UNIVERSITY OF CALIFORNIA
Los Angeles

Computational methods to analyze large-scale genetic studies of complex human traits

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Huwenbo Shi

2018

ABSTRACT OF THE DISSERTATION

Computational methods to
analyze large-scale genetic studies of
complex human traits

by

Huwenbo Shi
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Bogdan Pasaniuc, Chair

Large-scale genome-wide association studies (GWAS) have produced a rich resource of genetic data over the past decade, urging the need to develop computational and statistical methods that analyze these data. This dissertation presents four statistical methods that model the correlation structure between genetic variants and its effect on GWAS summary association statistics to help understand the genetic basis of complex human traits and diseases.

The first method employs the multivariate Bernoulli distribution to model haplotype data, allowing for higher-order interactions among genetic variants, and shows better accuracy in predicting DNase I hypersensitivity status.

The second method partitions heritability into small regions on the genome using GWAS summary statistics data, while accounting for complex correlation structures among genetic variants, and uncovers the genetic architectures of complex human traits and diseases.

Extending the second method into pairs of traits, the third method partitions genetic correlation into small genomic regions using GWAS summary statistics data, and provides insights into the shared genetic basis between pairs of traits.

Finally, the fourth method dissects population-specific and shared causal genetic variants of complex traits in two continental populations, using GWAS summary statistics data obtained from samples of different ethnicities, and reveals differences in genetic architectures of two continental populations.

The dissertation of Huwenbo Shi is approved.

Arash Amini

Rita Cantor

Päivi Pajukanta

Kenneth Lange

Janet Sinsheimer

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2018

To my parents and grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my thesis advisor, Bogdan Pasaniuc, for his patience, rigor, and encouragement, in cultivating me to become an independent researcher. His advice and support paved the way to my future career.

I would also like to thank both current and previous members of the Bogdan lab: Gleb Kichaev, Nicholas Mancuso, Claudia Giambartolomei, Robert Brown, Megan Roytman, Ruth Johnson, Kathryn Burch, Valerie Arboleda, Malika Freund, Arunabha Majumdar, Megan Major, Robert Smith, Tommer Schwarz, Wen-Yun Yang, and Page Goddard, for helpful and productive intellectual discussions, and more importantly, for the fun that they brought to my life as a Ph.D. student.

I owe special thanks to my previous Bruins in Genomics (BIG) summer students: Sarah Spendlove, Astrid Manuel, Natalie Dong, Christian Torres, and Anthony Fernandez for helping out on my research projects during the summer.

My adventure as a Ph.D. student wouldn't have been a smooth sail without the generous help and advice from my committee members: Janet Sinsheimer, Kenneth Lange, Päivi Pajukanta, Rita Cantor, Arash Amini, and our friends and collaborators both at UCLA and other institutions: Sriram Sankararaman, Jonathan Flint, Na Cai, Jeremy Rotman, Artur Jaroszewicz, Eleazar Eskin, Eric Sobel, Alexander Gusev, Alkes Price, and Zhaozhong Zhu. I owe them a big thank you.

I also owe a big thank you to the Chinese community at UCLA, USC, and Caltech: Zhicheng Pan, Zijun Zhang, Liangke Gou, Hanjun Cheng, Ning Wang, Chengyang Wang, Yuanyuan Wang, Yang Pan, Peng He, Ruyi Huang, Xiyu Yi, Wanlu Liu, Linghao Gao, Junhui Hu, Jiatong Chen, Jiajin Li, Weixian Deng, Jihui Sha, Kai Fu, Yida Zhang, and Yan Gao, without whom my life as a Ph.D. student would have been much less colorful.

Last but not least, I am deeply grateful to my father Guochao Shi, my mother Bingbing Hu, and all other members of my family, for their understanding and support during my time as a

Ph.D. student. I couldn't imagine completing the Ph.D. journey without the understanding and support from my family.

Thank you all!

VITA

2009 – 2013

B.S. (Computer Science), University of California Los Angeles.

Summer 2012

Software Development Engineer Intern, Amazon.com, Seattle, Washington

2014 – present

Graduate Student Researcher, Bogdan Pasaniuc Lab, University of California Los Angeles.

Winter 2018

Teaching Assistant for Computer Science CM122. Algorithms in Bioinformatics and Systems Biology, University of California Los Angeles.

PUBLICATIONS

Shi H., Burch K., Johnson R., Freund M., Kichaev G., Mancuso N., Manuel A., Dong N., Pasaniuc B., "Dissecting genetic architectures of complex traits specific to and shared by East Asian and European populations" in prep

Freund M., Burch K, **Shi H.**, Mancuso N., Kichaev G., Garske K., Pan D., Pajukanta P., Pasaniuc B., Arboleda V., "Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits" in prep

Mancuso N., Kichaev G., **Shi H.**, Freund M., Giambartolomei C., Gusev A., Pasaniuc B., "Probabilistic fine-mapping of transcriptome-wide association studies" bioRxiv 2018

Johnson R., **Shi H.**, Pasaniuc B., Sankararaman S., "A Unifying framework for joint trait analysis under a non-infinitesimal model." ISMB 2018

Giambartolomei C., Liu J., Zhang W., Hauberg M., **Shi H.**, Pickrell J., Jaffe A.E., the CommonMind Consortium, Pasaniuc B., Roussos P. "A Bayesian framework for multiple trait colocalization from summary association statistics." Bioinformatics 2018.

Shi, H., Mancuso, N., Spendlove, S., Pasaniuc, B. "Local genetic correlation gives insights into the shared genetic architecture of complex traits." The American Journal of Human Genetics 2017.

Mancuso, N., **Shi, H.**, Goddard, P., Kichaev, G., Gusev, A., Pasaniuc, B. "Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits." The American Journal of Human Genetics 2017.

Gusev A., **Shi H.**, Kichaev G., Price A., Pasaniuc B., et al. "Genomic functional atlas of prostate cancer heritability in European and African Americans reveals extensive tissue-specific regulation." Nature Communication 2016.

Shi, H., Kichaev, G., Pasaniuc, B. "Contrasting the genetic architecture of 30 complex traits from summary association data." The American Journal of Human Genetics 2016.

Gusev A., Ko Arthur., **Shi H.**, Bhatia G., Price A., Pajukanta P., Pasaniuc B., et al. "Integrative approaches for large-scale transcriptome-wide association studies." Nature Genetics 2016.

Shi, H., Pasaniuc, B., Lange, K. "A multivariate Bernoulli model to predict DNaseI hypersensitivity status from haplotype data." Bioinformatics 2015.

Orozco L.D., Morselli M., Rubbi L., Guo W., Go J., **Shi H.**, Lopez D., Furlotte N.A., Bennett B.J., Farber C.R., Ghazalpour A., Zhang M.Q., Bahous R., Rozen R., Lusis A.J., Pellegrini M. "Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice." Cell Metabolism 2015

Pasaniuc B., Zaitlen N., **Shi H.**, Bhatia G., Gusev A., Pickrell J., Hirschhorn J., Strachan D.P., Patterson N., Price A.L. "Fast and accurate imputation of summary statistics enhances evidence of functional enrichment." Bioinformatics 2014

CHAPTER 1

Introduction

Complex human traits and diseases are driven by both genetic and environmental factors. Since genetics is intrinsic to every person, studying the genetic basis of complex traits offers an unbiased approach to understand the biological mechanisms behind complex traits. A conceptually simple but highly effective approach to assess the contribution of genetics on complex traits is genome-wide association study, which scans for association between each genetic variant and complex trait. The drastic decrease in sequencing and genotyping technologies enabled genome-wide association studies at a large scale, which have produced a rich resource of genetic data over the past decade. These large-scale genetic studies revealed both valuable biological insights and challenges in genetic studies, urging the need to develop new and efficient computational and statistical methods to analyze these data. The next four chapters introduce methods for modeling linkage and its effect on GWAS results.

The non-random crossover during meiosis creates dependencies betweens alleles on haplotypes, sequences of alleles on one copy of chromosome, inducing correlation (linkage disequilibrium) between the alleles. Modeling linkage disequilibrium in haplotypes has a wide range of applications in population inference and disease gene discovery. The hidden Markov models (HMM) traditionally used for haplotypes[84] are hindered by the dubious assumption that dependencies occur only between consecutive pairs of variants. In Chapter 2, we apply the multivariate Bernoulli (MVB) distribution [30] to model haplotype data. The MVB distribution relies on interactions among all sets of variants, thus allowing for the detection and exploitation of long-range and higher-order interactions [30]. We discuss penalized estimation and present an efficient algorithm for fitting sparse versions of the MVB distribution to

haplotype data. Finally, we showcase the benefits of the MVB model in predicting DNaseI hypersensitivity (DH) status – an epigenetic mark describing chromatin accessibility– from population-scale haplotype data. We fit the MVB model to real data from 59 individuals on whom both haplotypes and DH status in lymphoblastoid cell lines are publicly available. The model allows prediction of DH status from genetic data (prediction $R^2 = 0.12$ in cross-validations). Comparisons of prediction under the MVB model with prediction under linear regression (best linear unbiased prediction or BLUP) and logistic regression demonstrate that the MVB model achieves about 10% higher prediction R^2 than the two competing methods in empirical data.

Linkage disequilibrium has profound impact on genome-wide association studies (GWAS) of complex traits [20]. Modeling the effect of linkage disequilibrium on between GWAS results is crucial for the correct interpretation of important quantities in genetics, such as heritability, the fraction of variance in trait explained by genetic variation. While GWAS have identified thousands of genetic variants associated with complex human traits and diseases, a large fraction of heritability of complex traits remain unexplained by genetic variants identified through GWAS [95]. A possible explanation to this discrepancy is that most genetic variants have effect too small to be detected at the current sample size. Variance components methods that estimate the aggregate contribution of large sets of variants to the heritability of complex traits have yielded important insights into the genetic architecture of common diseases. In Chapter 3, we introduce new methods that estimate the total variance in trait explained by the typed variants at a single locus in the genome (local SNP-heritability) from summary GWAS data while accounting for linkage disequilibrium (LD) among variants. We apply our new estimator to ultra large-scale GWAS summary data of 30 common traits and diseases to gain insights into their local genetic architecture. First, we find that common SNPs have a high contribution to the heritability of all studied traits. Second, we identify traits for which the majority of the SNP heritability can be confined to a small percentage of the genome. Third, we identify GWAS risk loci where the entire locus explains significantly more variance in the trait than the GWAS reported variants. Finally, we identify loci that explain significant amount of heritability across multiple traits.

The rich resource of genetic data curated through GWAS also facilitate the understanding of shared genetic basis between pairs of complex traits, which has been traditionally quantified through genetic correlation, correlation between complex traits driven by genetic variations [18]. Although genetic correlations between complex traits provide valuable insights into epidemiological and etiological studies, a precise quantification of which genomic regions disproportionately contribute to the genome-wide correlation is currently lacking. In Chapter 4, we introduce ρ-HESS, a technique to quantify the correlation between pairs of traits due to genetic variation at a small region in the genome. Our approach only requires GWAS summary data, and makes no distributional assumption on the causal variant effect sizes while accounting for linkage disequilibrium (LD) and overlapping GWAS samples. We analyzed large-scale GWAS summary data across 36 quantitative traits, and identified 25 genomic regions that contribute significantly to the genetic correlation among these traits. Notably, we find 6 genomic regions that contribute to the genetic correlation of 10 pairs of traits that show negligible genome-wide correlation, further showcasing the power of local genetic correlation analyses. Finally, we report the distribution of local genetic correlations across the genome for 55 pairs of traits that show putative causal relationships.

Genome-wide association studies (GWAS) have been predominantly performed in European populations, limiting the transferability of GWAS results into other populations. The recent increase in the number GWASs in non-European populations creates opportunities for trans-ethnic studies to improve disease mapping, fine mapping, risk prediction, and transferability of GWAS results. Differences in linkage disequilibrium patterns of two continental populations arose during the history of evolution, and need to be accounted for in trans-ethnic genetic studies. A quintessential theme of trans-ethnic genetic studies is the degree of genetic overlap of a complex trait across two populations. In Chapter 5, we introduce POSC, a method to dissect genetic architectures that are specific to a continental populations and those are shared by both populations. We applied POSC on summary statistics data of large-scale GWAS of anthropometric, hematological, immunological, and psychiatric traits and diseases, obtained from samples of East Asian and European descent. We show that complex traits harbor genetic architectures that are both specific to a population and

shared by both populations. We also quantify enrichment of population-specific and shared causal variants in regions of genes specifically expressed across 53 GTEx tissues, and find that there are enrichments of both population-specific and shared causal variants.

CHAPTER 2

A multivariate Bernoulli model to predict DNaseI hypersensitivity status from haplotype data

2.1 Introduction

Accidents of history and variable recombination rates have divided the human genome into blocks of shared recent ancestry [27, 31, 50]. Ancestry sharing manifests itself in complex haplotype patterns and strong dependencies among variants. (Recall that a haplotype summarizes the sequence of alleles displayed by the sampled markers in a narrow genomic region of a particular chromosome [77].) Therefore, modeling haplotype data is of paramount importance for a wide range of problems in population genetics and disease gene discovery [24, 67, 68, 81, 86, 92, 99, 107, 121, 126, 130, 136, 153].

Haplotypes have been traditionally analyzed by hidden Markov models (HMMs) [84, 172], with emissions corresponding to observed genotypes and transitions to recombination events. Although HMMs for haplotypes undergird many efficient and accurate algorithms for haplotype phasing [137], genotype imputation [17, 69, 85], and identity-by-descent detection [16], they suffer from the drawback of modeling only dependencies between consecutive variants. This assumption leads to the unrealistic conclusion that the previous variant and

This chapter is published in Shi et al., Bioinformatics 2015 [141]

the next variant are independent given the current variant. Ignoring dependencies among non-consecutive markers makes it difficult to detect and exploit long range correlations and higher-order interactions among variants. These complex dependencies definitely exist in the human genome and are important factors in genetic studies [131, 163].

The current paper applies the multivariate Bernoulli (MVB) distribution to haplotype data. The MVB distribution captures the entire spectrum of dependencies among the entries of random binary vectors of length N [30]. The observed haplotypes at N nearby SNPs (single nucleotide polymorphisms) can be thought of as realizations of such a process. Since there are 2^N possible haplotyes for N SNPs, the MVB distribution requires an unsustainable exponential number of parameters. Vast amounts of training data or clever algorithms cannot compensate for this combinatorial explosion. Here we investigate a Poisson re-parameterization of the MVB distribution and impose an $\ell 1$-norm penalty to enforce sparsity in parameter estimation. These steps allow us to devise an efficient coordinate ascent algorithm for learning the MVB parameters from haplotype data while restricting the number of parameters to a manageable level.

We showcase the utility of the MVB model by predicting an individual's DNaseI hypersensitivity status from haplotypes observed near known DNaseI hypersensitivity sites. DNaseI hypersensitivity (DH) status is a mark of open chromatin and flags genomic regions where the DNA is accessible to the DNaseI enzyme. These regions, such as transcription start sites, correlate with active DNA regulation. DH status is usually assayed through DNase-Seq, a genome-wide high-throughput technology that sequences genomic regions sensitive to DNaseI [94]. Recent research [36] suggests that genetic variants control this epigenetic mark. Since DH status can be naturally encoded as a binary variable, the MVB model offers a natural way to integrate DH status and local haplotype data. In predicting DH status from haplotypes, the MVB model allows all allelic sets to contribute regardless of the order of the participating SNPs and the physical distances separating them.

Our analysis of data from the 1000 Genomes project [27] demonstrates the superiority of the sparse MVB distribution in model fitting. In practice, interactions beyond order three play

little role in determining haplotype frequencies in these data. Our new cyclic coordinate descent algorithm for estimating the MVB interaction parameters converges quickly and reliably. The MVB model also turns out to be pertinent to predicting DH status from haplotype data at known DH sites [34]. On a sample of just 59 subjects, cross-validation under the MVB yields a prediction R^2 of 0.12 for dichotomized DH levels. As expected, the accuracy of DH prediction decreases as extraneous predictors are added. Finally, prediction under the MVB achieves about 10% better accuracy than prediction by linear regression (best unbiased linear predictor or BLUP) and logistic regression. Thus, the MVB model is recommended for prediction of binary epigenetic status from local haplotype data.

2.2 Methods

2.2.1 The multivariate Bernoulli distribution as a model for haplotype data

The multivariate Bernoulli (MVB) distribution extends the univariate Bernoulli distribution to binary vectors of fixed length N [30]. The density $\Pr(Y = \mathbf{y}) = p_{(y_1,\ldots,y_N)}$ of such a discrete random vector Y depends on 2^N probabilities $p_{(0,0,\ldots,0)}$, $p_{(0,0,\ldots,1)}, \ldots,$ $p_{(1,1\ldots,1)}$ specific to the different realizations of Y. For example, the bivariate Bernoulli distribution consists of four realizations $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ specified by four probabilities $p_{(0,0)}$, $p_{(0,1)}$, $p_{(1,0)}$, and $p_{(1,1)}$. By definition the conditional distribution of a subvector, say $(Y_1, Y_2, \ldots, Y_k)$, given the complementary subvector, say $(Y_{k+1}, \ldots, Y_N)$, is also MVB. In the bivariate case, conditioning on either Y_1 or Y_2 produces a standard univariate Bernoulli distribution. There is an alternative parameterization that captures interactions and is conducive to parsimony. This parameterization substitutes subsets of $\{1, \ldots, N\}$ for binary vectors. To the realization $\mathbf{y}$ we correspond the index set $A = \{i : y_i = 1\}$. The natural parameters f_C of the MVB model are indexed by interaction subsets C, and the density function $\Pr(Y = \mathbf{y})$ is written as the ratio

$$\Pr(A) = \frac{\exp\left(\sum_{C \subseteq A} f_C\right)}{\sum_B \exp\left(\sum_{C \subseteq B} f_C\right)} = \frac{\exp\left(S_A\right)}{\sum_B \exp\left(S_B\right)}, \tag{2.1}$$

where we define $S_A = \sum_{C \subseteq A} f_C$ for notational simplicity. The denominator $\sum_B \exp(S_B)$ is the appropriate normalizing constant.

The haplotypes spanning N bi-allelic SNPs can be represented as binary vectors of length N. We adopt the convention that $y_i = 0$ indicates the major allele and $y_i = 1$ indicates the minor allele at SNP i. One can obviously model the distribution of haplotypes in a population as MVB. The major advantage of the MVB is its ability to incorporate interactions in the recovery of haplotype frequencies. The number of parameters in both the naive and interaction parameterizations grows exponentially fast in N. However, the interaction parameterization organizes interactions by level and suggests limiting model complexity by imposing an upper bound on interaction level. The next section introduces a lasso penalty that in combination with maximum likelihood estimation eliminates superfluous interactions and keeps the number of levels in check.

2.2.2 Estimating MVB parameters from haplotype data

To estimate haplotype frequencies and ultimately infer missing haplotypes, one can randomly sample a population and count the number X_A of haplotypes of each type A. For a fixed sample size M, the X_A jointly follow a multinomial distribution with M total counts and the count probabilities $\Pr(A)$ displayed in equation (2.1). Alternatively, one can adopt a Poisson rather than a multinomial sampling framework. The two share the assumption of independent samples but differ in whether the total sample size is random (Poisson) or fixed (multinomial). The law of small numbers justifies the equivalence of the two frameworks. The Poisson setting invokes a mean sample size μ, which is estimated by the observed sample size $\sum_A X_A$. One can show [78] that the random variables X_A are independent and Poisson distributed with means $\mu_A = \mu \Pr(A)$.

In the Poisson framework, it is easier work with the interaction parameters by setting $\mu_A = \exp(S_A) = \exp(\sum_{B \subseteq A} f_B)$ and ignoring μ and the normalizing constant $\sum_B \exp(S_B)$. In effect, these are absorbed into the empty set parameter $f_\varnothing$. Independence of the X_A now

yields the likelihood

$$\mathcal{L}(\mathbf{f}|\mathbf{X}) = \prod_A \frac{(\mu_A)^{X_A}}{X_A!} \exp(-\mu_A), \tag{2.2}$$

where $\mathbf{X} = (X_A)$ and $\mathbf{f} = (f_A)$ are the vectors of haplotype counts and interaction parameters, respectively. Taking logarithms produces the loglikelihood

$$\ell(\mathbf{f}|\mathbf{X}) = \sum_A f_A \sum_{B \supseteq A} X_B - \sum_A \exp(S_A) - \sum_A \log X_A!. \tag{2.3}$$

It is natural to estimate the MVB parameter vector $\mathbf{f} = (f_A)$ by maximizing $\ell(\mathbf{f}|\mathbf{X})$.

Unless N is small and the sample size M is large, estimating all 2^N MVB parameters is an exercise in over-fitting. To achieve parsimony, we append an ℓ_1-norm (lasso) penalty to the loglikelihood. Any reasonable model should include the low-order parameters f_A with $|A| \leqslant 1$, where $|A|$ denotes the cardinality of the set A. Hence, we maximize the penalized loglikelihood

$$F(\mathbf{f}) = \sum_A f_A \sum_{B \supseteq A} X_B - \sum_A \exp(S_A) - \lambda \sum_{|A| \geqslant 2} |f_A|. \tag{2.4}$$

Here, λ is a tuning constant determining the strength of the penalty. Increasing λ increases the sparsity of the estimated parameter vector. The analogy with lasso-guided regression is obvious. The new objective function $F(\mathbf{f})$ is concave and directionally differentiable. It has kinks introduced by the terms $|f_A|$. We recommend maximization by coordinate ascent.

2.2.3 Coordinate ascent algorithm

Coordinate ascent maximizes the objective function one parameter at a time holding other parameters fixed. Cycling through the parameters continues until the objective value converges or a maximum number of iterations is reached. Algorithm 1 outlines the coordinate ascent algorithm for estimating model parameters.

