetics* **86,** 23–33. ISSN: 1537-6605 (Jan. 2010).

152. Pasaniuc, B. *et al.* Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. en. *PLOS Genetics* **7.** Publisher: Public Library of Science, e1001371. ISSN: 1553-7404. `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001371` (2024) (Apr. 2011).

153. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* **29,** 1407–1415. ISSN: 1367-4803. `https://doi.org/10.1093/bioinformatics/btt166` (2024) (June 2013).

154. Chimusa, E. R. *et al.* Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. eng. *Human Molecular Genetics* **23,** 796–809. ISSN: 1460-2083 (Feb. 2014).

155. Smith, E. N. *et al.* Genome-wide association study of bipolar disorder in European American and African American individuals. eng. *Molecular Psychiatry* **14,** 755–763. ISSN: 1476-5578 (Aug. 2009).

156. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 1628. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-020-15464-w` (2023) (Apr. 2020).

157. Hoggart, C. J. *et al.* BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. en. *Nature Genetics* **56.** Number: 1 Publisher: Nature Publishing Group, 180–186. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-023-01583-9` (2024) (Jan. 2024).

158. Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. en. *Nature Communications* **15.** Number: 1 Publisher: Nature Publishing Group, 1016. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-024-45135-z` (2024) (Feb. 2024).

159. Shang, L. *et al.* Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. *American Journal of Human Genetics* **106,** 496–512. ISSN: 0002-9297. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7118581/` (2024) (Apr. 2020).

160. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. en. *Nature* **460.** Number: 7256 Publisher: Nature Publishing Group, 748–752. ISSN: 1476-4687. https://www.nature.com/articles/nature08185 (2024) (Aug. 2009).

161. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** s13742–015–0047–8. ISSN: 2047-217X. https://doi.org/10.1186/s13742-015-0047-8 (2024) (Dec. 2015).

162. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58.** Publisher: [Royal Statistical Society, Wiley], 267–288. ISSN: 0035-9246. https://www.jstor.org/stable/2346178 (2024) (1996).

163. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. en. *Genetic Epidemiology* **41.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.22050, 469–480. ISSN: 1098-2272. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22050 (2024) (2017).

164. Hou, K. *et al. Admix-kit: An Integrated Toolkit and Pipeline for Genetic Analyses of Admixed Populations* en. Pages: 2023.09.30.560263 Section: New Results. Oct. 2023. https://www.biorxiv.org/content/10.1101/2023.09.30.560263v1 (2024).

165. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27,** 2304–2305. ISSN: 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150040/ (2024) (Aug. 2011).

166. Massarat, A. R. *et al.* Haptools: a toolkit for admixture and haplotype analysis. *Bioinformatics* **39,** btad104. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btad104 (2024) (Mar. 2023).

167. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. en. *Nature* **590.** Number: 7845 Publisher: Nature Publishing Group, 290–299. ISSN: 1476-4687. https://www.nature.com/articles/s41586-021-03205-y (2022) (Feb. 2021).

168. Das, S. *et al.* Next-generation genotype imputation service and methods. en. *Nature Genetics* **48.** Number: 10 Publisher: Nature Publishing Group, 1284–1287. ISSN: 1546-1718. https://www.nature.com/articles/ng.3656 (2022) (Oct. 2016).

169. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31,** 782–784. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btu704 (2022) (Mar. 2015).

170. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. eng. *American Journal of Human Genetics* **103,** 338–348. ISSN: 1537-6605 (Sept. 2018).

171. Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with FLARE. English. *The American Journal of Human Genetics* **110.** Publisher: Elsevier, 326–335. ISSN: 0002-9297, 1537-6605. `https://www.cell.com/ajhg/abstract/S0002-9297(22)00544-4` (2024) (Feb. 2023).

172. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32,** 1479–1485. ISSN: 1367-4803. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4866519/` (2024) (May 2016).

173. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19,** 1655–1664. ISSN: 1088-9051. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752134/` (2024) (Sept. 2009).

174. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. en. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 1776. ISSN: 2041-1723. `https://www.nature.com/articles/s41467-019-09718-5` (2024) (Apr. 2019).

175. Uren, C., Hoal, E. G. & Möller, M. Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genetics* **21,** 40. ISSN: 1471-2156. `https://doi.org/10.1186/s12863-020-00845-3` (2024) (Apr. 2020).

176. Qi, Y. *et al.* Pseudogenes in Cardiovascular Disease. *Frontiers in Molecular Biosciences* **7,** 622540. ISSN: 2296-889X. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7902774/` (2024) (Feb. 2021).

177. Lu, M. *et al.* TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies. *Nucleic Acids Research* **51,** D1179–D1187. ISSN: 0305-1048. `https://doi.org/10.1093/nar/gkac821` (2024) (Jan. 2023).

178. Díez-Obrero, V. *et al.* Transcriptome-Wide Association Study for Inflammatory Bowel Disease Reveals Novel Candidate Susceptibility Genes in Specific Colon Subsites and Tissue Categories. *Journal of Crohn's and Colitis* **16,** 275–285. ISSN: 1873-9946. `https://doi.org/10.1093/ecco-jcc/jjab131` (2024) (Feb. 2022).

179. Leníček, M. *et al.* The Relationship Between Serum Bilirubin and Crohn's Disease. *Inflammatory Bowel Diseases* **20,** 481–487. ISSN: 1078-0998. `https://doi.org/10.1097/01.MIB.0000440817.84251.98` (2024) (Mar. 2014).

180. Homuth, G. *et al.* Extensive alterations of the whole-blood transcriptome are associated with body mass index: results of an mRNA profiling study involving two large population-based cohorts. *BMC Medical Genomics* **8,** 65. ISSN: 1755-8794. `https://doi.org/10.1186/s12920-015-0141-x` (2024) (Oct. 2015).

181. Mortlock, S. *et al.* Tissue specific regulation of transcription in endometrium and association with disease. *Human Reproduction* **35,** 377–393. ISSN: 0268-1161. `https://doi.org/10.1093/humrep/dez279` (2024) (Feb. 2020).

182. Dinsdale, N. L. & Crespi, B. J. Endometriosis and polycystic ovary syndrome are diametric disorders. en. *Evolutionary Applications* **14.** Publisher: John Wiley & Sons, Ltd, 1693–1715. ISSN: 1752-4571. `https://onlinelibrary.wiley.com/doi/10.1111/eva.13244` (2024) (July 2021).

183. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. en. *Nature Genetics* **52.** Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic association study;Genome-wide association studies Subject_term_id: genetic-association-study;genome-wide-association-studies, 541–547. ISSN: 1546-1718. `https://www.nature.com/articles/s41588-020-0613-6` (2022) (May 2020).