Line 5 of Algorithm 1 requires finding $\arg\max_{f_A} F(\mathbf{f})$. To update f_A when $|A| \leqslant 1$, we set

Algorithm 1 coordinate ascent algorithm for fitting the MVB

Let $\mathcal{C}$ be the collection of possible haplotypes of length N
Initialize f_A to 0 for all $A \in \mathcal{C}$
while stop condition fails **do**
  for A in $\mathcal{C}$ **do**
    $f_A = \arg\max_{f_A} F(\mathbf{f})$
  end for
end while

the partial derivative of $F(\mathbf{f})$

$$\frac{\partial}{\partial f_A} F(\mathbf{f}) = \sum_{B \supseteq A} X_B - e^{f_A} \sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C} \tag{2.5}$$

with respect to f_A equal to 0. This yields the update

$$f_A = \ln \frac{\sum_{B \supseteq A} X_B}{\sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C}}. \tag{2.6}$$

When $|A| \geqslant 2$, the supergradient

$$\begin{aligned} \frac{\partial}{\partial f_A} F(\mathbf{f}) = & \sum_{B \supseteq A} X_B - e^{f_A} \sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C} \\ & - \lambda \begin{cases} 1 & \text{if } f_A > 0 \\ [-1, 1] & \text{if } f_A = 0 \\ -1 & \text{if } f_A < 0 \end{cases} \end{aligned} \tag{2.7}$$

must contain 0 [79]. Equating it to 0 yields the update

$$f_A = \begin{cases} 0 & |c| \leqslant \lambda \\ \ln \frac{\sum_{B \supseteq A} X_B - \lambda}{\sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C}} & c > \lambda \\ \ln \frac{\sum_{B \supseteq A} X_B + \lambda}{\sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C}} & c < -\lambda \end{cases} \tag{2.8}$$

for the criterion $c = \sum_{B \supseteq A} X_B - \sum_{B \supseteq A} e^{\sum_{C \subseteq B,\, C \neq A} f_C}$.

In view of the summations over $B \supseteq A$ in the denominators of equation (2.6) and equation (2.8), each coordinate ascent update takes nearly $O(2^N)$ operations. This computational load restricts estimation to MVB models with small N, say $N \leqslant 15$. Once parameters are estimated, prediction under the MVB is relatively straightforward. The normalizing constant in formula (2.1) must be calculated, but this can be done once and the result stored.

2.2.4 Best linear unbiased predictor (BLUP)

Part of our evaluation of the MVB involves comparison of DNaseI hypersensitivity (DH) prediction on simulated data. The simulated DH status y_i of an individual i was constructed as a linear combination of individual i's SNP alleles and SNP pairwise interactions weighted by effect sizes β_j and β_{jk}. In symbols

$$y_i = \sum_j \beta_j h_{ij} + \sum_{\{j,k\}} \beta_{jk} h_{ij} h_{ik} + \varepsilon_i, \tag{2.9}$$

where h_{ij} is the SNP predictor (standardized version of 0 or 1) of individual i at SNP j, $h_{ij}h_{ik}$ is the SNP interaction of individual i for the pair of SNPs j and k, and ε_i is an independent normally distributed error term. Simplified versions of the model ignore the pairwise interactions and take all $\beta_{jk} = 0$.

To make predictions under the linear model, we first estimate the effect sizes β_j and β_{jk} from training data set and then predict the phenotype (DH status) of each individual in the test data, substituting estimated parameters for true parameters. For notational brevity, let $H = (H_{SNP},\ H_{INT})$ be the block matrix of single SNP and interaction SNP predictors across the training set; for each subject i and SNPs j and k, the matrix H_{SNP} has entries (h_{ij}), and the matrix H_{INT} has entries $(h_{ij}h_{ik})$. The effect sizes β_j and β_{jk} are estimated by the least squares formula

$$\hat{\beta} = (H^T H)^{-1} H^T y. \tag{2.10}$$

Finally, the best linear unbiased predictor (BLUP) $\hat{y}_i$ of DH status for an individual i is

computed via

$$\hat{y}_i = \sum_j \hat{\beta}_j h_{ij} + \sum_{\{j,k\}} \hat{\beta}_{jk} h_{ij} h_{ik}. \tag{2.11}$$

2.2.5 Logistic regression (LOGIT)

We also compared the MVB model with logistic regression (LOGIT); unlike linear regression, logistic regression directly models binary outcomes. Under logistic regression, the probability of the DH status y_i of individual i given his/her SNP alleles (h_{ij}) and pairwise interactions ($h_{ij}h_{ik}$) is

$$\Pr(y_i = y) = \left(\frac{e^{c_i}}{1+e^{c_i}}\right)^y \left(\frac{1}{1+e^{c_i}}\right)^{1-y}, \tag{2.12}$$

where $c_i = \alpha_0 + \sum_j \alpha_j h_{ij} + \sum_{\{j,k\}} \alpha_{jk} h_{ij} h_{ik}$. Here the α's are the regression coefficients in logistic regression. As with linear regression, one can simplify the model by ignoring pairwise interactions and taking all $\alpha_{jk} = 0$. To estimate the parameters of the model, one maximizes the likelihood

$$\prod_{\{i:y_i=1\}} \frac{e^{c_i}}{1+e^{c_i}} \prod_{\{i:y_i=0\}} \frac{1}{1+e^{c_i}}. \tag{2.13}$$

over the entire sample. Prediction of the DH status of individual i relies on the the predicted probability

$$\hat{y}_i = \frac{e^{\hat{c}_i}}{1+e^{\hat{c}_i}}, \tag{2.14}$$

of $y_i = 1$, where $\hat{c}_i$ is the same as c_i except for substitution of estimated regression coefficients for true coefficients.

2.2.6 Hidden Markov model (HMM) for haplotypes

A hidden Markov model (HMM) views a haplotype $\mathbf{h}$ of length N as a mosaic of haplotypes from a set $\mathcal{H}$ of R reference haplotypes [84]. The $N \times R$ HMM states (i, j) capture the particular reference haplotype j occurring at SNP i. A transition matrix $\mathbf{K}$ models recombination events and controls how switches occur between haplotypes in meiosis. The entries

$K[(ij),(kl)]$ of the transition matrix are 0 unless $k = i + 1$. For neighboring SNPs, the entries depend on the distance between the SNPs. Thus, the larger the distance, the larger the transition probability for $j \neq l$. The emission probabilities $\Pr(h_i|(ij))$ allow for mistyping and occasional mutation events. Inferences based on HMM are achieved efficiently through the forward, backward, and Viterbi algorithms, all of which have complexity $O(NR^2)$. We adopt the latest IMPUTE2 [69, 66] implementation of HMM for comparison purposes.

2.3 Results

2.3.1 Assessment of MVB on 1000 Genome haplotypes

In an initial set of experiments, we used the 1000 Genomes EUR (European) haplotypes (505 individuals) to investigate the performance of the MVB model and our coordinate descent algorithm for fitting it to data. We randomly selected 50 regions on chromosome 1, each containing 15 SNPs, and fit the MVB under various settings. The first setting imposed no constraint on the maximum order (max $|A|$) of the interaction sets A. Thus, in effect, we estimated all $2^{15} = 32,768$ parameters. Figure 2.1 shows that the regularization constant λ has a significant effect on the magnitude of parameters, especially for f_A's where $|A| \geqslant 4$. For example, as λ increases from 0.0 to 0.5, the sum $\sum_{|A|=4} |f_A|$ of estimated parameters decreases from 87.5 to 30 for interaction sets with $|A| = 4$. Furthermore, Figure 2.2 indicates that the average value of $|f_A|$ converges to 0 as $|A|$ tends to $N = 15$. Thus, we conclude that the lower-order interactions f_A predominate in determining haplotype frequencies.

We also recorded the number of iterations until convergence of the coordinate descent algorithm. The algorithm invariably converges within 20 to 30 iterations. See Figure 2.3 for typical results. Finally, Table 2.2 shows that the bulk of computational time is taken in estimating MVB parameters; once model parameters are estimated, applying the model to making predictions is relatively trivial.

Next we investigated how well the MVB fits the selected 1000 Genomes haplotypes using just

lower-order interactions. To measure goodness of fit, we computed the Euclidean distance between the haplotype frequencies recovered by the MVB model as given in equation (2.1) and the haplotype frequencies observed in the data. Table 2.1 demonstrates that the MVB model requires only the lower-order interactions terms to accurately fit typical data. Because $\lambda = 0.25$ attains the best fit across interaction level bounds ($|A| \leqslant b$), we set λ to 0.25 in all future experiments.

2.3.2 Prediction of DNaseI hypersensitivity status in simulations

To simulate binary DNaseI hypersensitivity (DH) data, we took the 1,010 EUR (European) haplotypes of the 1000 Genome project [27] and simulated 20,000 haploid individuals at 200 randomly selected 20Kbp regions on chromosome 1 [150]. From each region we selected 15 SNPs with minor allele frequency above 1%. From the 15 chosen SNPs we randomly selected m causal SNPs and n pairs of interaction SNPs and simulated continuous DH values according to the linear model sketched in Section 2.2.4. Prior to simulation we standardized the SNP predictors h_{ij} and $h_{ij}h_{ik}$ to have mean 0 and variance 1. The regression coefficients for the causal SNPs and SNP pairs were sampled as $\beta_j \sim N(0, h^2/m)$ and $\beta_{jk} \sim N(0, h^2_{int}/n)$ and the noise for each DH variable as $\varepsilon_i \sim N(0, 1-(h^2+h^2_{int}))$, where h^2 and h^2_{int} denote the variance of DH values explained by single variants and interactions, respectively. Finally, we converted the continuous DH values to binary DH values by imposing a threshold chosen so that 20% of the binary DH values were elevated (status 1 rather than status 0).

For testing under the MVB model, we constructed binary vectors of length 16 by concatenating each 15-SNP haplotype and a corresponding simulated binary DH status. Given the tuning constant $\lambda = 0.25$, this allows us to estimate the f_A parameters. To predict DH status given observed SNP haplotypes, one simply computes a conditional probability under the MVB model. In one set of MVB trials, we limited the interaction level to $|A| \leqslant 2$, for a total of 137 parameters. In a second set of trials, we limited the interaction level $|A| \leqslant 3$, for a total of 697 parameters. One can compare MVB prediction to BLUP and LOGIT prediction based on the same SNP haplotypes and interaction model. For BLUP and LOGIT, we also

tested a model involving SNPs and interactions between adjacent SNPs.

In linear regression, equation (2.10) supplies effect sizes, and equation (2.11) supplies predicted values. In logistic regression, equation (2.14) supplies predicted values. For estimation and prediction under HMM, we concatenated DH status as a pseudo SNP at the end of each 15-SNP haplotype to avoid changing the SNP interactions in the original haplotype. We also set the physical distance between the pseudo SNP and the last SNP to be the average distance between consecutive pairs of SNPs in the original 15-SNP haplotype. We employed half of the simulated individuals as reference panel and ran HMM with IMPUTE2 default settings on the other half to obtain predicted DH status. All 200 simulations summarized below involve two causal SNPs ($m = 2$) and 2 causal SNP interactions ($n = 2$) for 200 randomly sampled individuals. Of these 200 people, 100 served as training individuals and 100 as validation individuals.

We first investigated performance of MVB, BLUP, LOGIT, and HMM prediction for varying h^2 for a fixed interaction h^2_{int} of 0.1. Figure 2.4 shows that prediction R^2 achieved by all models increases as h^2 increases. However, the MVB model consistently achieves higher prediction R^2 than BLUP, LOGIT, and HMM under both settings, suggesting that the MVB model is capable of yielding more accurate estimates of effect sizes for prediction.

Notably as h^2 increases, the improvement in prediction R^2 also increases. In other words, as the effect of a single SNP increases, the comparative advantage of the MVB model over BLUP, LOGIT, and HMM increases.

Next we investigated the accuracy of these approaches at varying h^2_{int} values. Figure 2.5 demonstrates that for all pairs of h^2 and h^2_{int}, the MVB model also achieves higher prediction R^2 than BLUP, LOGIT, and HMM.

Finally we investigated the number of samples required for accurate prediction. Figure 2.6 shows that although the MVB model requires more parameters than BLUP, LOGIT, and HMM, it is able to outperform these models even if the training sample size is small. This suggests that the MVB model is less sensitive to noise. Notably, HMM under-performs

both MVB and LOGIT in most simulation settings, suggesting that HMM is less capable of detecting long range interactions for reasonable sample sizes. Across all simulated data sets, we observe no major difference in prediction R^2 between the two MVB settings. This is to be expected since only pairwise interactions are simulated.

2.3.3 Predicting DNaseI hypersensitivity status in empirical data

We now turn to real data on DH status and reach similar conclusions. The data set in question [36] contains normalized DNaseI hypersensitivity (DH) scores for 70 YRI (Yorubas in Ibadan, Nigeria) individuals at 1.5 million 100-bp genomic windows. These windows cover the 5% of the human genome with the highest DNaseI sensitivity. About half of the windows are expected to be truly sensitive to DNaseI [14]; 8,902 windows have associated dsQTLs (SNPs showing significant correlations with DH scores across individuals [36]). We dichotomized DH scores by placing scores above the threshold of 0.0 in one category and scores below the threshold of 0.0 in the complementary category. Among the 70 YRI individuals in the sample, 59 are also in the 1000 Genome project [27] and have fully phased haplotypes. We accordingly used the haplotypes and the binary DH status of these 59 individuals to evaluate the MVB model. For computational reasons, we selected one haplotype for each individual and restricted our analysis to 250 random DH sites and the 377 DH sites with associated dsQTLs on chromosome 22.

In genomic windows with associated dsQTLs, the dsQTLs are on average about 8,000 base pairs (10 SNPs) away from their windows. This action at a distance renders it difficult for HMMs to accurately capture interactions between dsQTLs and their genomic windows. Because sequence order is an important factor for HMMs, the question also arises of where to place binary DH status (a pseudo SNP) in the haplotype. For this reason, we excluded HMM from comparisons on real data.

To avoid over-fitting, we assessed prediction accuracy by leave-one-out cross-validation. Thus, we estimated parameters using data from 58 (all but one) training individuals and

predicted DH status for the remaining validation individual. Repeating this process across all 59 individuals allowed us to compare predicted and true DH status. The results can be summarized in a squared Pearson correlation (prediction R^2). Prior to parameter estimation in each of the 59 folds, we selected a small number of relevant SNP predictors by linear regression and forward selection. Our selection procedure excluded SNPs with minor allele frequency below 1% or at a distance of 1 Mbp or greater from the center of the window. Each successive SNP entering the candidate list provided the greatest reduction of the current residual sum of squares.

Given a candidate set of SNP predictors P in the MVB model, we created binary haplotype vectors of length $|P| + 1$ from the SNPs and the binary DH status. We considered at most second-order interactions and set the penalty constant λ to 0.25. For BLUP and LOGIT, we considered three models, one limited to single SNPs, one involving both single SNPs and two-way interactions, and one involving single SNPs and only interactions between adjacent SNPs.

Figure 2.7a shows the prediction R^2 obtained through leave-one-out cross-validation averaged over the 250 randomly selected windows. Due to overfitting and our small sample size, the average prediction R^2 decreases for all methods as the number of predictors $|P|$ increases. The MVB model achieves higher prediction R^2 than BLUP and LOGIT over both settings. We repeated the same experiment on the 377 windows with associated dsQTLs. Again the MVB model consistently achieves higher prediction R^2 than BLUP and LOGIT (see Figure 2.7b). Figures 2.7c and 2.7d depict the distribution of prediction R^2's under each model. It is clear that the MVB models achieve more high prediction R^2's (greater than 0.2) than BLUP and LOGIT. One can legitimately conclude that the MVB model predicts DH status better than BLUP and LOGIT. Table 2.3 summarizes the average and standard error of prediction R^2 for some representative experiments.

2.4 Discussion

The current paper presents the multivariate Bernoulli (MVB) distribution as a vehicle for modeling haplotype data. Because the number of distinct haplotypes observed in a narrow genomic region tends to be small, the MVB model is typically wildly over-parameterized. To achieve parsimony, we propose a lasso penalty within a Poisson sampling framework. The penalized MVB model encourages the detection and exploitation of higher-order interactions among the underlying SNPs. In contrast to Markovian models, interactions extend beyond nearest neighbor and pairwise interactions. The interaction parameterization adopted here is more natural than the naive MVB parameterization implicitly seen in BLUP and LOGIT. Empirically, the interaction parameterization extracts more haplotype information and predicts with better accuracy.

Our application of the MVB model to predict DNaseI hypersensitivity (DH) status from observed haplotypes supports the utility of the model. We show that the MVB model achieves better accuracy than BLUP and LOGIT in predicting simulated DH status. The overall prediction R^2 achieved by MVB, BLUP, and LOGIT on real DH status suggests substantial heritability of this epigenetic signal.

In likelihood evaluation and parameter estimation, the computational complexity of the MVB models scales like 2^N for N SNPs. This harsh reality limits the applicability of the model to a small number of variants. Fortunately, even for small N, the MVB model offers valuable insights into genomic data. The MVB model may well be critical in predicting binary gene expression when a small number of causal variants localize within a gene. In particular, MVB profiles in cases and controls may help in fine mapping traits in genome-wide association studies. Overcoming the computational limits of the MVB model limit is high on our research agenda. Once this task is accomplished, it will be possible to apply the MVB model to pre-phasing, a technique for improving genotype imputation by first imputing haplotypes [66]. We conjecture that Monte Carlo methods will play a decisive role in extending the range of the model to larger N. Finding an efficient sampling scheme

to approximate the normalization constant $\sum_B \exp(S_B)$ is of paramount importance and doubtless the place to start in accelerating algorithm performance.

2.5 Tables

		λ				
max $\|A\|$	**no. param.**	0.0	**0.25**	0.5	0.75	1.0
1	16	0.348	**0.348**	0.348	0.348	0.348
2	121	0.137	**0.072**	0.073	0.074	0.075
3	576	0.120	**0.054**	0.055	0.056	0.056
4	1,941	0.120	**0.055**	0.056	0.057	0.058

Table 2.1: **Euclidean distance between haplotype frequencies recovered by the MVB model and haplotype frequencies observed in data for different values of max $|A|$ and λ.**

max $\|A\|$	**Learning (sec/iter)**	**Prediction (sec/pred)**
1	0.2	< 0.01
2	1.1	< 0.01
3	4.4	0.01
4	13.7	0.02

Table 2.2: **Learning time (second per iteration) and prediction time (second per prediction), averaged over 50 loci.**

	$\|P\|$	MVB($\|A\| \leqslant 2$)	LOGIT	BLUP
RANDOM	1	.112±.015	.093±.013	.097±.013
	2	.109±.015	.106±.015	.100±.014
dsQTL	1	.120±.015	.108±.015	.114±.015
	2	.102±.014	.100±.015	.096±.014

Table 2.3: **Average prediction R^2 and standard error for $|P| \leqslant 2$ over 250 randomly selected windows (RANDOM) and 377 windows with dsQTLs (dsQTL).**

2.6 Figures

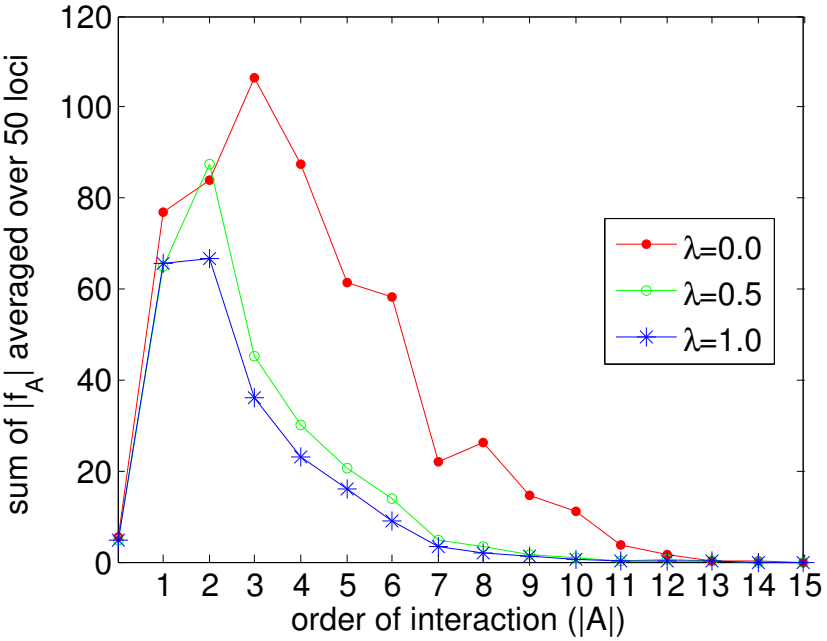


Figure 2.1: **Sum of $|f_A|$'s averaged over 50 regions as a function of $|A|$.**

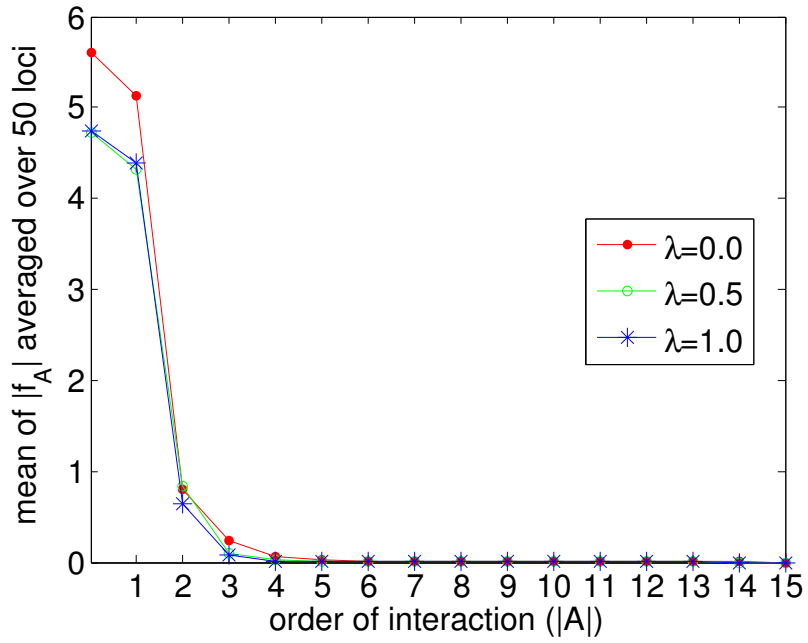


Figure 2.2: **Mean of $|f_A|$'s averaged over 50 loci as a function of $|A|$.**

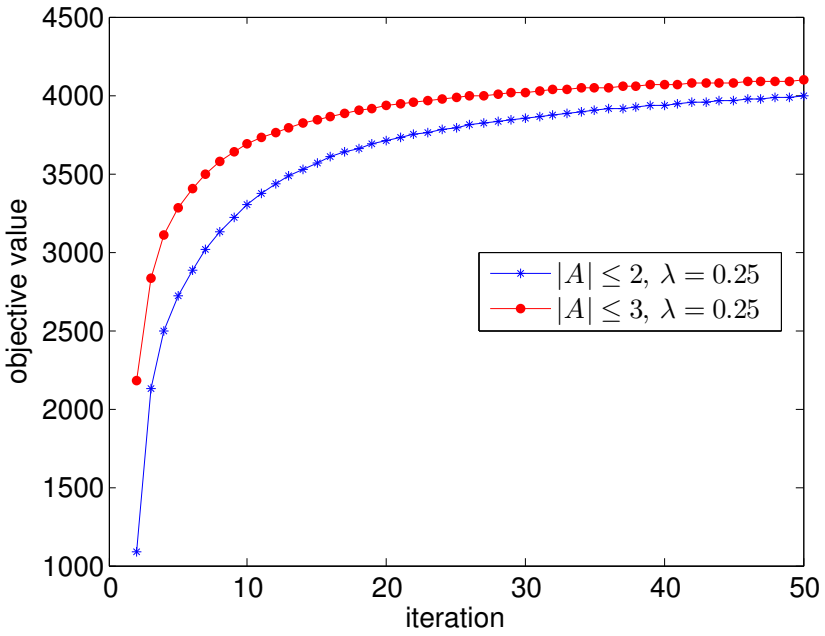


Figure 2.3: **Objective value averaged over 50 loci at each iteration of the coordinate ascent algorithm for different values of max $|A|$.**

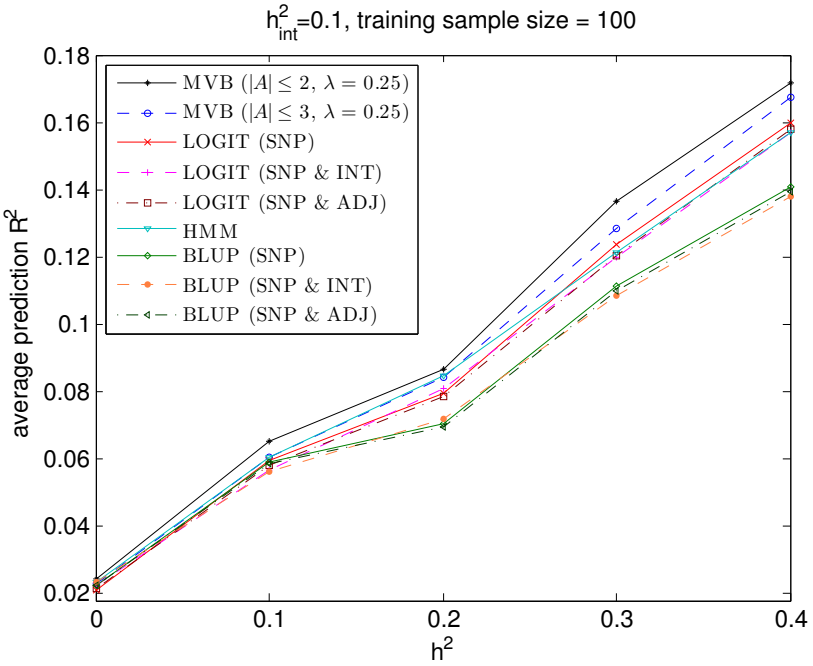


Figure 2.4: **Prediction R^2 across 100 validation individuals averaged over 200 regions for MVB, BLUP, LOGIT, and HMM as a function of h^2 when h^2_{int} is fixed at 0.1.**

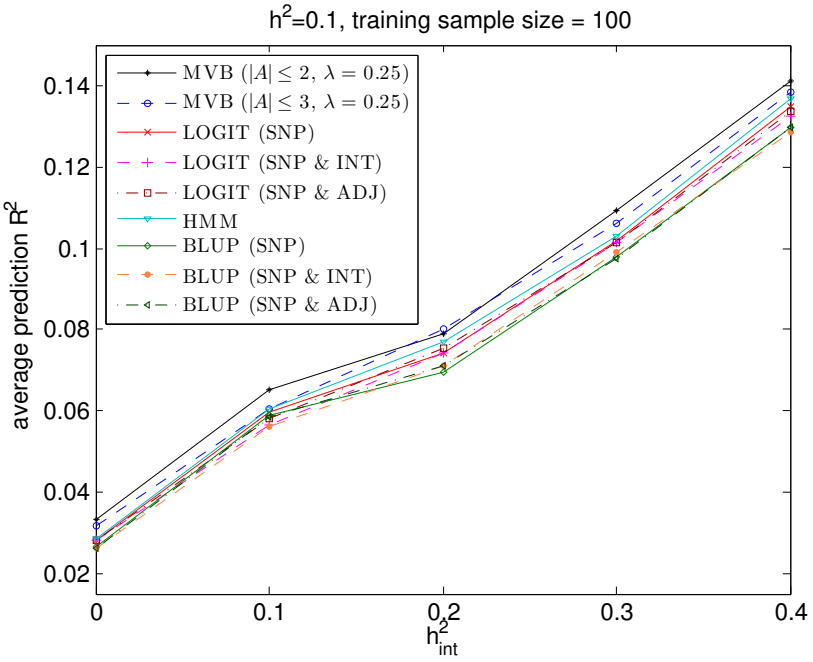


Figure 2.5: **Prediction R^2 across 100 validation individuals averaged over 200 regions for MVB, BLUP, LOGIT, and HMM as a function of h^2_{int} when h^2 is fixed at 0.1.**

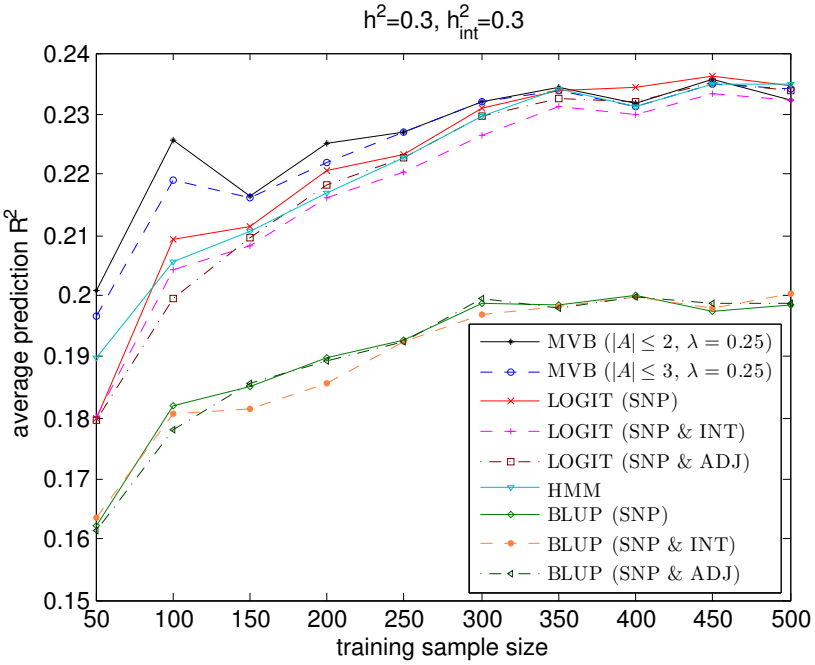
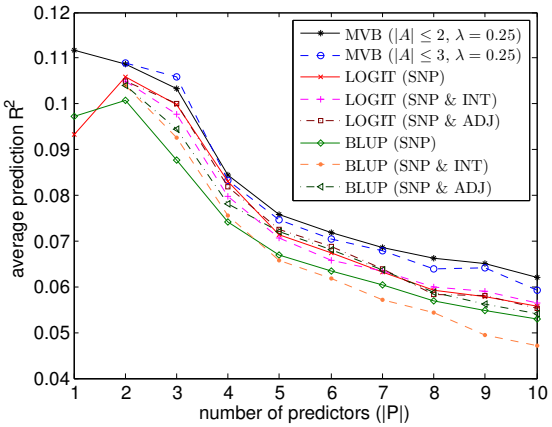
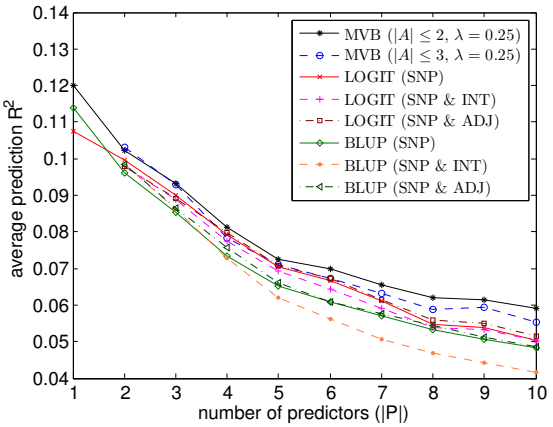


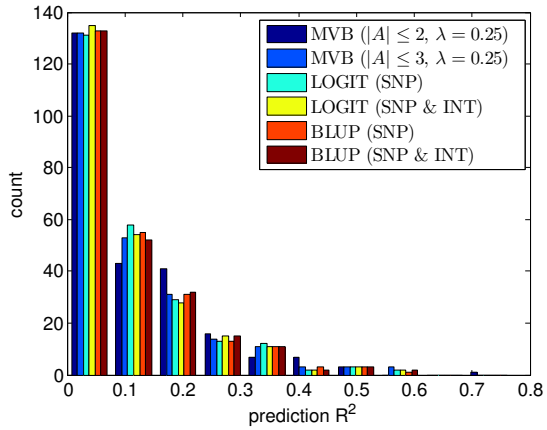
Figure 2.6: **Prediction R^2 across validation individuals averaged over 200 regions for the MVB, BLUP, LOGIT, and HMM as a function of training sample sizes when h^2 and h^2_{int} are both fixed at 0.3.**



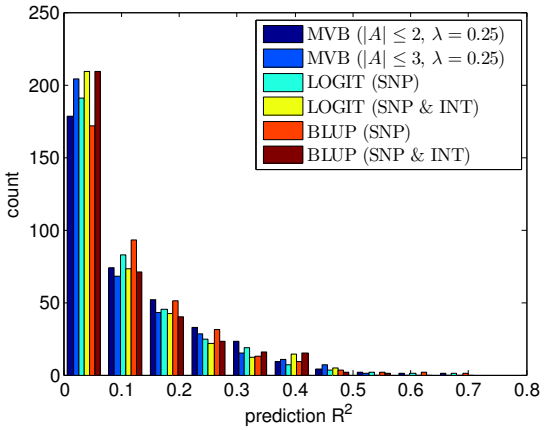
(a) Prediction R^2 averaged over 250 randomly selected windows



(b) Prediction R^2 averaged over 377 windows with dsQTLs



(c) Distribution of prediction R^2 for 250 randomly selected windows



(d) Distribution of prediction R^2 for 377 windows with dsQTLs

Figure 2.7: **Prediction R^2 for MVB, BLUP, and LOGIT.** Here "SNP" refers to the experiment involving only single SNPs, "SNP & INT" refers to the experiment involving both SNPs and all two-way interactions, and "SNP & ADJ" refers to the experiment involving both SNPs and only interactions between adjacent SNPs. Figure 2.7a and 2.7b show the average prediction R^2 over different windows as a function of the number of true predictors $|P|$. Figure 2.7c and 2.7d show the distribution of prediction R^2 for the highest average prediction R^2 over all $|P|$. For $|P| = 2$, the experiments "SNP & INT" and "SNP & ADJ" are identical.

CHAPTER 3

Contrasting the genetic architecture of 30 complex traits from summary association datas

3.1 Introduction

Large-scale genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) associated with hundreds of traits and diseases [90, 165, 29, 164]. However, only a fraction of the variance in trait can be explained by the risk SNPs reported by GWAS. The so-called "missing heritability problem" is in part due to the stringent significance threshold imposed in GWAS, which neglects variants of small effect that fail to reach the genome-wide significance level at current sample sizes. As an alternative, variance component (SNP-heritability) analysis aggregates the effect of all SNPs regardless of their significance [168] and has yielded important insights into the genetic architecture of complex traits [19, 44, 170, 91, 54, 117].

Heritability has been traditionally estimated using twins or pedigree [13] information with more recent works showing that SNP-based heritability (i.e. proportion of variance in trait explained by a given set of SNPs) can be estimated from unrelated individuals [170]. Standard approaches for SNP-heritability estimation rely on estimating the genetic relationships

This chapter is published in Shi et al., American Journal of Human Genetics 2016 [140]

between pairs of individuals (estimated genome-wide or across a subset of the genome) [170, 59, 53]. Therefore, these analyses require individual-level genotype data which prohibits their applicability to ultra-large GWAS that, due to privacy concerns, is typically available only at the summary level. To solve this bottleneck, recent methods have shown that SNP-heritability, both genome-wide as well as for different functional categories in the genome, can be accurately estimated using only summary GWAS data [19, 44]. Although these methods have enabled powerful analyses making insights into genetic basis of complex traits, they rely on the infinitesimal model assumption (i.e. all SNPs contribute to the trait) which is invalid at most risk loci [19, 44]. To overcome this drawback, alternative approaches have proposed to impose a prior on the sparsity of effect sizes to further increase SNP-heritability estimation accuracy [178]. A potentially more robust approach is to not assume any distribution for the effect sizes at causal variants and treat them as fixed effects in the estimation procedure. Indeed, recent works have shown that SNP-heritability estimation can be performed under maximum-likelihood from polygenic scores under a fixed-effect model assuming no LD among SNPs [117].

Here, we introduce Heritability Estimator from Summary Statistics (HESS), an approach to estimate the variance in trait explained by all typed SNPs at a single locus in the genome while accounting for linkage disequilibrium (LD) among SNPs. We build upon recent works[45, 117] that treat causal effect sizes as fixed effects and model the genotypes at the locus as random correlated variables. Our estimator can be viewed as a weighted summation of the squares of the projection of GWAS effect sizes onto the eigenvectors of the LD matrix at the considered locus, where the weights are inversely proportional to the corresponding eigenvalues. Through extensive simulations, we show that HESS is unbiased when in-sample LD is available regardless of disease architecture (i.e. number of causals and distribution of effect sizes). We extend our method to use LD estimated from reference panels [28] and show that a principal components based regularization of the LD matrix [57] yields approximately unbiased and more consistent estimates of local SNP-heritability as compared to existing methods [19].

We applied HESS to partition common SNP heritability at each locus in the genome using GWAS summary data for 30 traits spanning over 10 million SNPs and 2.4 million phenotype measurements. First, we show that common SNPs explain a large fraction of the total familial heritability estimated from twin studies, ranging anywhere from 20% to 90% across the studied quantitative traits. Second, we showcase the utility of local SNP-heritability estimates in finding loci that explain more variance in trait than the top associated SNP at the locus – an effect likely due to multiple signals of association. Third, we contrast the polygenicity of all 30 traits by comparing the fraction of total SNP-heritability attributable to loci with highest local SNP-heritability. We find that most of the 30 selected traits are highly polygenic with a small number of traits driven by a small number of loci. Finally, we report 36 "heritability hotspots" – regions of genome that attain a significant contribution to the SNP-heritability of multiple traits. Taken together, our results give insights into traits where further GWAS and/or fine-mapping studies are likely to recover a significant amount of the missing heritability.

3.2 Materials and methods

3.2.1 Overview of methods

We introduce estimators for the variance in trait explained by typed variants at a single locus (local SNP-heritability, $h^2_{g,local}$) from summary GWAS data (i.e. Z-scores, effect sizes and their standard errors). We derive our estimator under the assumption that effect sizes at typed variants are fixed and genotypes are drawn from a distribution with a pre-specified covariance structure. The covariance, (i.e. pairwise correlation between any variants at a locus, LD) can be estimated in-sample, from the genotype data in GWAS, or from external reference panels (e.g. 1000 Genomes Project[28]). Our estimator can be viewed as a weighted summation of the squared projections of GWAS effect sizes onto the eigenvectors of the LD matrix at the considered locus. The finite sample size of the GWAS studies as well as the reference panels used to estimate LD induces statistical noise that needs to be accounted

for to obtain an accurate estimation. Since the top projections make up the bulk of the summation, truncated-SVD lends itself as the appropriate regularization method to account for noise in the estimated LD matrices. Finally we extend our approach to consider multiple independent loci each contributing to the trait and show how our local estimator can be employed when the total genome-wide SNP-heritability is known (or estimated from other methods).

3.2.2 Estimating SNP-heritability at a single locus from GWAS summary data

Let $y_i = \mathbf{x_i}^T\boldsymbol{\beta} + \epsilon_i$, where y_i is the trait value for individual i, $\mathbf{x_i}$ are the standardized (i.e. 0 mean and unit variance) genotypes of individual i at p typed SNPs in the locus, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is the vector of fixed effect sizes for the p SNPs, and $\epsilon_i \sim N(0, \sigma_e^2)$ is the environmental effect. Assuming that $\boldsymbol{\beta}$ is fixed and $\boldsymbol{X}$ is random, the phenotypic variance is

$$\mathrm{V}ar[\boldsymbol{y}] = \mathrm{V}ar[\boldsymbol{X}\boldsymbol{\beta}] + \sigma_e^2 = \boldsymbol{\beta}^T \mathrm{C}ov[\boldsymbol{X}]\boldsymbol{\beta} + \sigma_e^2 = \boldsymbol{\beta}^T\boldsymbol{V}\boldsymbol{\beta} + \sigma_e^2 \tag{3.1}$$

where $\boldsymbol{V}$ is a $p \times p$ variance-covariance matrix of the genotype vector (i.e. the LD matrix). If we make a standard assumption that the phenotypes are standardized (i.e. $\mathrm{V}ar[\mathbf{y}] = 1$), it follows that the amount of variance contributed by the p SNPs to the trait (i.e. local SNP-heritability) is $h^2_{g,local} = \boldsymbol{\beta}^T\boldsymbol{V}\boldsymbol{\beta}$. If the true effect size vector $\boldsymbol{\beta}$ and the LD matrix $\boldsymbol{V}$ are given, then computing $h^2_{g,local}$ is trivial. In reality, however, the vector $\boldsymbol{\beta}$ is unknown and is estimated in GWAS involving n samples and p SNPs, where $\hat{\beta}_{gwas,i}$ is estimated as the marginal standardized regression coefficient for SNP i

$$\begin{aligned}\hat{\beta}_{gwas,i} &= \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{y} = \frac{1}{n}\boldsymbol{X}_i^T\left(\begin{bmatrix}\boldsymbol{X}_1 & \cdots & \boldsymbol{X}_p\end{bmatrix}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) \\ &= \begin{bmatrix}\frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{X}_1 & \cdots & \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{X}_p\end{bmatrix}\boldsymbol{\beta} + \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{\epsilon} = \sum_{j=1}^{p} r_{ij}\beta_j + \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{\epsilon}\end{aligned} \tag{3.2}$$

where $\boldsymbol{X}_i$ denotes standardized genotypes for SNP i across the n individuals, and r_{ij} denotes the LD between SNPs i and j. Extending to p SNPs at the locus, if follows that $\hat{\boldsymbol{\beta}}_{gwas} = \boldsymbol{V}\boldsymbol{\beta} +$

$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon}$ where $\boldsymbol{V}$ is the LD matrix. With $\boldsymbol{\beta}$ fixed and $\boldsymbol{\epsilon}$ random, $\hat{\boldsymbol{\beta}}_{gwas}$ is a random variable with $\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}] = \mathrm{E}[\boldsymbol{V}\boldsymbol{\beta} + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon}] = \boldsymbol{V}\boldsymbol{\beta}$, and $\mathrm{C}ov[\hat{\boldsymbol{\beta}}_{gwas}] = \mathrm{V}ar[\boldsymbol{V}\boldsymbol{\beta} + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon}] = \frac{1}{n^2}\boldsymbol{X}^T\,\mathrm{C}ov[\boldsymbol{\epsilon}]\boldsymbol{X} = \frac{\sigma_e^2}{n}\boldsymbol{V} = \frac{1-h^2_{g,local}}{n}\boldsymbol{V}$. By central limit theorem, $\hat{\boldsymbol{\beta}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\beta}, \frac{1-h^2_{g,local}}{n}\boldsymbol{V}\right)$.

As GWAS sample size (n) increases, $\hat{\boldsymbol{\beta}}_{gwas}$ converges to $\boldsymbol{\beta}_{gwas} = \boldsymbol{V}\boldsymbol{\beta}$. By simple substitution in Equation (3.1) it follows that an estimator for $h^2_{g,local}$ is

$$(\boldsymbol{\beta}^T_{gwas}\boldsymbol{V}^{-1})\boldsymbol{V}(\boldsymbol{V}^{-1}\boldsymbol{\beta}_{gwas}) = \boldsymbol{\beta}^T_{gwas}\boldsymbol{V}^{-1}\boldsymbol{\beta}_{gwas} \tag{3.3}$$

Unfortunately, the finite sample size of GWAS induces statistical noise in the estimation of $\boldsymbol{\beta}_{gwas}$ which leads to biased estimation if we simply replace $\boldsymbol{\beta}_{gwas}$ with $\hat{\boldsymbol{\beta}}_{gwas}$ above, as $\mathrm{E}[\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas}] = tr(\boldsymbol{V}^{-1}\,\mathrm{C}ov[\hat{\boldsymbol{\beta}}_{gwas}]) + \boldsymbol{\beta}^T\boldsymbol{V}\boldsymbol{\beta}$. However, we can correct for the bias $tr(\boldsymbol{V}^{-1}\,\mathrm{C}ov[\hat{\boldsymbol{\beta}}_{gwas}])$ as follows.

Let $\hat{h}^2_{g,local}$ be an unbiased estimator of $h^2_{g,local}$, then by definition $E[\hat{h}^2_{g,local}]$ must satisfy $E[\hat{h}^2_{g,local}] = h^2_{g,local}$. Then it follows that

$$\mathrm{E}[\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas}] = tr\left(\frac{1-h^2_{g,local}}{n}\boldsymbol{V}^{-1}\boldsymbol{V}\right) + h^2_{g,local} = \frac{1-\mathrm{E}[\hat{h}^2_{g,local}]}{n}p + \mathrm{E}[\hat{h}^2_{g,local}]. \tag{3.4}$$

A sufficient condition for Equation (3.4) to hold is $\frac{1-\hat{h}^2_{g,local}}{n}p + \hat{h}^2_{g,local} = \hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas}$. Solving for $\hat{h}^2_{g,local}$ gives an unbiased estimator for $h^2_{g,local}$

$$\hat{h}^2_{g,local} = \frac{n\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas} - p}{n-p}. \tag{3.5}$$

Following quadratic form theory [41], the variance of $\hat{h}^2_{g,local}$ is

$$\mathrm{V}ar[\hat{h}^2_{g,local}] = \left(\frac{n}{n-p}\right)^2\left(2p\left(\frac{1-h^2_{g,local}}{n}\right) + 4h^2_{g,local}\right)\left(\frac{1-h^2_{g,local}}{n}\right). \tag{3.6}$$

Since $h^2_{g,local}$, the true local SNP-heritability, is unknown, we use $\hat{h}^2_{g,local}$ instead. For $h^2_{g,local}$

near 0, $\mathrm{V}ar[\hat{h}^2_{g,local}] \approx \frac{4}{(n-p)^2} h^2_{g,local} + \frac{2p}{(n-p)^2}$ through Taylor expansion around 0. Thus, the plug in principle yields an estimation of $\mathrm{V}ar[\hat{h}^2_{g,local}]$ approximately equal to the truth in the expectation. For small $\hat{h}^2_{g,local}$ (as expected for most loci and traits) $\mathrm{V}ar[\hat{h}^2_{g,local}]$ is dominated by $\frac{2p}{(n-p)^2}$.

3.2.3 Accounting for rank deficiencies in the LD

In the above derivation we made the assumption that the inverse of the LD matrix $\boldsymbol{V}$ exists. In practice, however, due to pairs of SNPs in perfect LD, $\boldsymbol{V}$ is usually rank deficient, and thus $\boldsymbol{V}^{-1}$ does not exist. In such cases we use the Moore-Penrose pseudoinverse [9] $\boldsymbol{V}^{\dagger}$. Let $q = r\mathit{ank}(\boldsymbol{V})$, by rank decomposition, $\boldsymbol{V} = \boldsymbol{V}_A \boldsymbol{V}_B$, where $\boldsymbol{V}_A \in \boldsymbol{R}^{p\times q}$ and $\boldsymbol{V}_B \in \boldsymbol{R}^{q\times p}$ are matrices with full column rank and full row rank respectively, then $\mathrm{t}r(\boldsymbol{V}^{\dagger}\boldsymbol{V}) = \mathrm{t}r(\boldsymbol{V}_B^{\dagger}\boldsymbol{V}_A^{\dagger}\boldsymbol{V}_A\boldsymbol{V}_B) = \mathrm{t}r(\boldsymbol{V}_B\boldsymbol{V}_B^{\dagger}\boldsymbol{V}_A^{\dagger}\boldsymbol{V}_A) = \mathrm{t}r(\mathbf{I}_q) = q$. Accounting for rank-deficient LD matrix, we obtain an unbiased estimator, $\hat{h}^2_{g,local} = \frac{n\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}-q}{n-q}$. We make the same adjustment (replacing p with q) in the variance estimator for $\hat{h}^2_{g,local}$.

3.2.4 Adjusting for noise in external reference LD

When genotype data of GWAS samples is not available, we substitute the in-sample LD matrix $\boldsymbol{V}$ with external reference LD matrix $\hat{\boldsymbol{V}}$ estimated from the 1000 Genomes Project [28] using a population that matches the GWAS samples. However, due to limited sample size, external reference LD matrices contain statistical noise that biases our estimate. We apply truncated-SVD regularization to remove noise from external reference LD matrix as follows.

First note that $\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas} = \sum_{i=1}^{q} s_i = \sum_{i=1}^{q} \frac{1}{w_i}(\hat{\boldsymbol{\beta}}^T_{gwas}\boldsymbol{u}_i)^2$, where w_i and $\boldsymbol{u}_i$ are the eigenvalues and eigenvectors of the LD matrix $\boldsymbol{V}$, and $q = r\mathit{ank}(\boldsymbol{V})$. For external reference LD matrix $\hat{\boldsymbol{V}}$ with eigenvalues and eigenvectors $\hat{w}_i$ and $\hat{\boldsymbol{u}}_i$, the same decomposition holds except that s_i is replaced by $\hat{s}_i = \frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}^T_{gwas}\hat{\boldsymbol{u}}_i)^2$. In our previous works [122, 72], we propose

to regularize $\hat{\boldsymbol{V}}$ using ridge regression penalty. This regularization method is equivalent to replacing $\hat{w}_i$ with $\hat{w}_i + \lambda$, where λ is the ridge regression penalty. The ridge regression penalty shrinks the quadratic term $\hat{\boldsymbol{\beta}}_{gwas}\hat{\boldsymbol{V}}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}$ towards 0, which can lead to downward bias. We also notice that a large λ is needed to drive down the noise ($\hat{s}_i$ for large i), which diminishes the true signal at the same time. Here we show through simulations that most of the signal in $\hat{\boldsymbol{\beta}}_{gwas}^T\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}$ comes from s_i where $i \ll q$ and that $\hat{s}_i \approx s_i$ for $i \ll q$ (see Figure 3.1). These results motivate us to apply truncated-SVD to remove noise in $\hat{\boldsymbol{V}}$, i.e. we estimate $\hat{\boldsymbol{\beta}}_{gwas}^T\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\beta}}_{gwas}$ by $\sum_{i=1}^{k} 1/\hat{w}_i(\hat{\boldsymbol{\beta}}_{gwas}^T\hat{\boldsymbol{u}}_i)^2$, where $k \ll q$. Let $g(\hat{\boldsymbol{\beta}}_{gwas}, k) = \sum_{i=1}^{k}\frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}_{gwas}^T\hat{\boldsymbol{u}}_i)^2$, through eigen-decomposition of $\hat{\boldsymbol{V}}$, it can be shown that

$$\mathrm{E}[g(\hat{\boldsymbol{\beta}}_{gwas}, k)] = \frac{k(1-h^2_{g,local})}{n} + \sum_{i=1}^{k}\hat{w}_i(\hat{\mathbf{u}}_i^T\boldsymbol{\beta})^2. \tag{3.7}$$

Since the true local SNP-heritability is $h^2_{g,local} = \sum_{i=1}^{q} w_i(\boldsymbol{u}_i^T\boldsymbol{\beta})^2$, assuming $\hat{\boldsymbol{u}}_i = \boldsymbol{u}_i$ for $i \ll q$, Equation (3.7) is an approximation of $h^2_{g,local}$ with bias $\frac{k(1-h^2_{g,local})}{n}$. Correcting for this bias yields the estimator for the single-locus case

$$\tilde{h}^2_{g,local} = \frac{ng(\hat{\boldsymbol{\beta}}_{gwas}, k) - k}{n-k}. \tag{3.8}$$

In theory, the variance of $\tilde{h}^2_{g,local}$ is $\mathrm{V}ar[\tilde{h}^2_{g,local}] \approx \frac{4}{(n-k)^2}\hat{h}^2_{g,local} + \frac{2k}{(n-k)^2}$. In practice, however, this gives an underestimation of the truth. Thus, we replace k with $q = \mathrm{r}ank(\boldsymbol{V})$.

3.2.5 Extension to multiple independent loci

For genomes partitioned into m independent loci, the linear model for individual i's trait value becomes $y_i = \boldsymbol{x}_{i,1}^T\boldsymbol{\beta}_1 + \ldots + \boldsymbol{x}_{i,m}^T\boldsymbol{\beta}_m + \epsilon_i$ where $\boldsymbol{x}_{i,j}$ denotes the genotypes at the p_i SNPs in the i-th locus for individual i, and $\boldsymbol{\beta}_i$ denotes the effect sizes of SNPs in this locus. Based on the revised model, we decompose $\mathrm{V}ar[\mathbf{y}]$ into

$$\mathrm{V}ar[\mathbf{y}] = \mathrm{V}ar[\mathbf{X}_1\boldsymbol{\beta}_1] + \cdots + \mathrm{V}ar[\mathbf{X}_m\boldsymbol{\beta}_m] + \sigma_e^2 = h^2_{g,local,1} + \cdots + h^2_{g,local,m} + \sigma_e^2, \tag{3.9}$$

where $h^2_{g,local,i}$ denotes the local SNP-heritability contributed by the i-th locus. In the case of multiple independent loci, the noise term σ^2_e is equal to $1 - \sum_{j=1}^{m} h^2_{g,local,j}$. Thus, in order to correct for the bias generated by σ^2_e, one need to know $h^2_{g,local,j}$ for all j. Accounting for bias and adjusting for noise in external reference LD ($\hat{\boldsymbol{V}}_i$) following strategies outlined in previous sections, we arrive at the estimator,

$$\hat{h}^2_{g,local,i} = \frac{ng(\hat{\boldsymbol{\beta}}_{gwas,i}, k_i) - (1 - \sum_{j=1, j\neq i}^{m} \hat{h}^2_{g,local,j})k_i}{n - k_i}, \tag{3.10}$$

which defines a system of linear equations involving m variables ($\hat{h}^2_{g,local,i}$) and m equations. A similar system of linear equations can be solved to obtain the variance estimate,

$$\mathrm{V}ar[\hat{h}^2_{g,i}] = \left(\frac{n}{n-k_i}\right)^2 \left(2k_i\frac{\hat{\sigma}^2_e}{n} + 4\hat{h}^2_{g,local,i}\right)\frac{\hat{\sigma}^2_e}{n} + \left(\frac{k_i}{n-k_i}\right)^2 \sum_{j=1, j\neq i}^{m} \mathrm{V}ar[\hat{h}^2_{g,local,j}], \tag{3.11}$$

where $\hat{\sigma}^2_e = 1 - \sum_{j=1}^{m} \hat{h}^2_{g,local,j}$.

In the special case when $k_1 = \cdots = k_m = k$ (i.e. all loci use the same number of eigenvectors in the truncated-SVD regularization of LD matrices), Equation (3.10) simplifies as follows: $\hat{h}^2_g = \sum_{i=1}^{m} \hat{h}^2_{g,local,i} = \sum_{i=1}^{m} \frac{ng(\hat{\boldsymbol{\beta}}_{gwas,i},k) - (1-\hat{h}^2_g + \hat{h}^2_{g,local,i})k}{n-k} = \frac{n}{n-k}\sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, k) - \frac{k}{n-k}(m - m\hat{h}^2_g + \hat{h}^2_g)$, yielding the following estimate for the total genome-wide SNP-heritability:

$$\hat{h}^2_g = \frac{n}{n-mk}\sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, k) - \frac{mk}{n-mk}, \tag{3.12}$$

with variance:

$$\mathrm{V}ar[\hat{h}^2_g] = \left(\frac{n}{n-mk}\right)^2 \sum_{i=1}^{m} \mathrm{V}ar[g(\hat{\boldsymbol{\beta}}_{gwas,i}, k)] \approx \left(\frac{n}{n-mk}\right)^2 \frac{2mk}{(n-k)^2}. \tag{3.13}$$

Thus, if k is chosen such that $n - mk$ is small (i.e. $\frac{n}{n-mk}$ large) the genome-wide SNP-heritability estimates becomes unstable with large variance. To ensure stable estimates and reduce variance (at the cost of some bias) we recommend choosing k such that $\frac{n}{n-mk}$ is less than 2 when using our estimator for genome-wide estimation.

3.2.6 Known genome-wide SNP-heritability

In many cases, the total genome-wide SNP-heritability estimate (h_g^2) and its variance ($\mathrm{V}ar[h_g^2]$) of a trait are known (e.g. estimated from individual-level data). In those cases, one can simply plug h_g^2 into Equation (3.10) to obtain local estimates of heritability $h_{g,local,i}^2$:

$$\hat{h}_{g,local,i}^2 = g(\hat{\boldsymbol{\beta}}_{gwas,i}, k) - \frac{k}{n}(1 - h_g^2), \tag{3.14}$$

from which we conclude

$$\mathrm{V}ar[\hat{h}_{g,local,i}^2] = \mathrm{V}ar[g(\hat{\boldsymbol{\beta}}_{gwas,i}, k)] + \left(\frac{k}{n}\right)^2 \mathrm{V}ar[h_g^2]. \tag{3.15}$$

In general, the sum of local SNP-heritability $\hat{h}_g^2 = \sum_{i=1}^m \hat{h}_{g,local,i}^2$ is not necessarily equal to h_g^2 due to variance in $\hat{h}_{g,local,i}^2$. Since $\mathrm{V}ar[\hat{h}_g^2] = \mathrm{V}ar[\sum_{i=1}^m \hat{h}_{g,local,i}^2] \approx \frac{2mk}{(n-k)^2} + \left(\frac{mk}{n}\right)^2 \mathrm{V}ar[h_g^2]$, we recommend choosing k such that $\frac{mk}{n}$ is less than 0.5 to ensure stable estimate and reduce variance. We assessed the local SNP-heritability estimation with or without known genome-wide SNP-heritability using the height GWAS data (see Table 1) with a previously reported h_g^2=0.50[165]. The local SNP-heritability estimates were virtually indistinguishable between the two approaches ($R = 1.0$).

3.2.7 Simulation framework

We use HAPGEN2 [150] to simulate genotypes for 50,000 individuals starting from half of the 505 European (EUR) individuals in the 1000 Genomes Project [28] for SNPs with minor allele frequency (MAF) greater than 5% in randomly selected regions spanning 0.75 Mb to 1.5Mb on chromosome 1. We reserve the other half of the EUR individuals as external reference panel. From the simulated genotypes of the 50,000 individuals, we then simulate phenotypes based on the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is the standardized genotype matrix with mean 0 and variance 1 at each column.

We investigated the performance of our method under a wide-range of simulations. We first select a subset C of $|C|$ causal SNPs at random and then simulate the effect sizes at these SNPs as $\boldsymbol{\beta}_C \sim N(\mathbf{0}, \frac{h^2}{|C|}\mathbf{I}_{|C|})$, where h^2 is the heritability to be simulated. We draw $\boldsymbol{\epsilon}$ from $N(\mathbf{0}, (1-h^2)\boldsymbol{I}_n)$ such that $\mathrm{E}[\boldsymbol{y}] = 0$, $\mathrm{V}ar[\boldsymbol{y}] = 1$, and that the SNP-heritability for this locus is h^2. For fixed $\boldsymbol{\beta}$, we then generate replications of trait values $\boldsymbol{y}$ by re-drawing $\boldsymbol{\epsilon}$. Finally, we compute summary statistics, $\hat{\boldsymbol{\beta}}_{gwas}$, following procedures outlined in previous sections. We simulate 500 set of summary statistics for each simulation scenario. Although within each of the 500 set of simulated summary statistics, C and $\boldsymbol{\beta}$ are fixed, they vary across different set of simulations.

We also investigated simulations where $\boldsymbol{\beta}$ varies across simulated individuals. In each of the 500 set of simulated GWAS summary statistics, we first select a subset C of $|C|$ causal SNPs at random. Then, for each individual, we draw $\boldsymbol{\beta}_{C,i}$ from $N(0, \boldsymbol{\alpha}_i h^2)$ for $i = 1, \cdots, |C|$, where $\boldsymbol{\alpha}$ governs the proportion of heritability contributed by each SNP and satisfies $\sum_{i=1}^{|C|} \boldsymbol{\alpha}_i = 1$. In the special case when $\boldsymbol{\alpha}_i = \frac{1}{|C|}$ for all i, each causal SNP contributes the same proportion of heritability. Here, C and $\boldsymbol{\alpha}$ are fixed in each set of simulation but vary across the 500 set of simulations.

Since in simulations, we assume that all SNPs are typed and that environmental effect ($\boldsymbol{\epsilon}$) is drawn independently for each individual, cryptic relatedness among individuals in the 1000 Genomes Project [28] will have minimal effect on our estimates.

3.2.8 Empirical data sets

We obtained publicly available GWAS summary over European ancestry data for 30 traits from 11 GWAS consortia (see Table 3.3). For quality control, we restricted our analysis to GWAS studies involving at least 20,000 samples, and excluded sex chromosomes. We used the definition of independent loci as defined in [11] (1.6 Mb on the average). To reduce statistical noise in LD matrix, we focused on estimating heritability attributable to common SNPs (i.e. SNPs with MAF greater than 5% in the European 1000 Genomes data[28]). Prior

to estimating heritability, we also removed SNPs with ambiguous alleles as compared to the reference panel (Table 3.3) and applied our estimator as defined in Equation (3.10). For each trait, we choose k, the number of eigenvectors used to estimate local heritability across all loci, based on sample size of the GWAS (see Methods) – a large k is used for GWAS with large sample size, and a small k is used for GWAS with small sample size. To avoid inflation due to noise in LD, we cap k at a maximum of 50 (see Table 3.4). To ensure stable estimates, we also recommend filtering out eigenvectors with corresponding eigenvalues less than 1.

Most GWAS apply genomic control (GC) factor (λ_{gc}) to χ^2 statistics to correct for inflation due to population structure [158] and publish GC-corrected effect size estimation ($\hat{\boldsymbol{\beta}}_{gwas,gc}$). And we note that all the summary GWAS data we analyze in this work were adjusted for population structure to various degrees, and had at least one round of genomic correction. However, recent works [19, 171] show that λ_{gc} can not distinguish between inflation and true polygenicity and overestimates the correction factor needed for population stratification. Although dividing the χ^2 statistics by λ_{gc} has little effect on computing the ratios between local and genome-wide heritability [44], it can result in underestimation of both local and genome-wide SNP-heritability – when applied on GC-corrected summary data directly, our method can produce negative and uninformative local and total SNP-heritability estimates. To account for this, we first estimate λ_{gc} from summary GWAS data and re-inflate the effect sizes ($\hat{\boldsymbol{\beta}}_{gwas,gc}$) with estimated $\sqrt{\lambda_{gc}}$ before obtaining local SNP-heritability estimates. We estimate λ_{gc} based on the observation that at a locus with no heritability (i.e. $h^2_{g,local,i} = 0$), $\mathrm{E}[\hat{\boldsymbol{\beta}}^T_{gwas,gc,i}\boldsymbol{V}_i^{\dagger}\hat{\boldsymbol{\beta}}_{gwas,gc,i}] = \frac{1}{\lambda_{gc}}\frac{q_i}{n}$, where $\hat{\boldsymbol{\beta}}_{gwas,gc,i} = \frac{\hat{\boldsymbol{\beta}}_{gwas,i}}{\sqrt{\lambda_{gc}}}$ denotes GC-corrected effect size vector, and that $\mathrm{E}[\hat{\boldsymbol{\beta}}^T_{gwas,i}\boldsymbol{V}_i^{\dagger}\hat{\boldsymbol{\beta}}_{gwas,i}] = \frac{q_i}{n}$, where $\hat{\boldsymbol{\beta}}_{gwas,i}$ is the vector of effect size estimation without GC correction. To estimate λ_{gc}, we treat the bottom 50% of all loci with the smallest estimated local SNP-heritability as loci having $h^2_{g,local,i} = 0$, and regress the vector $(\frac{q_i}{n})$ against the vector $(\hat{\boldsymbol{\beta}}^T_{gwas,gc,i}\boldsymbol{V}_i^{\dagger}\hat{\boldsymbol{\beta}}_{gwas,gc,i})$. We note that using the bottom 50% of all loci is a conservative measure to account for ascertainment in choosing loci and can result in estimated λ_{gc} less than 1. In practice, we only re-inflate $\hat{\boldsymbol{\beta}}_{gwas,gc}$ if the estimated λ_{gc} is

greater than 1. We report estimated λ_{gc} for all 30 traits in Table S1. Overall, our estimated λ_{gc} is consistent with the reported λ_{gc}. For example, our estimated λ_{gc} for BMI (1.33), HDL (1.13), LDL (1.16), TC (1.16), and TG (1.18) are consistent with the reported λ_{gc} for BMI (1.38) [90] and lipid traits (1.10-1.15) [29].

We define GWAS hits as SNPs with p-values less than 5×10^{-8}. To avoid overestimation due to LD tagging, for each locus, we only select the most significant (i.e. smallest p-value) GWAS hit as the index SNP. Heritability attributable to index SNPs, $\hat{h}^2_{gwas}$, is then estimated as $\sum_{i=1}^{I} \hat{\beta}_i^2$, where $\hat{\beta}_i$ is effect size of the i-th index SNP, and I the number of index SNPs. We estimate the variance of $\hat{h}^2_{gwas}$ as $\mathrm{V}ar[\hat{h}^2_{gwas}] = \sum_{i=1}^{I} \mathrm{V}ar[\hat{\beta}_i^2] = \sum_{i=1}^{I} \mathrm{V}ar[(Z_i/\sqrt{n})^2] = \sum_{i=1}^{I} \mathrm{V}ar[\frac{1}{n}\chi_i^2] = 2I/n^2$.

For case-control traits, an adjustment factor is needed to correct for ascertainment [82]. We note that this adjustment factor is derived based on the infinitesimal model, and does not apply to our method, which assumes a fixed effect model. Therefore, we only report unadjusted heritability estimates for case-control traits. However, we note that ratio between local to genome-wide SNP-heritability is not affected by this scaling factor.

3.3 Results

3.3.1 Performance of HESS in simulations

We used simulations to assess the performance of our proposed approach under a variety of disease architectures. First, we confirmed that by accounting for rank deficiency in the LD matrix we obtain unbiased estimation whereas the approach that uses the number of SNPs to correct for bias generated by the quadratic form [45] leads to a severe under-estimation of heritability. Second, we find that using the top 10-50 eigenvectors of the LD matrix (see Methods) provides a good approximation for the estimated heritability when LD is estimated from external reference panels (Figure 3.1).

Since we use approximately independent loci [11], we also assessed potential bias due to cross-tagging of heritability resulting from LD across adjacent loci. We simulated summary statistics based on 10,000 randomly selected SNPs spread across the entire chromosome 22, with 20% of these SNPs being causal and total SNP-heritability varying from 2% to 10%. For each simulation scenario, we simulate 500 set of summary statistics, and obtain local SNP-heritability estimates using equation (3.10). We obtain total SNP-heritability estimate by summing all local SNP-heritability estimates. We find that using the top $k = 30$ eigenvectors in the truncated-SVD regularization of LD matrices, HESS yields downwardly biased estimate of total SNP-heritability estimate, whereas at $k = 50$ HESS is approximately unbiased (Figure 3.2). Therefore, we use $k = 50$ as the default unless otherwise noted.

Next, we compared HESS to the recently proposed LD-score regression (LDSC)[19, 44] method that provides estimates of heritability from GWAS summary data. Although LDSC is not designed for local analyses due to model assumptions on polygenicity, it is able to estimate the variance in trait attributable to any sets of SNPs. As expected, in our simulations, where all individuals share the same effect size vector ($\boldsymbol{\beta}$), we find that LDSC is sensitive to the underlying polygenicity and, in general, yields biased estimation of heritability. In contrast, HESS provides an unbiased estimation of heritability across all simulated disease architectures when in-sample LD is available. For example, in simulations where 20% of the variants at the locus are causal explaining 0.05% heritability, HESS yields an estimate of 0.054% (s.e. 0.004%) as compared to 0.025% (s.e. 0.0009%) for LDSC (Figure 3.3). We attribute this to the fact that HESS does not make any assumption on the distribution of effect sizes at causal variants by treating them as fixed effects in the model. When LD from the sample is unavailable and has to be estimated from reference panels, both methods are biased with HESS (with $k = 30, 50$ eigenvectors in the truncated-SVD regularization of the LD matrix) yielding results closer to simulated heritability than LDSC at randomly selected loci with different width (Figure 3.3). Similar results were obtained in simulations where the $\boldsymbol{\beta}$ is drawn independently for each individual. This is expected because conditional on a fixed $\boldsymbol{\beta}$, HESS is unbiased (i.e. $\mathrm{E}[\hat{h}_g^2|\boldsymbol{\beta}] = h_g^2$), then the expectation of HESS estimate integrating over all possible $\boldsymbol{\beta}$ is still unbiased (i.e. $\mathrm{E}[\hat{h}_g^2] = \mathrm{E}[\mathrm{E}[\hat{h}_g^2|\boldsymbol{\beta}]] = \mathrm{E}[h_g^2] = h_g^2$).

Finally, unlike LDSC that employs a jack-knife approach to estimate variance in the estimated heritability (thus requiring multiple loci), HESS provides a variance estimator following quadratic form theory (see Methods). Since external reference LD is typically computed based on much smaller samples than in-sample LD, subtle patterns in in-sample LD cannot be captured by external reference LD. Thus, external reference LD matrices usually have lower rank than their corresponding in-sample LD matrices, resulting in under-estimation of $\mathrm{V}ar[\hat{h}^2_{g,local,i}]$ (see Equation (3.11)). We verify this in simulations and find that the variance estimator yields unbiased estimates when in-sample LD is available and under-estimates theoretical variance when external reference LD is used. We also note that cryptic relatedness in GWAS samples can drive down the effective sample size (n), thus our estimates of standard errors could be deflated for GWAS where the effective sample size is significantly smaller than the actual sample size.

3.3.2 Common variants explain a large fraction of heritability

Having demonstrated the utility of HESS in simulations, we next applied our method to empirical GWAS summary data across 30 complex traits and diseases spanning more than two million phenotypic measurements (see Methods, Table 3.3, Table S1). We estimated the local SNP-heritability at 1,703 approximately-independent loci [11] using European individuals of the 1000 Genomes to estimate LD [28]. We first investigated the total contribution of common variants (MAF $> 5\%$) to the heritability of complex traits. We summed up the local estimates provided by our method to obtain an estimate for the total genome-wide heritability for all genotyped SNPs. For traits where the SNP-heritability was previously reported we find a broad consistency between our estimate and the existing estimates from the literature (see Table 3.3). For example, HESS estimates a genome-wide SNP heritability (h^2_g) of 16.5% (s.e. 0.5%) for BMI and 59.4% (s.e. 0.3%) for height as compared to previously reported estimates of 21.6% (2.2%) for BMI [90] and 62.5% for height [165]. We also find that our total SNP-heritability estimates broadly correlates with those obtained by LDSC ($R = 0.78$). Most importantly, we find that common SNPs explain a large fraction of the

previously reported familial heritability for all quantitative traits we interrogated ranging from 21% for Forearm BMD to 94% for HDL(Table 3.3). Although we observe a very high contribution of common SNPs to case-control traits as well, we note that our estimator can be biased due to ascertainment in this case (see Methods).

3.3.3 Hidden heritability at known risk loci

Recent works[54, 96] have shown that the total heritability explained by all variants at the GWAS risk loci ($h^2_{g,local,gwas}$) is higher than heritability explained by GWAS index SNPs (h^2_{gwas}). This suggests that a fraction of the missing heritability is due to multiple causal variants or poor tagging of hidden causal variants at known risk loci. We used HESS to quantify the gap between these two estimates of heritability at known risk loci. We find several traits for which $h^2_{g,local,gwas}$ is significantly larger than h^2_{gwas}. For example, $h^2_{g,local,gwas}$ is over two fold higher (32.0%, s.e. 0.2%) than h^2_{gwas} (13.9%, s.e. 0.002%) for height (Table 3.3). The difference can be accounted by incomplete tagging of hidden causal variant(s) or allelic heterogeneity (i.e., multiple causal variants). Indeed, conditional analysis identified 36 GWAS loci that contain multiple signals of associations (for a total of 87 GWAS risk SNPs at these loci) for height[169]. Restricting to the 28 loci that contain at least 2 of the 87 GWAS risk SNPs, we estimate $h^2_{g,local,gwas}$=4.6%(s.e. 0.06%), a 2.4-fold increase over h^2_{gwas}=1.9% (s.e. 0.003%). These loci, 5.8% of all GWAS loci for height, contribute to 14.2% of the difference between $h^2_{g,local,gwas}$ and h^2_{gwas} across all loci, thus suggesting that the difference is likely due to multiple signals of association. To confirm this hypothesis we applied a conditional analysis from summary GWAS data using GCTA-COJO [169] for the traits HDL, TG, RA, and SCZ. We observe that a moderate fraction (2% – 16%) of GWAS loci show multiple signals of association (see Table 3.2) thus confirming that contrasting $h^2_{g,local,gwas}$ with h^2_{gwas} is indicative of multiple signals of association.

In contrast, the majority of traits show similar $\hat{h}^2_{g,local,gwas}$ and $\hat{h}^2_{gwas}$ (see Table 1) suggesting a single causal variant at these loci very well tagged by the index GWAS variant. For example, it is known that LDL is strongly regulated by a single non-coding functional variant at the

SORT1 locus [29, 110] and that bone mineral density trait (FN) is strongly regulated by WNT16 [177, 75]. We also observe traits (e.g. MCH, MCV, RBC) for which $\hat{h}^2_{g,local,gwas}$ is estimated to be less than $\hat{h}^2_{gwas}$. This seemingly contradictory result is due to the fact that fewer eigenvectors in the truncated-SVD regularization of LD matrices were used to estimate $\hat{h}^2_{g,local,gwas}$ for GWAS with small sample sizes (see Table S2), resulting in downward bias (see Methods).

3.3.4 Contrasting polygenicity across multiple complex traits

Most studied common traits exhibit a strong polygenic architecture (i.e. an abundance of loci of small effect contributing to trait)[91, 90, 165, 29] . We recapitulate this observation using the HESS analysis and find a strong correlation between chromosome length and the fraction of heritability it explains for most traits we analyze here (Figures 3.4, 3.5). We also observe, consistent with previous findings [25], regions such as FTO on chromosome 16 and HLA on chromosome 6 contributing disproportionately to the fraction of heritability for HDL, BMI, and RA, respectively.

Next, we sought to quantify the variability in polygenicity across traits. We rank order loci based on their estimated local SNP-heritability, sum their contribution and plot it versus the percentage of genome they occupy (Figure 3.6). For highly polygenic traits, we expect the cumulative fraction of total SNP-heritability to be proportional to the fraction of genome covered, whereas for less polygenic traits, we expect to see a small fraction of the genome accounting for a large fraction of total SNP heritability. For example, in schizophrenia and height the top 1% of the loci with the highest local SNP-heritability contribute to 4.2%(s.e. 1.0%) and 6.5%(s.e. 1.5%) of the total SNP-heritability of these traits, respectively. This is consistent with previous reports on the degree of polygenicity of these traits [91, 165, 29]. At the other extremes, RA and lipid traits (HDL, LDL, TC, TG) have a lower degree of polygenicity, with the top 1% of loci accounting for 14-30% of the total SNP heritability. However, the low polygenicity of RA is mostly driven by the HLA region on chromosome 6. After removing local SNP-heritability estimates at loci overlapping the HLA region for

all traits, we observe that RA shows a moderate degree of polygenicity for the rest of the genome. We also note that the different degrees of polygenic signals across traits reflect both a difference in disease architecture (i.e. distribution of effect sizes) as well as a difference in the sample sizes for the GWAS summary data.

A different perspective of polygenicity is to restrict to GWAS risk loci (as they clearly contain risk variants) and contrast the proportion of explained variance with the proportion of the genome they occupy. We observe a wide distribution across traits reflecting diverse genetic architectures as well as different sample sizes for the GWAS performed for these traits. For example, approximately 30% of loci across the genome harbor a risk variant for height and account for 50% to the total SNP-heritability (a 1.5-fold enrichment). On the other hand, while only 5% of the loci contain GWAS risk variants for HDL, these loci collectively explain 25% of the SNP-heritability of HDL (a 4.6-fold enrichment) (Figure 3.7).

3.3.5 Loci that contribute to heritability of multiple traits

It has been previously established that a number of the 30 traits investigated in this study share a genetic basis [18]. Correlating local SNP-heritability estimates across the entire genome can serve as a proxy for the magnitude of pleiotropy and we can identify pairs of traits whose genetic components tend localize within the same regions of the genome. Motivated by this, we searched for specific pleiotropic loci which we define as loci that contribute significant non-zero SNP-heritability (one-tailed p-value < 0.05, Bonferroni corrected for 1,703 loci) for at least 3 out of the 30 analyzed traits. In total, we identified 36 such loci distributed genome-wide (see Figure 3.9).

As expected, the HLA region (chr6:26-34M), displays strong pleiotropic signal, particularly for immunologically relevant phenotypes (see Figure 3.9). For instance, the locus chr6:32-33M contributes significant amount of SNP-heritability for 8 traits, with exceptionally strong signals for RA, UC, and IBD (see Figure 3.9). We also observe several other pleiotropic loci, including chr2:199M-202M, contributing to AM, SCZ, and Height; chr6:134-136M, contribut-

ing to multiple red blood cell traits; and chr19:45-46M, contributing to multiple lipid traits. It's well known that there exist genetic correlations among red blood cell traits [160, 46, 23] as well as among lipid traits [29, 18]. Interestingly, previous research has also revealed that early age at menarche is associated with later onset of schizophrenia [26]. Our results suggest that these genetic correlations and associations may be caused in part by the pleiotropic effect of these loci.

We note that the selection of traits can bias the identification of pleitropic loci towards over-represented traits such as height and lipid traits. Nevertheless, local SNP-heritability analysis is still a useful tool to quantify the fraction of total SNP heritability contributed by a single loci and provide valuable insights into identifying pleiotropic loci.

3.4 Discussion

We have presented HESS, an unbiased estimator of local SNP-heritability from GWAS summary data. We extend existing work [45] that estimate heritability under the fixed-effect model by proposing to regularize external reference LD matrix via truncated-SVD and generalizing the estimator to multiple independent loci. Through extensive simulations, we demonstrate that HESS is unbiased given in-sample LD and yields more consistent and less biased local SNP-heritability estimates than LDSC given external reference LD. We applied HESS on GWAS summary data of 30 complex traits from 12 GWAS consortia and showed that our results recapitulate previous findings. We then used these local SNP-heritability estimate to contrast polygenicity of complex traits, find loci with multiple causal variants, and identify heritability hot spots. We note that enrichment of heritability at GWAS risk loci could be leveraged into prioritizing GWAS or fine-mapping; for example, traits with small enrichment of heritability at GWAS risk loci are more suitable for larger GWAS, whereas traits with large enrichment of heritability at known risk loci could be investigated further through fine-mapping.

In this work, we focus on estimating local heritability attributable to common autosomal

variants (MAF $> 5\%$), ignoring potential heritability captured by the sex chromosomes and rare variants. We also note that our heritability estimates for case-control traits are not adjusted for ascertainment as it is unclear whether adjustment derived based on the infinitesimal model can be directly applied for the fixed-effect model. Thus, our reported heritability estimation for case-control traits can be biased due to ascertainment. Future work that addresses local heritability estimation including both common and rare variants, sex chromosomes, as well as adjustment of heritability estimates under the fixed-effect model for case-control traits will further improve the utility of our approach.

We conclude with several caveats and limitations of our work. First, our method relies on independent LD blocks, which are often hard to define due to LD across multiple loci. In this work, we attempt to minimize LD leakage by defining approximately independent loci using principled approaches. Second, when only external reference LD is available, our method can yield biased heritability estimate as well as its variance estimate, due to external reference LD having lower rank than its corresponding in-sample LD as well as cryptic relatedness in GWAS samples. This makes precise hypothesis testing difficult. However, with in-sample LD and larger reference panels such as the Haplotype Reference Consortium [103], this bias will be reduced as LD can be inferred more precisely. We also note that our estimated λ_{gc} can be a potential source of bias, thus our genome-wide estimate should be interpreted with caution. Third, to obtain stable estimate, the number of eigenvectors used (k) in the truncated-SVD regularization should be chosen based on the sample size of GWAS study – GWAS with large sample size can afford large k, whereas GWAS with small sample size should use a small k. We recommend applying our method on summary data obtained from GWAS studies involving around or above 50,000 samples. For GWAS with small sample size, when genome-wide SNP-heritability is known, one can still apply Equation (3.14) to obtain stable local heritability estimate. We also note that although using the same number of eigenvectors for all loci facilitates the study of the statistical properties of our estimator, this approach may not be optimal for all loci. We conjecture that selecting k using more principled approach (e.g. based on the distribution of eigenvalues) may reduce bias, and we leave such investigation as future work.

3.5 Tables

Trait	h^2_g	h^2_{pub}	h^2_g/h^2_{pub}	h^2_{gwas}	$h^2_{g,local,gwas}$	$h^{2*}_{g,local,gwas}$	**Enrichment**[a]
BMI (Body Mass Index) [90]	16.5(0.5)	42 [61]	0.39	1.6(0.001)	3.1(0.1)	3.1(0.1)	3.7(0.4)
Height (Height) [165]	59.4(0.3)	69 [61]	0.86	13.9(0.002)	32.0(0.2)	24.0(0.2)	1.5(0.1)
HB (Haemoglobin) [160]	17.9(2.1)	37 [47]	0.48	2.2(0.003)	1.9(0.3)	1.8(0.3)	7.6(1.4)
MCH (Mean Cell Haemoglobin) [160]	29.3(2.2)	52 [88]	0.56	7.2(0.003)	6.2(0.4)	6.1(0.4)	9.9(1.9)
MCHC (MCH Concentration) [160]	10.9(2.5)	48 [62]	0.23	0.4(0.003)	0.5(0.2)	0.5(0.2)	6.7(1.8)
MCV (Mean Cell Volume) [160]	26.3(2.0)	52 [88]	0.51	6.5(0.004)	5.7(0.4)	5.6(0.4)	8.1(1.3)
PCV (Packed Cell Volume) [160]	16.7(2.5)	30 [47]	0.56	1.4(0.003)	0.9(0.2)	0.8(0.2)	6.0(1.4)
RBC (Red Blood Cell Count) [160]	22.0(2.3)	56 [88]	0.39	3.6(0.004)	2.6(0.3)	2.6(0.3)	6.4(1.6)
PLT (Number of Platelets) [52]	27.5(1.5)	57 [47]	0.48	3.5(0.003)	3.9(0.3)	3.9(0.3)	5.7(0.9)
FG (Fasting Glucose) [38]	22.3(2.3)	66 [145]	0.34	2.6(0.002)	1.7(0.2)	1.6(0.2)	8.0(2.5)
FI (Fasting Insulin) [38]	19.9(2.4)	36 [133]	0.55	–	–	–	–
HBA1C (HBA1C) [147]	20.8(2.3)	75 [145]	0.28	1.8(0.003)	0.9(0.2)	0.9(0.2)	6.6(1.9)
HOMA-B (HOMA-B) [38]	20.3(2.4)	72 [106]	0.28	0.6(0.001)	0.4(0.1)	0.4(0.1)	7.5(1.9)
HOMA-IR (HOMA-IR) [38]	19.9(2.4)	38 [133]	0.52	–	–	–	–
HDL (High Density Lipoprotein) [29]	39.4(0.9)	42 [174]	0.94	5.8(0.002)	10.7(0.2)	10.5(0.2)	4.6(1.3)
LDL (Low Density Lipoprotein) [29]	33.0(1.0)	40 [174]	0.82	7.8(0.002)	8.4(0.2)	8.3(0.2)	5.1(0.9)
TC (Total Cholesterol) [29]	35.5(0.9)	50 [35]	0.71	8.0(0.002)	9.3(0.2)	9.3(0.2)	4.3(0.6)
TG (Triglycerides) [29]	34.8(0.9)	40 [40]	0.87	5.2(0.002)	8.0(0.2)	8.0(0.2)	5.8(1.4)
EY (Education Years) [134]	19.9(0.8)	40 [134]	0.50	0.1(0.002)	0.2(0.0)	0.2(0.0)	3.2(1.4)
FA (Forearm BMD) [176]	17.4(2.2)	84 [3]	0.21	0.3(0.001)	0.5(0.1)	0.5(0.1)	22.4(7.7)
FN (Femoral Neck BMD) [176]	24.1(2.1)	84 [3]	0.29	2.0(0.003)	2.0(0.2)	2.0(0.2)	7.1(1.0)
LS (Lumbar Spine) [176]	25.1(2.0)	84 [3]	0.30	2.2(0.003)	2.2(0.3)	2.2(0.3)	6.1(0.8)
AM (Age at Menarche) [124]	27.8(0.7)	49 [156]	0.57	2.6(0.002)	3.8(0.2)	3.7(0.2)	2.9(0.2)
COL (College) [134]	19.4(0.8)	40 [134]	0.48	0.1(0.001)	0.1(0.0)	0.1(0.0)	3.5(0.9)
RA (Rheumatoid Arthritis) [114]	66.3(0.9)	55 [58]	1.21	11.2(0.003)	22.0(0.3)	22.1(0.3)	9.8(4.3)
SCZ (Schizophrenia) [113]	64.5(0.7)	81 [152]	0.80	6.2(0.004)	9.2(0.2)	9.2(0.2)	2.3(0.1)
CD (Crohn's Disease) [89]	35.9(1.8)	53 [159]	0.68	3.8(0.002)	5.9(0.4)	5.9(0.4)	4.8(0.7)
IBD[c] (Inflammatory Bowel Disease) [89]	35.3(1.4)	–	–	4.9(0.002)	6.7(0.3)	6.6(0.3)	4.6(0.5)
UC (Ulcerative Colitis) [89]	31.9(2.1)	58 [159]	0.55	2.7(0.002)	4.1(0.3)	4.1(0.3)	5.4(1.0)
T2D (Type 2 Diabetes) [109]	25.4(1.6)	26 [128]	0.98	1.3(0.002)	1.1(0.2)	1.1(0.2)	3.9(0.7)

Table 3.1: **Total SNP heritability estimates and the amount of h^2_g attributable to loci containing GWAS index SNPs ($h^2_{g,local,gwas}$) and index SNPs only (h^2_{gwas}).** $h^{2*}_{g,local,gwas}$ is the same as $h^2_{g,local,gwas}$ except that GWAS index SNPs are excluded in the computation. In Table S2, we report $h^{2\dagger}_{g,local,gwas}$, obtained by excluding all GWAS hits. We also report familial heritability (h^2_{pub}) estimates obtained from twin or family studies. We list case-control traits where our estimate of h^2_g is biased due to ascertainment at the bottom of the table. [a]Similar to [44], we define enrichment as the ratio between the fraction of h^2_g attributable to $h^{2*}_{g,local,gwas}$ and the fraction of genome covered by these loci. We obtain standard errors by jackknife over the loci. [b]IBD refers to the union of CD and UC.

Trait	**No. GWAS hit loci**	**No. GWAS loci with multiple signals**	$\hat{h}^2_{g,local,gwas}$ (%)	$\hat{h}^2_{gwas}$ (%)	**Fraction** (%)
HDL (High Density Lipoprotein) [29]	92	15	6.1(0.14)	2.8(0.003)	67.3
TG (Triglycerides) [29]	66	9	4.6(0.12)	3.0(0.002)	57.1
RA (Rheumatoid Arthritis) [114]	51	4	14.8(0.19)	4.3(0.005)	97.3
SCZ (Schizophrenia) [113]	103	2	0.28(0.003)	0.17(0.003)	3.6

Table 3.2: **GCTA-COJO[169] analysis on summary statistics for the traits HDL, TG, RA, and SCZ.** We define loci with multiple association signals as loci containing at least 2 of the risk SNPs reported by GCTA-COJO. Here, $\hat{h}^2_{g,local,gwas}$ and $\hat{h}^2_{gwas}$ are computed restricting to the loci with multiple association signals. Fraction refers to the fraction of difference between $\hat{h}^2_{g,local,gwas}$ and $\hat{h}^2_{gwas}$ across all loci that is accounted for by loci with multiple signals of association.

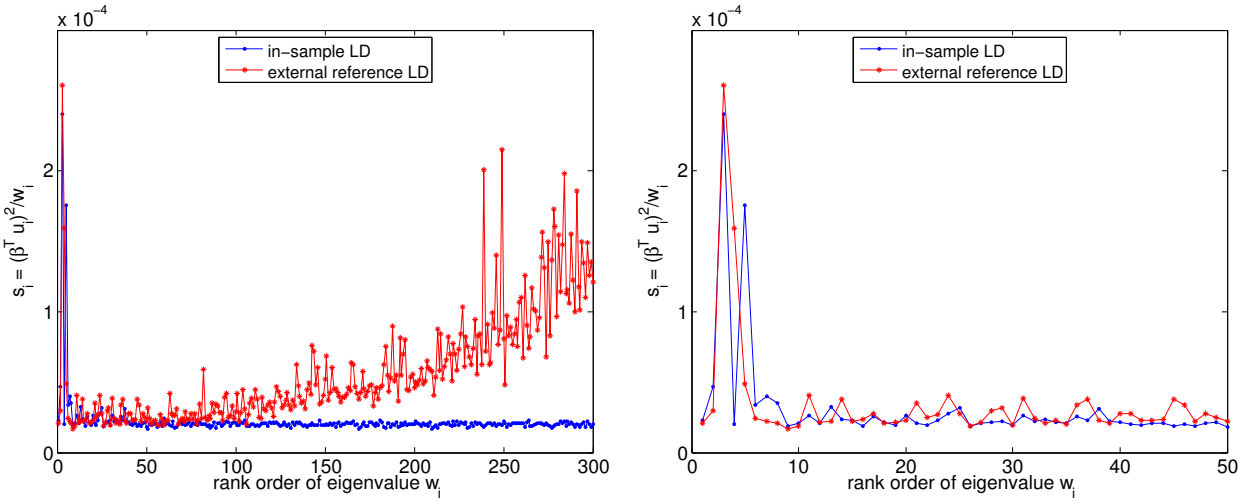
Trait	Sample size	No. SNPs	No. GWAS hits	No. index SNPs	Fraction[a]
BMI (Body Mass Index) [90]	229269	1859666	1851	79	5.31
Height (Height) [165]	244015	1854761	26374	476	31.15
HB (Haemoglobin) [160]	52666	1894024	459	24	1.38
MCH (Mean Cell Haemoglobin) [160]	44658	1892019	1585	37	2.25
MCHC (MCH Concentration) [160]	48252	1893281	223	15	0.9
MCV (Mean Cell Volume) [160]	49808	1893769	1602	46	3.08
PCV (Packed Cell Volume) [160]	46169	1893412	288	14	0.92
RBC (Red Blood Cell Count) [160]	46465	1892553	1132	31	2.1
PLT (Number of Platelets) [52]	66867	1954590	954	40	2.54
FG (Fasting Glucose) [38]	46186	1824182	290	12	0.97
FI (Fasting Insulin) [38]	46186	1822388	–	–	–
HBA1C (HBA1C) [147]	46368	1870395	187	11	0.6
HOMA-B (HOMA-B) [38]	46186	1820938	119	4	0.24
HOMA-IR (HOMA-IR) [38]	46186	1821061	–	–	–
HDL (High Density Lipoprotein) [29]	96335	1805617	3445	92	6.28
LDL (Low Density Lipoprotein) [29]	91529	1803637	2971	76	4.87
TC (Total Cholesterol) [29]	96596	1805676	4039	91	5.98
TG (Triglycerides) [29]	92768	1803908	3149	91	3.95
EY (Education Years) [134]	126559	1788888	11	4	0.25
FA (Forearm BMD) [176]	53236	4725343	152	3	0.18
FN (Femoral Neck BMD) [176]	53236	4637340	867	21	1.21
LS (Lumbar Spine) [176]	53236	4636561	1077	24	1.39
AM (Age at Menarche) [124]	132989	1821879	2391	73	4.61
COL (College) [134]	126559	1792881	61	3	0.2
RA (Rheumatoid Arthritis) [114]	14361/43923	4265540	19575	51	3.06
SCZ (Schizophrenia) [113]	32405/42221	4772186	8113	103	6.9
CD (Crohn's Disease) [89]	17897/33977	4822932	5179	54	3.48
IBD[b] (Inflammatory Bowel Disease) [89]	13769/33977	4823603	9243	70	4.17
UC (Ulcerative Colitis) [89]	31666/33977	4823578	5114	42	2.45
T2D (Type 2 Diabetes) [109]	12171/56862	1806359	236	13	1.0

Table 3.3: **Details of the summary GWSA data for the 30 analyzed traits.** [a]Fraction refers to the fraction of genome with GWAS hits. [b]IBD refers to the union of CD and UC. For case-control traits, we list sample size as No. cases / No. controls.

Trait	h_g^2**(HESS)**	**k**	**Estimated** λ_{gc}	$h^{2\dagger}_{g,local,gwas}$	**Enrichment**[a]	h_g^2**(LDSC)**
BMI (Body Mass Index) [90]	16.5(0.5)	50	1.33	2.45(0.11)	3.22(0.27)	14.0(0.9)
Height (Height) [165]	59.4(0.3)	50	1.00	23.86(0.20)	1.73(0.05)	33.0(1.7)
HB (Haemoglobin) [160]	17.9(2.1)	16	1.29	1.40(0.28)	6.19(1.38)	27.4(1.4)
MCH (Mean Cell Haemoglobin) [160]	29.3(2.2)	14	1.32	3.16(0.39)	6.71(1.28)	39.5(2.6)
MCHC (MCH Concentration) [160]	10.9(2.5)	15	1.30	0.40(0.25)	5.41(1.70)	21.6(0.9)
MCV (Mean Cell Volume) [160]	26.3(2.0)	15	1.31	3.08(0.39)	5.66(0.91)	35.2(2.1)
PCV (Packed Cell Volume) [160]	16.7(2.5)	14	1.31	0.64(0.25)	4.71(1.26)	31.4(1.5)
RBC (Red Blood Cell Count) [160]	22.0(2.3)	14	1.32	1.61(0.35)	4.48(0.82)	34.2(1.7)
PLT (Number of Platelets) [52]	27.5(1.5)	20	1.26	2.41(0.25)	4.04(0.44)	30.2(1.4)
FG (Fasting Glucose) [38]	22.3(2.3)	14	1.20	0.66(0.21)	3.58(1.11)	27.6(1.6)
FI (Fasting Insulin) [38]	19.9(2.4)	14	1.19	0.10(0.06)	15.41(0.00)	24.0(1.0)
HBA1C (HBA1C) [147]	20.8(2.3)	14	1.24	0.69(0.20)	5.31(1.89)	31.8(1.2)
HOMA-B (HOMA-B) [38]	20.3(2.4)	14	1.19	0.06(0.12)	1.26(0.68)	24.2(1.1)
HOMA-IR (HOMA-IR) [38]	19.9(2.4)	14	1.20	0.11(0.06)	16.17(0.00)	24.9(1.1)
HDL (High Density Lipoprotein) [29]	39.4(0.9)	29	1.13	4.33(0.24)	2.78(0.26)	33.4(7.5)
LDL (Low Density Lipoprotein) [29]	33.0(1.0)	27	1.16	3.97(0.24)	3.34(0.33)	27.0(4.5)
TC (Total Cholesterol) [29]	35.5(0.9)	29	1.16	5.27(0.25)	3.19(0.28)	27.2(3.8)
TG (Triglycerides) [29]	34.8(0.9)	28	1.18	3.76(0.21)	3.69(0.47)	31.4(5.2)
EY (Education Years) [134]	19.9(0.8)	38	1.05	0.15(0.04)	3.20(1.45)	12.6(0.5)
FA (Forearm BMD) [176]	17.4(2.2)	16	1.18	0.19(0.10)	9.90(4.33)	20.6(0.9)
FN (Femoral Neck BMD) [176]	24.1(2.1)	16	1.17	1.43(0.25)	5.39(0.81)	26.7(1.2)
LS (Lumbar Spine) [176]	25.1(2.0)	16	1.17	1.61(0.26)	4.70(0.61)	26.7(1.1)
AM (Age at Menarche) [124]	27.8(0.7)	40	1.05	3.18(0.17)	2.60(0.16)	16.5(0.7)
COL (College) [134]	19.4(0.8)	38	1.08	0.13(0.04)	3.34(0.98)	11.6(0.5)
RA (Rheumatoid Arthritis) [114]	66.3(0.9)	18	1.20	5.98(0.32)	5.82(1.29)	34.0(8.7)
SCZ (Schizophrenia) [113]	64.5(0.7)	22	1.00	8.36(0.21)	2.20(0.15)	43.7(1.4)
CD (Crohn's Disease) [89]	35.9(1.8)	16	1.12	3.64(0.37)	3.47(0.43)	31.4(2.1)
IBD[c] (Inflammatory Bowel Disease) [89]	35.3(1.4)	20	1.09	4.45(0.33)	3.66(0.39)	26.3(1.5)
UC (Ulcerative Colitis) [89]	31.9(2.1)	15	1.11	2.96(0.36)	4.35(0.73)	28.5(1.3)
T2D (Type 2 Diabetes) [109]	25.4(1.6)	19	1.19	0.71(0.16)	2.63(0.49)	24.5(1.1)

Table 3.4: **Total SNP-heritability for the 30 traits obtained by HESS and LDSC.** To obtain LDSC estimate, we compute LD scores for all SNPs with MAF greater than 5% using the same reference panel as used by HESS. Since HESS does not account for population stratification, we obtain LDSC estimate without the intercept. $h^{2\dagger}_{g,local,gwas}$ refers to the estimated SNP-heritability attributable to loci containing GWAS hit after all GWAS hits are removed. [a]We define enrichment as the ratio between the fraction of h_g^2 attributable to $h^{2\dagger}_{g,local,gwas}$ and the fraction of genome covered by these loci. We obtain standard errors by jack-knife over the loci.

3.6 Figures



(a) (b)

Figure 3.1: $s_i = (\hat{\boldsymbol{\beta}}^T \mathbf{u}_i)^2/w_i$ **as a function of the rank order of eigenvalue** w_i **obtained under in-sample LD (blue, rank=974) and external reference LD (red, rank=251) for a locus containing 1,377 SNPs.** Each point represents the mean of s_i over 500 simulations. Figure 3.1a displays the first 300 s_i. Figure 3.1b focuses on the first 50 s_i.

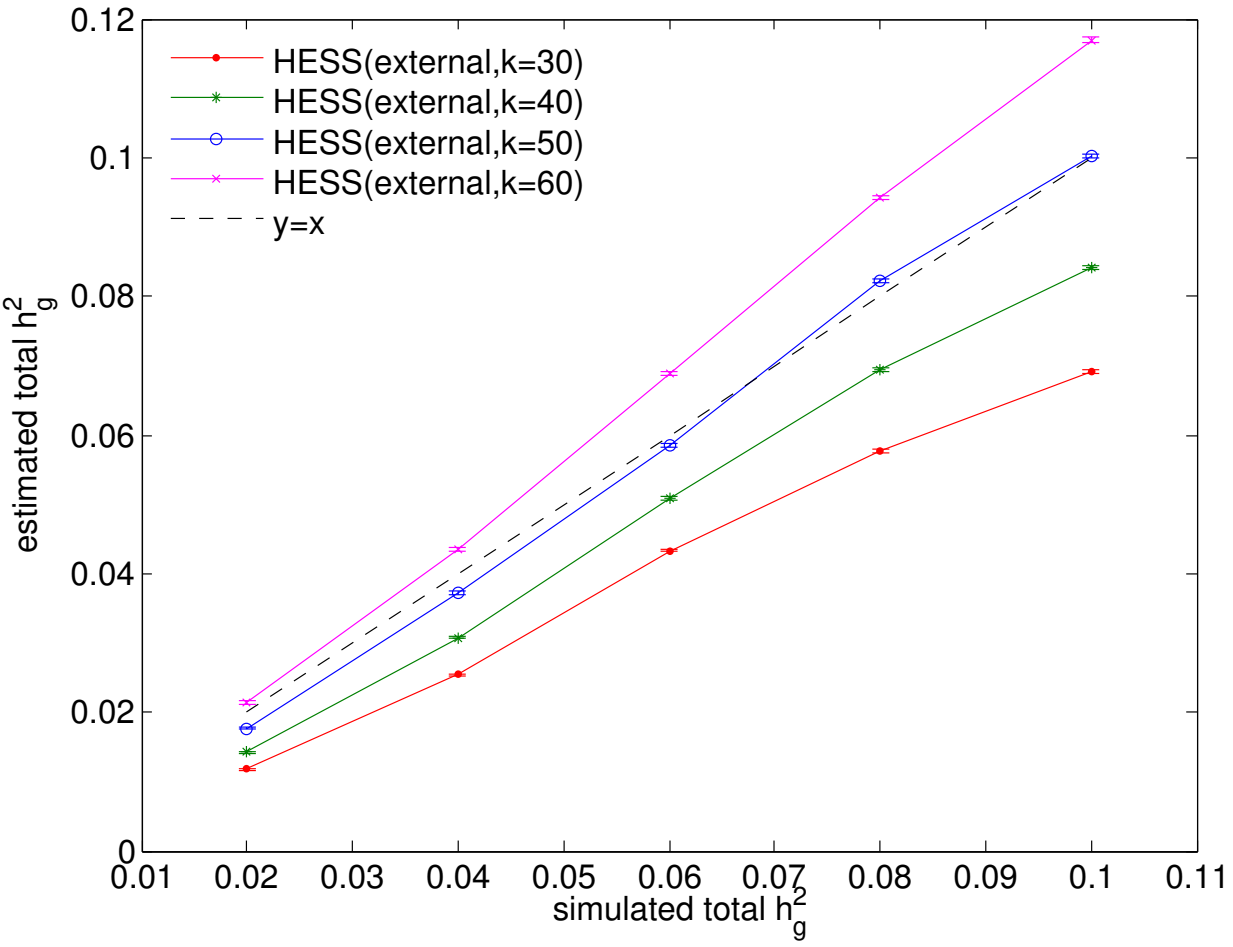


Figure 3.2: **Total SNP-heritability estimates in the whole chromosome simulation for different number (k) of eigenvectors included.** We see a slight downward bias when k is small (e.g. $k = 30$), and upward bias when k is large (e.g. $k = 60$). When $k = 50$, we attain approximately unbiased estimate of total SNP-heritability.

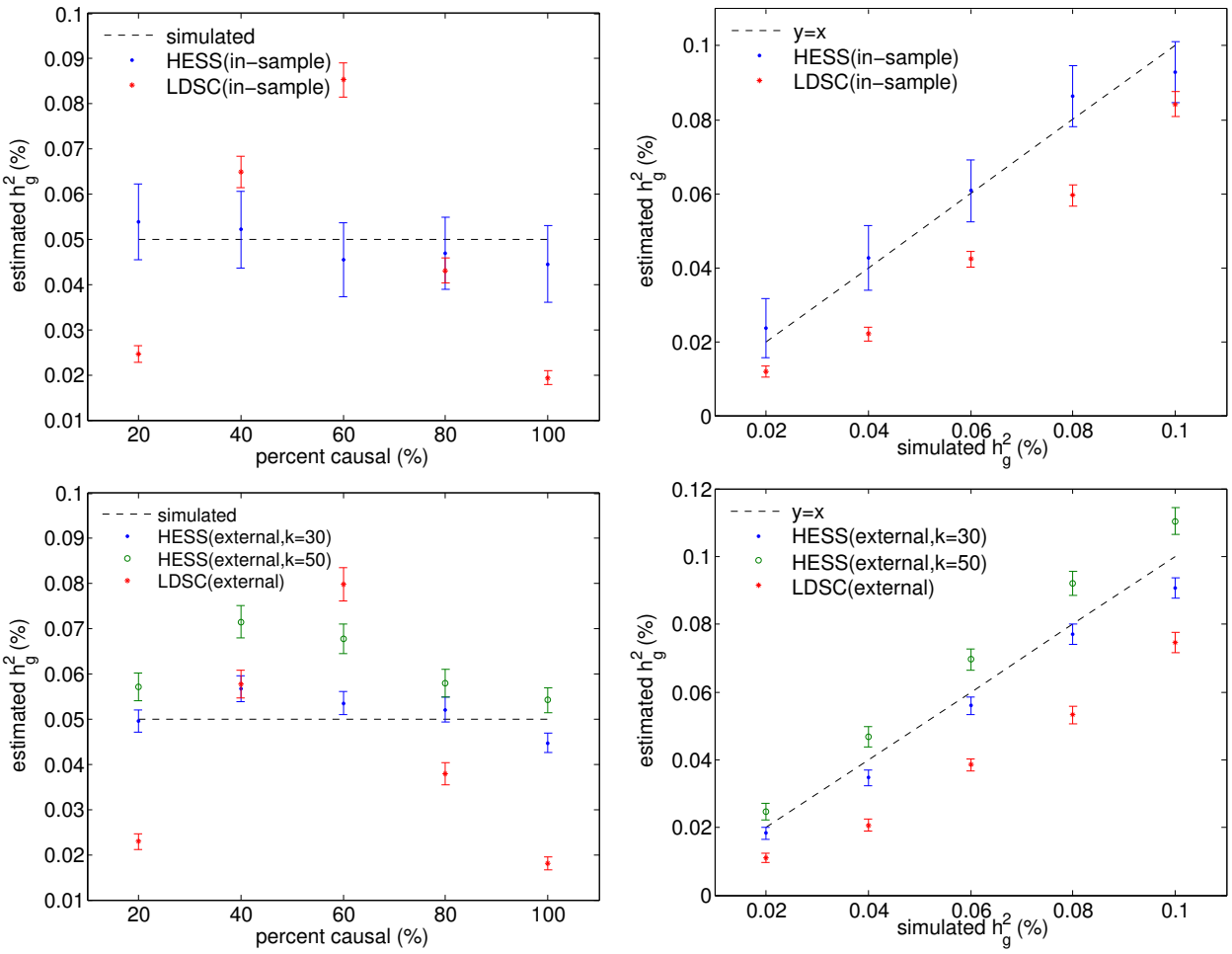


Figure 3.3: **HESS provides superior accuracy over LDSC in estimating local heritability.** HESS attains unbiased estimates when in-sample LD is used (top) and approximately unbiased estimates when reference LD is used (bottom). Mean and standard errors in these figures are computed based on 500 simulations, each involving 50,000 simulated GWAS data sets.

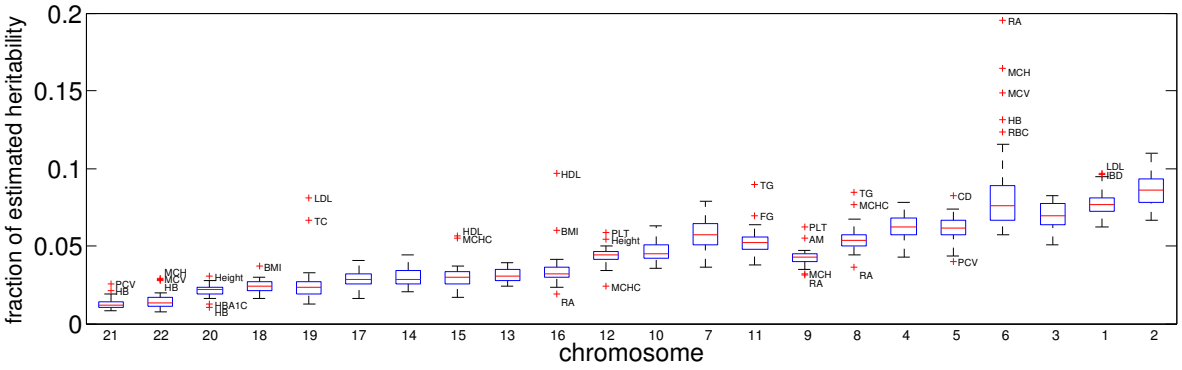


Figure 3.4: **Fraction of h_g^2 per chromosome across the 30 traits studied.** Here, the chromosomal heritability is obtained by summing local heritability at loci within the chromosome. For each chromosome we plot the box plots of estimates at the 30 considered traits. Chromosomes are ordered by size. With some notable exceptions, all traits show a strong polygenic signature of genetic architecture.

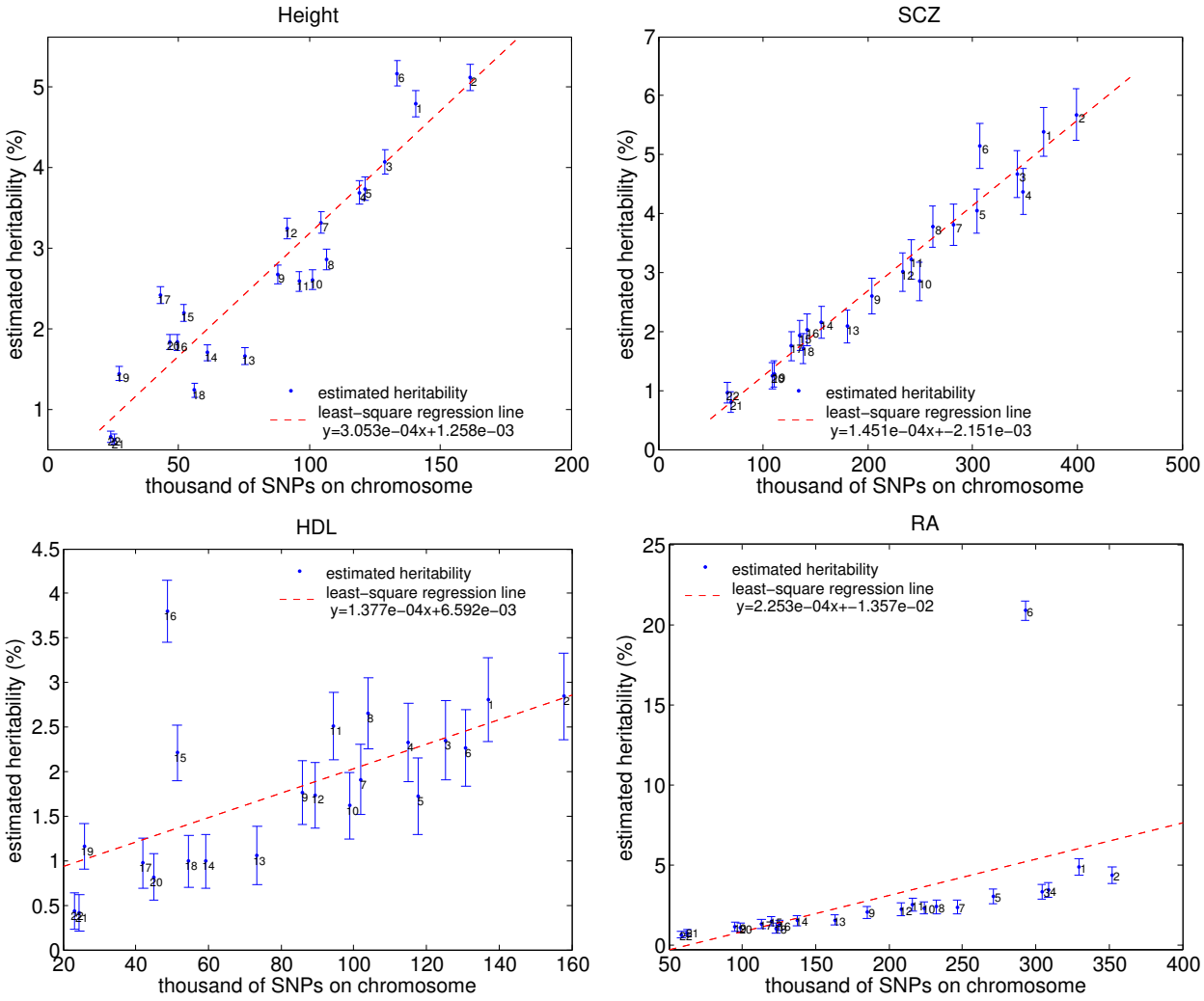


Figure 3.5: **Heritability attributable to each chromosome for four example traits.** The chromosomal heritability is obtained by summing local heritability at loci within the chromosome. Standard error is obtained by taking the square root of the sum of variance estimation.

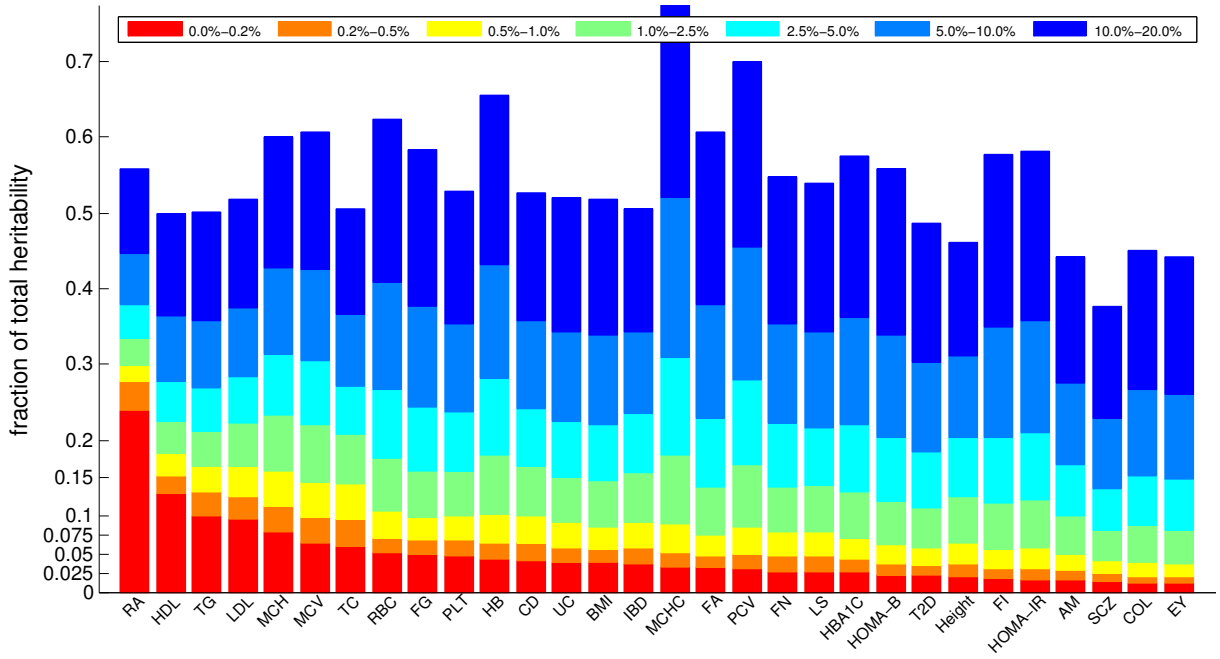


Figure 3.6: **Stacked bar plot showing the percentage of total heritability attributable to different fractions of genome.** We rank ordered all genomic loci by their explained heritability and quantified the fraction of total heritability attributable to different percentile ranges. Traits with high polygenicity tend to have bars with height proportional to bin size (e.g. Height and SCZ), whereas less polygenic traits tend to have bars much larger than bin size (e.g. RA and HDL).

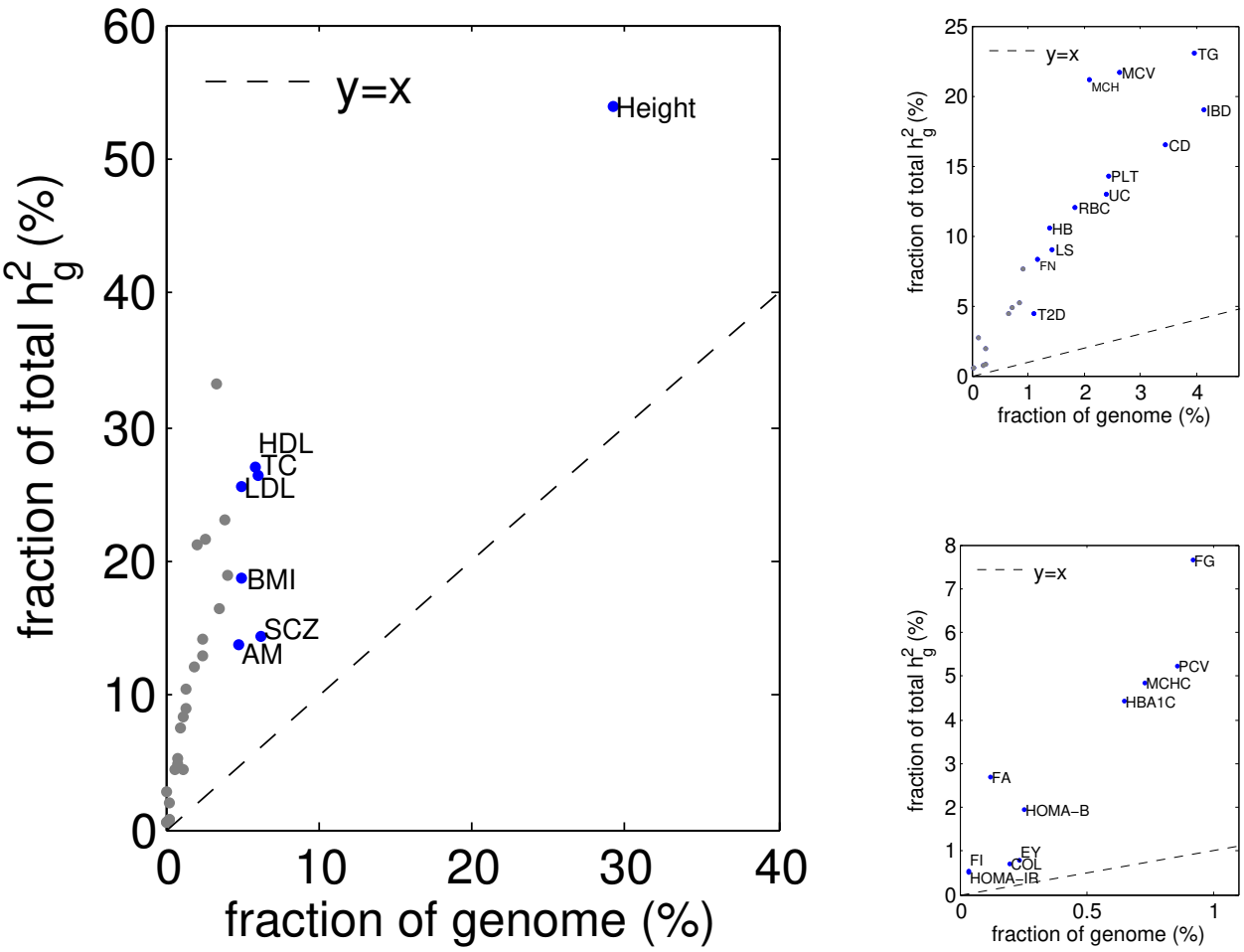


Figure 3.7: **Fraction of h_g^2 explained by all loci that contain a GWAS hit versus the fraction of genome covered by these loci.** Images on the right focus successively on the traits near the bottom left.

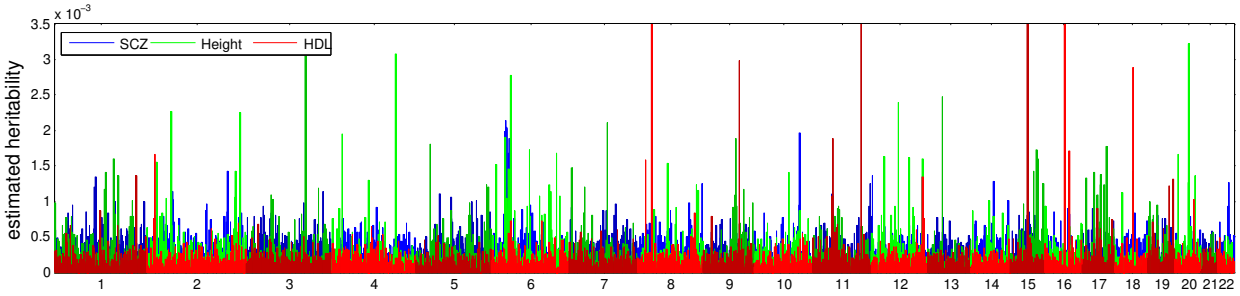


Figure 3.8: **Manhattan-style plots of regional heritability across the genome for the traits Height, HDL, and SCZ.**

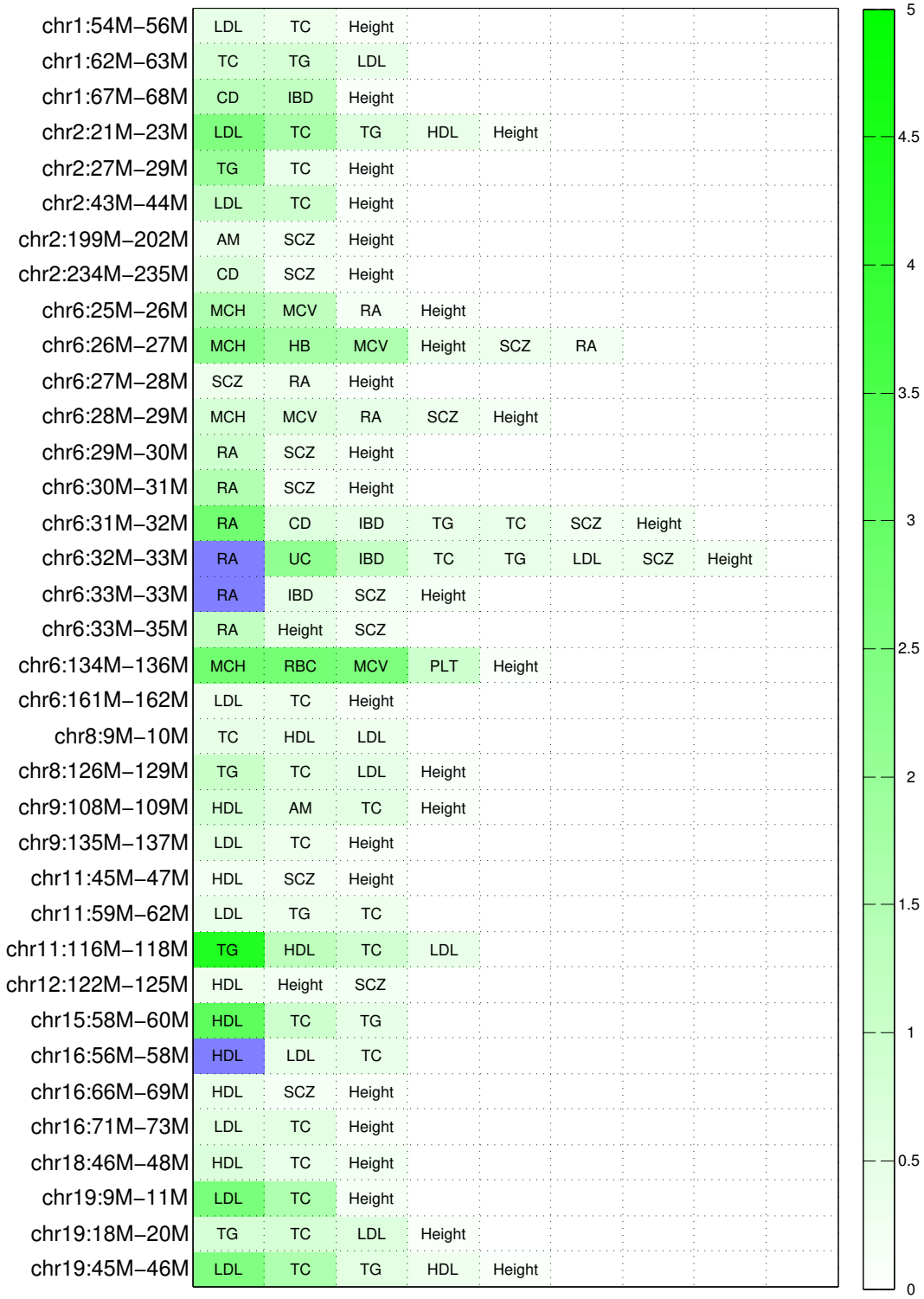


Figure 3.9: **Heat map showing the fraction of total SNP-heritability (h_g^2) contributed by each of the 36 "pleiotropic" loci.** For each locus, we only display traits to which the locus contributes significant amount of heritability. We mark traits to which the locus contributes more than 5% of the total SNP heritability in dark blue.

CHAPTER 4

Local genetic correlation gives insights into the shared genetic architecture of complex traits

4.1 Introduction

Genomic regions that harbor variants contributing to multiple traits provide valuable insights into the underlying biological mechanisms with which genetic variation impacts complex traits [49, 125, 98, 55, 129, 139, 162]. Therefore, both de novo discovery of such regions as well as the quantification of the correlation in effect sizes at known shared regions are important to epidemiological and etiological studies. For example, genetic variants associated with multiple traits in genome-wide associations studies (GWAS) can be used as instrumental variables in Mendelian randomization analyses to suggest causal relationships among complex traits [80, 146, 162, 32]. Unfortunately, many risk variants are left undetected by existing GWAS due to a combination of high polygenicity (i.e. many variants of small effects) and sample sizes which limits the power to detect genetic variants of small effect[168]. To improve accuracy at sub-GWAS significant regions, recent works [49, 125] proposed to utilize the posterior probability of two traits sharing a causal variant at a given risk region to detect genetic overlap. Although powerful in detecting shared genetic risk variants, the posterior probability does not convey the direction or magnitude of the genetic effect at the

This chapter is published in Shi et al., American Journal of Human Genetics 2017 [**?**]

overlapped genomic regions [49, 125]. Alternative approaches have used genetic correlation (i.e. correlation of the genetic components of two traits), that summarizes both direction and magnitude of effects, to gain insights into genetic overlap of complex traits[83, 19, 112]. Traditional methods to estimate genetic correlation are hindered by the lack of availability of large-scale individual-level data due to privacy concerns as they require individual genotype and trait measurements on the same set of individuals [112, 83, 59]. More recent works have shown that GWAS summary data (i.e., effect sizes and standard errors at all variants typed in the study) are sufficient to estimate genome-wide genetic correlation under a polygenic trait architecture by aggregating information across all typed variants in the study[18, 119].

In this work, we investigate the correlation between traits due to typed genetic variants from a small region in the genome (i.e. local genetic correlation) as means to identify genomic regions that contribute disproportionately to the genetic sharing between traits. We introduce methods that estimate the local genetic correlation from GWAS summary data while allowing for overlapping GWAS samples and linkage disequilibrium (LD) among variants. We partition the genome-wide genetic sharing across approximately independent LD regions of 1.6Mb in width on average[12]. To allow for a broad range of causal effect sizes, our approach makes no distributional assumptions on the causal effect sizes by treating them as fixed quantities. Our method can be viewed as a natural extension to pairs of traits of recently proposed methods that quantify local SNP-heritability from GWAS summary data under a fixed-effect model[140].

We illustrate the utility of local genetic correlation through an analysis of GWAS summary data of 36 quantitative complex traits. We identify 25 genomic regions that show significant local genetic correlation across 27 pairs of traits; e.g., region chr2:21-23M that harbors *APOB* (MIM 107730) shows a significant genetic correlation for the pair of traits High Density Lipoprotein (HDL) and Triglycerides (TG). Notably, 6 (out of the 25) regions show significant local genetic correlation although the genome-wide genetic correlation is not significantly different from 0; e.g. region chr6:134-136M shows a significant in local genetic correlation for mean cell volume (MCV) and platelet count (PLT) although the genome-wide genetic

correlation MCV-PLT is negligible (0.02, 95% CI [-0.04, 0.07]). This shows that these traits are correlated at a local level (e.g., due to pleiotropy and/or shared pathways) that are not reflected in the genome-wide correlation (due to balancing effect of other loci; e.g., positive correlation partially canceling a negative correlation, see Figure 4.1). Regions with significant local genetic correlations can also be used to identify new risk loci. For example, although the region chr8:9.2M-9.6M shows a significant local genetic correlation between HDL and LDL, although it does not harbor GWAS variant for HDL and LDL . Finally, we explore putative causal relations between all the 36 studied traits using a recently proposed approach[125] and report 55 instances of pairs with putative causality. For most of these pairs, we show that the local genetic correlation ascertained for GWAS signals specific to each trait is consistent with the putative causal relation while providing a directly interpretable quantity of the magnitude of effect.

4.2 Material and methods

4.2.1 Overview of methods

Genetic covariance measures the similarity between a pair of traits driven by genetic variations, and enjoys wide applications in understanding relations between complex traits[60, 22, 19]. Genetic covariance is traditionally estimated as a single measure across the entire genome to capture the genome-wide contribution of genetic variations to the correlation between phenotypes. Here, we introduce local genetic covariance, the similarity between pairs of traits driven by genetic variations localized at a specific region in the genome (e.g., one LD block), as a principled way to partition the shared genetic risk between traits. For example, a high genome-wide genetic covariance can be driven by one genomic region containing a shared risk variant, or by a large number of regions each with a small contribution reflecting putative causal relations (where all risk variants for one trait are risk variants for the other trait) and/or pleiotropy (risk variants contributing to both traits through shared pathways) (see Figure 4.1). Whereas genetic covariance quantifies the magnitude of co-variation of the

genetic components of two traits in their original scale, *genetic correlation* quantifies covariation in a standardized scale, and is therefore comparable across pairs of traits and/or genomic regions for which magnitude of effect size may differ. As a motivating example, consider two traits modeled by $\phi = x_1\beta_1 + x_2\beta_2 + \epsilon$ and $\psi = x_1\gamma_1 + x_2\gamma_2 + \delta$, where x_1 and x_2 represent two independent SNPs. In the special case where $\boldsymbol{\gamma}$ is proportional to $\boldsymbol{\beta}$ by a factor of α, i.e. $\boldsymbol{\gamma} = \alpha\boldsymbol{\beta}$, the genetic covariance between the two traits is $\alpha(\beta_1^2 + \beta_2^2)$, and is governed by α. However, the genetic correlation between the two traits is always 1 for positive α (-1 for negative α) regardless of the magnitude of α.

We start by defining local genetic covariance under the fixed effect model, making a distinction between genetic covariance and covariance of the causal effects, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (see below). We then describe methods to estimate genetic covariance followed by an approach to standardize the local genetic covariance to estimate local genetic correlation.

4.2.2 Local genetic covariance under fixed-effect model

Let $\phi = \boldsymbol{x}^\intercal\boldsymbol{\beta}+\epsilon$ and $\psi = \boldsymbol{x}^\intercal\boldsymbol{\gamma}+\delta$ be two traits measured at an individual, standardized so that $\mathrm{E}[\phi] = \mathrm{E}[\psi] = 0$ and $\mathrm{V}ar[\phi] = \mathrm{V}ar[\psi] = 1$, where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$ are the fixed effect size vectors for the two traits; $\boldsymbol{x} \in \mathbb{R}^p$, the genotype vector of the individual at p SNPs, standardized so that $\mathrm{E}[\boldsymbol{x}] = \mathbf{0}$, and $\mathrm{V}ar[\boldsymbol{x}] = \boldsymbol{V}$, the LD matrix; and ϵ, δ, random environmental effects independent of $\boldsymbol{x}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, with $\mathrm{E}[\epsilon] = \mathrm{E}[\delta] = 0$, $\mathrm{V}ar[\epsilon] = \sigma_\epsilon^2$, $\mathrm{V}ar[\delta] = \sigma_\delta^2$, and $\mathrm{C}ov[\epsilon, \delta] = \rho_e$. Under these assumptions, one can decompose the phenotypic covariance, ρ, between ϕ and ψ into a summation of genetic covariance and environmental covariance, as

$$\begin{aligned}
\rho &= \mathrm{C}ov[\phi, \psi] = \mathrm{E}[\phi\psi] - \mathrm{E}[\phi]\,\mathrm{E}[\psi] = \mathrm{E}[(\boldsymbol{x}^\intercal\boldsymbol{\beta} + \epsilon)(\boldsymbol{x}^\intercal\boldsymbol{\gamma} + \delta)^\intercal] \\
&= \mathrm{E}[(\boldsymbol{x}^\intercal\boldsymbol{\beta})(\boldsymbol{x}^\intercal\boldsymbol{\gamma})] + \mathrm{E}[\epsilon\delta] = \mathrm{C}ov[\boldsymbol{x}^\intercal\boldsymbol{\beta}, \boldsymbol{x}^\intercal\boldsymbol{\gamma}] + \mathrm{C}ov[\epsilon, \delta] \\
&= \boldsymbol{\beta}^\intercal\,\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^\intercal]\boldsymbol{\gamma} + \mathrm{C}ov[\epsilon, \delta] = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma} + \rho_e,
\end{aligned} \tag{4.1}$$

where $\rho_g = \mathrm{C}ov[\boldsymbol{x}^\intercal\boldsymbol{\beta}, \boldsymbol{x}^\intercal\boldsymbol{\gamma}] = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma}$ is the genetic covariance between the two traits (i.e. covariance between the genetic components of the two traits, $\boldsymbol{x}^\intercal\boldsymbol{\beta}$ and $\boldsymbol{x}^\intercal\boldsymbol{\gamma}$), and ρ_e the

environmetal covariance (i.e. covariance between the environmental effects of two traits, ϵ and δ). The magnitude and sign of local genetic covariance can be interpreted as the effect and direction of the local genetic component of one trait on that of the other. Thus, given the true effect size vectors, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and the LD matrix $\boldsymbol{V}$, one can obtain ρ_g by plugging in these quantities.

4.2.3 Genetic covariance versus covariance of the causal effects

An alternative approach to the covariance of the genetic components of the traits, is to quantify the *covariance (correlation) of the causal effects* (i.e. $\rho_{g,causal} = \boldsymbol{\beta}^\intercal\boldsymbol{\gamma}$). In the special case where there is no LD (i.e. $\boldsymbol{V} = \boldsymbol{I}$, the identity matrix), genetic covariance and covariance of the causal effects coincide, $\rho_g = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma} = \boldsymbol{\beta}^\intercal\boldsymbol{I}\boldsymbol{\gamma} = \boldsymbol{\beta}^\intercal\boldsymbol{\gamma} = \rho_{g,causal}$. However, in general genetic covariance is different from covariance of the causal effects as function of the LD between the causal variants. More importantly, high local genetic covariance does not necessarily imply high covariance of the causal effects. In fact, high genetic covariance can be attained even when causal variants are different between the traits. To illustrate the difference, consider an example involving 2 SNPs. Let $\boldsymbol{\beta} = (1, 0)$ and $\boldsymbol{\gamma} = (0, 1)$ be the causal effect vectors of the two traits, i.e. the two traits have two distinct set of causal variants. And let

$$\boldsymbol{V} = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix}$$

be the LD matrix between the SNPs. In this example, the covariance of the causal effects is $\rho_{g,causal} = \boldsymbol{\beta}^\intercal\boldsymbol{\gamma} = 0$, whereas the genetic covariance is $\rho_g = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma} = 0.9$. Thus, at a region where the causal variants are distinct for the two traits, covariance of the causal effects is always zero, whereas genetic covariance may be non-zero depending on the LD (see Figure 4.2). The two definitions measure genetic sharing at different levels of resolution. Local genetic covariance measures sharing at regional level giving a measure of how similar the regional genetic components are between the two traits, and has applications in predicting the regional genetic component of one trait from that of the other. In contrast, local causal

effect covariance measures sharing at an individual SNP level giving a measure of how similar the causal effects are between the two traits. Consider a scenario where two traits are each driven locally by a different SNP in the same gene. In this case, the local causal effect covariance is zero since the two traits share no causal SNP. However, the local genetic covariance is non-zero if the two SNPs are in LD, which induces similarity in the genetic component of the two traits, and is an indication of the gene being shared across the two traits. Although in this work we focus on genetic covariance, for completeness we discuss an estimator for covariance of the causal effects ($\rho_{g,causal}$) in Appendix.

4.2.4 Estimating local genetic covariance from GWAS summary data

In two GWASs involving n_1 individuals for trait 1 ($\boldsymbol{\phi}$), n_2 individuals for trait 2 ($\boldsymbol{\psi}$), and n_s shared individuals, we assume

$$\begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{\phi}_s \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X}_s \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_s \end{bmatrix}, \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\psi}_s \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{X}'_s \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\delta}_s \end{bmatrix}, \tag{4.2}$$

where $(\boldsymbol{\phi}, \boldsymbol{\phi}_s) \in \mathbb{R}^{n_1}$ and $(\boldsymbol{\psi}, \boldsymbol{\psi}_s) \in \mathbb{R}^{n_2}$ are the standardized trait values of all individuals in each GWAS; $(\boldsymbol{Y}, \boldsymbol{X}_s) \in \mathbb{R}^{n_1 \times p}$, $(\boldsymbol{Z}, \boldsymbol{X}'_s) \in \mathbb{R}^{n_2 \times p}$, column standardized genotype matrices of all individuals in each GWAS, where $\boldsymbol{X}_s$ and $\boldsymbol{X}'_s$ represent the genotype matrices for the same set of individuals and SNPs but standardized differently in each GWAS; $(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_s) \in \mathbb{R}^{n_1}$, $(\boldsymbol{\delta}, \boldsymbol{\delta}_s) \in \mathbb{R}^{n_2}$, environmental effects of all individuals in each GWAS. We use the subscript 's' to represent individuals shared by both GWASs. We further assume that $\mathrm{E}[\boldsymbol{\epsilon}] = \mathrm{E}[\boldsymbol{\delta}] = \mathrm{E}[\boldsymbol{\epsilon}_s] = \mathrm{E}[\boldsymbol{\delta}_s] = \boldsymbol{0}$, $\mathrm{V}ar[\boldsymbol{\epsilon}] = \mathrm{V}ar[\boldsymbol{\epsilon}_s] = \sigma^2_\epsilon \boldsymbol{I}$, $\mathrm{V}ar[\boldsymbol{\delta}] = \mathrm{V}ar[\boldsymbol{\delta}_s] = \sigma^2_\delta \boldsymbol{I}$, $\mathrm{C}ov[\boldsymbol{\epsilon}, \boldsymbol{\delta}] = \boldsymbol{0}$, and $\mathrm{C}ov[\boldsymbol{\epsilon}_s, \boldsymbol{\delta}_s] = \rho_e \boldsymbol{I}$.

In a traditional GWAS, we obtain marginal effect size estimates, $\hat{\boldsymbol{\beta}}_{gwas}$ and $\hat{\boldsymbol{\gamma}}_{gwas}$, as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{gwas} &= \frac{1}{n_1}\left[\boldsymbol{Y}^\intercal\ \boldsymbol{X}_s^\intercal\right]\begin{bmatrix}\boldsymbol{\phi}\\ \boldsymbol{\phi}_s\end{bmatrix} = \frac{1}{n_1}\left(\boldsymbol{Y}^\intercal\boldsymbol{Y} + \boldsymbol{X}_s^\intercal\boldsymbol{X}_s\right)\boldsymbol{\beta} + \frac{1}{n_1}\left(\boldsymbol{Y}^\intercal\boldsymbol{\epsilon} + \boldsymbol{X}_s^\intercal\boldsymbol{\epsilon}_s\right)\\ \hat{\boldsymbol{\gamma}}_{gwas} &= \frac{1}{n_2}\left[\boldsymbol{Z}^\intercal\ \boldsymbol{X'}_s^\intercal\right]\begin{bmatrix}\boldsymbol{\psi}\\ \boldsymbol{\psi}_s\end{bmatrix} = \frac{1}{n_2}\left(\boldsymbol{Z}^\intercal\boldsymbol{Z} + \boldsymbol{X'}_s^\intercal\boldsymbol{X'}_s\right)\boldsymbol{\gamma} + \frac{1}{n_2}\left(\boldsymbol{Z}^\intercal\boldsymbol{\delta} + \boldsymbol{X'}_s^\intercal\boldsymbol{\delta}_s\right).\end{aligned} \tag{4.3}$$

Assuming individuals in both GWASs are drawn from the same population with LD matrix $\boldsymbol{V}$, we have $\hat{\boldsymbol{\beta}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\beta}, \frac{\sigma_\epsilon^2}{n_1}\boldsymbol{V}\right)$, $\hat{\boldsymbol{\gamma}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\gamma}, \frac{\sigma_\delta^2}{n_2}\boldsymbol{V}\right)$. We also find

$$Cov[\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}] = \mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}\hat{\boldsymbol{\gamma}}_{gwas}^\intercal] - (\boldsymbol{V}\boldsymbol{\beta})(\boldsymbol{V}\boldsymbol{\gamma})^\intercal = \frac{\rho_e}{n_1 n_2}\mathrm{E}[\boldsymbol{X}_s^\intercal\boldsymbol{X'}_s] = \frac{\rho_e n_s}{n_1 n_2}\boldsymbol{V}, \tag{4.4}$$

where the last equality follows from Isserlis' theorem [70].

Under infinite sample sizes, $Var[\hat{\boldsymbol{\beta}}_{gwas}] = Var[\hat{\boldsymbol{\gamma}}_{gwas}] = Cov[\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}] = \mathbf{0}$, and we have $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_{gwas}$, $\boldsymbol{\gamma} = \boldsymbol{V}^{-1}\hat{\boldsymbol{\gamma}}_{gwas}$. Thus, local genetic covariance, $\rho_{g,local}$, can be computed as

$$\rho_{g,local} = (\hat{\boldsymbol{\beta}}_{gwas}^\intercal\boldsymbol{V}^{-1})\boldsymbol{V}(\boldsymbol{V}^{-1}\hat{\boldsymbol{\gamma}}_{gwas}) = \hat{\boldsymbol{\beta}}_{gwas}^\intercal\boldsymbol{V}^{-1}\hat{\boldsymbol{\gamma}}_{gwas}. \tag{4.5}$$

However, when sample sizes are finite, from bilinear form theory [138], the covariance between $\hat{\boldsymbol{\beta}}_{gwas}$ and $\hat{\boldsymbol{\gamma}}_{gwas}$ creates bias, resulting in

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}^\intercal\boldsymbol{V}^{-1}\hat{\boldsymbol{\gamma}}_{gwas}] = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma} + \frac{\rho_e}{n_1 n_2}tr\left(\boldsymbol{V}\right) = \boldsymbol{\beta}^\intercal\boldsymbol{V}\boldsymbol{\gamma} + \frac{p(\rho - \rho_{g,local})n_s}{n_1 n_2}, \tag{4.6}$$

Correcting for bias, we arrive at the unbiased estimator

$$\hat{\rho}_{g,local} = \frac{n_1 n_2\hat{\boldsymbol{\beta}}_{gwas}^\intercal\boldsymbol{V}^{-1}\hat{\boldsymbol{\gamma}}_{gwas} - n_s p\rho}{n_1 n_2 - n_s p}. \tag{4.7}$$

For rank-deficient LD matrix $\boldsymbol{V}$, one replaces $\boldsymbol{V}^{-1}$ with the pseudo-inverse ($\boldsymbol{V}^\dagger$) and p with

$q = rank(\boldsymbol{V})$, yielding the unbiased estimator

$$\hat{\rho}_{g,local} = \frac{n_1 n_2 \hat{\boldsymbol{\beta}}_{gwas}^{\intercal} \boldsymbol{V}^{\dagger} \boldsymbol{\gamma}_{gwas} - n_s q \rho}{n_1 n_2 - n_s q}. \tag{4.8}$$

Thus, in order to obtain an unbiased estimate of genetic covariance between a pair of traits, one needs to know their phenotypic covariance. When phenotypic covariance is not available, one can obtain an estimate from genome-wide summary association data using cross-trait LD Score regression [18],

$$\mathrm{E}[z_{\phi,j} z_{\psi,j} | l_j] = \frac{\sqrt{n_1 n_2}\rho_g}{p} l_j + \frac{\rho n_s}{\sqrt{n_1 n_2}}, \tag{4.9}$$

where $z_{\phi,j}$, $z_{\psi,j}$ are the Z-scores of SNP j in the two traits, and l_j the LD score of SNP j. Cross-trait LD Score regression regresses the product of Z-scores at each SNP against its LD score, l_j, and accounts for bias generated by overlapping samples through the intercept term, $\frac{\rho n_s}{\sqrt{n_1 n_2}}$ [18], from which one can obtain an estimate of phenotypic covariance, ρ.

In the special case when $\hat{\boldsymbol{\beta}}_{gwas}$ and $\hat{\boldsymbol{\gamma}}_{gwas}$ are obtained for the same trait on the same set of individuals (i.e. $\hat{\boldsymbol{\beta}}_{gwas} = \hat{\boldsymbol{\gamma}}_{gwas}$, $n_1 = n_2 = n_s$, $\rho = 1$) Equation (4.7) reduces to the local SNP-heritability estimator [140]. When $n_s = 0$ (i.e. no shared individuals between the GWASs), the unbiased estimator is simply $\hat{\rho}_{g,local} = \hat{\boldsymbol{\beta}}_{gwas}^{\intercal} \boldsymbol{V}^{-1} \hat{\boldsymbol{\gamma}}_{gwas}$. An interpretation for this simple formula is that in the absence of sample overlap, the covariance in the noise, ϵ and δ, is 0 and does thus not introduce bias into the estimate of $\rho_{g,local}$.

Following bilinear form theory [138], we can estimate the variance for $\hat{\rho}_{g,local}$ as,

$$\mathrm{V}ar[\hat{\rho}_{g,local}] = \left(\frac{n_1 n_2}{n_1 n_2 - n_s p}\right)^2 \left[\left(\frac{p \rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2 \sigma_\delta^2 p}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local}^2}{n_2} + \frac{\sigma_\epsilon^2 h_{g\psi,local}^2}{n_1} + 2\frac{n_s \rho_e \rho_{g,local}}{n_1 n_2}\right] \tag{4.10}$$

For rank deficient LD matrix with $rank(\boldsymbol{V}) = q$, one replaces p with q in Equation (4.10).

4.2.5 Accounting for statistical noise in LD estimates

Limited sample size of external reference panels creates statistical noise in the estimated LD matrix that biases our estimates. Following our previous work [140], we apply truncated-SVD regularization [57] to remove noise in external reference LD. We note that $\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\gamma}}_{gwas} = \sum_{i=1}^{q} s_i = \sum_{i=1}^{q} \frac{1}{w_i}(\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{u}_i)(\hat{\boldsymbol{\gamma}}_{gwas}^{\intercal}\boldsymbol{u}_i)$, where w_i, $\boldsymbol{u}_i$ are the eigenvalues and eigenvectors of the LD matrix $\boldsymbol{V}$, and $q = rank(\boldsymbol{V})$. We use $\hat{s}_i = \frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\hat{\boldsymbol{u}}_i)(\hat{\boldsymbol{\gamma}}_{gwas}^{\intercal}\hat{\boldsymbol{u}}_i)$, to denote the counterpart obtained from external reference LD matrix $\hat{\boldsymbol{V}}$. We show through simulations that the bulk of $\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\gamma}}_{gwas}$ comes from s_i where $i \ll q$ and that $s_i \approx \hat{s}_i$ for $i \ll q$, thus justifying truncated-SVD as an appropriate regularization method when only external reference LD ($\hat{\boldsymbol{V}}$) is available.

Let $g(\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}, k) = \sum_{i=1}^{k} \hat{s}_i = \sum_{i=1}^{k} \frac{1}{\hat{w}_i}(\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\hat{\boldsymbol{u}}_i)(\hat{\boldsymbol{\gamma}}_{gwas}^{\intercal}\hat{\boldsymbol{u}}_i)$, be the truncated-SVD regularized estimates for $\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{V}^{\dagger}\hat{\boldsymbol{\gamma}}_{gwas}$, then it can be shown that

$$\mathrm{E}[g(\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}, k)] = \frac{n_s k(\rho - \rho_g)}{n_1 n_2} + \sum_{i=1}^{k} \hat{w}_i(\boldsymbol{\beta}^{\intercal}\hat{\boldsymbol{u}}_i)(\boldsymbol{\gamma}^{\intercal}\hat{\boldsymbol{u}}_i). \tag{4.11}$$

Assuming $\hat{w}_i = w_i$ and $\hat{\boldsymbol{u}}_i = \boldsymbol{u}_i$ for $i \ll k$, Equation (4.11) is a biased approximation of $\rho_{g,local}$, with bias $\frac{n_s k(\rho-\rho_g)}{n_1 n_2}$. Correcting for the bias, we arrive at the estimator

$$\hat{\rho}_{g,local} = \frac{n_1 n_2 g(\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}, k) - n_s \rho k}{n_1 n_2 - n_s k}, \tag{4.12}$$

which has variance

$$\mathrm{V}ar[\hat{\rho}_{g,local}] = \left(\frac{n_1 n_2}{n_1 n_2 - n_s k}\right)^2 \left[\left(\frac{k\rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2 \sigma_\delta^2 k}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local}^2}{n_2} + \frac{\sigma_\epsilon^2 h_{g\psi,local}^2}{n_1} + 2\frac{n_s \rho_e \rho_{g,local}}{n_1 n_2}\right] \tag{4.13}$$

4.2.6 Extension to multiple independent regions

For genome partitioned into m regions, let

$$\begin{aligned}\phi &= \boldsymbol{x}_1^\intercal \boldsymbol{\beta}_1 + \cdots + \boldsymbol{x}_m^\intercal \boldsymbol{\beta}_m + \epsilon \\ \psi &= \boldsymbol{x}_1^\intercal \boldsymbol{\gamma}_1 + \cdots + \boldsymbol{x}_m^\intercal \boldsymbol{\gamma}_m + \delta,\end{aligned} \tag{4.14}$$

denote the phenotype measurements of two traits at an individuals, where we assume that SNPs in different pairs of regions are independent, i.e. $\mathrm{E}[\boldsymbol{x}_{ik}\boldsymbol{x}_{il}] = 0$ for all $i \neq j$, $k \in \{1, \cdots, p_i\}$, and $l \in \{1, \cdots, p_j\}$, where p_i and p_j are the number of SNPs in region i and j. Under these assumptions, we decompose the phenotypic covariance, ρ, between ϕ and ψ, into a summation of per-region genetic covariance and environmental covariance

$$\begin{aligned}\rho &= \mathrm{C}ov[\phi, \psi] = \mathrm{E}[(\boldsymbol{x}_1^\intercal \boldsymbol{\beta}_1 + \cdots + \boldsymbol{x}_m^\intercal \boldsymbol{\beta}_m + \epsilon)(\boldsymbol{x}_1^\intercal \boldsymbol{\gamma}_1 + \cdots + \boldsymbol{x}_m^\intercal \boldsymbol{\gamma}_m + \delta)^\intercal] \\ &= \mathrm{E}[(\boldsymbol{x}_1^\intercal \boldsymbol{\beta}_1)(\boldsymbol{x}_1^\intercal \boldsymbol{\gamma}_1)] + \cdots + \mathrm{E}[(\boldsymbol{x}_m^\intercal \boldsymbol{\beta}_m)(\boldsymbol{x}_m^\intercal \boldsymbol{\gamma}_m)] + \mathrm{E}[\epsilon\delta] \\ &= \sum_{i=1}^{m} \mathrm{C}ov[\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_i, \boldsymbol{x}_i^\intercal \boldsymbol{\gamma}_i] + \mathrm{C}ov[\epsilon, \delta] = \sum_{i=1}^{m} \boldsymbol{\beta}_i^\intercal \boldsymbol{V}_i \boldsymbol{\gamma}_i + \rho_e\end{aligned}, \tag{4.15}$$

where $\rho_{g,local,i} = \mathrm{C}ov[\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_i, \boldsymbol{x}_i^\intercal \boldsymbol{\gamma}_i] = \boldsymbol{\beta}_i^\intercal \boldsymbol{V}_i \boldsymbol{\gamma}_i$ is the local genetic covariance between the pair of traits attributed to genetic variants at region i. Following strategies outlined in previous sections, we arrive at the estimator for genetic covariance at the i-th region,

$$\hat{\rho}_{g,local,i} = \frac{n_1 n_2 g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k) - n_s(\rho - \sum_{j=1, j\neq i}^{m} \hat{\rho}_{g,local,j}) k_i}{n_1 n_2 - n_s k_i}, \tag{4.16}$$

which defines a system of linear equation involving m unknown variables and m equations. Following bilinear form theory, we obtain variance estimate for $\hat{\rho}_{g,local,i}$ as,

$$\begin{aligned}\mathrm{V}ar[\hat{\rho}_{g,local,i}] = &\left(\frac{n_1 n_2}{n_1 n_2 - n_s k_i}\right)^2 \left[\left(\frac{k_i \rho_e n_s}{n_1 n_2}\right)^2 + \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2} + \frac{\sigma_\delta^2 h_{g\phi,local,i}^2}{n_2} + \frac{\sigma_\epsilon^2 h_{g\psi,local,i}^2}{n_1} + 2\frac{n_s \rho_e \rho_{g,local,i}}{n_1 n_2}\right] \\ &+ \sum_{j=1, j\neq i}^{m} \left(\frac{n_s k_j}{n_1 n_2 - n_s k_i}\right)^2 \mathrm{V}ar[\hat{\rho}_{g,local,j}]\end{aligned} \tag{4.17}$$

which also defines a system of linear equations with m equations and m variables. In the special case where there is no sample overlap ($n_s = 0$), $\hat{\rho}_{g,local,i}$ reduces to $g(\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}, k)$ with $\mathrm{V}ar[\hat{\rho}_{g,local,i}] = \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2} + \frac{\sigma_\delta^2 h^2_{g\phi,local,i}}{n_2} + \frac{\sigma_\epsilon^2 h^2_{g\psi,local,i}}{n_1} \approx \frac{\sigma_\epsilon^2 \sigma_\delta^2 k_i}{n_1 n_2}$, i.e. both the local genetic covariance and its variance can be estimated independent of all other windows.

When $k_1 = \cdots = k_m = k$, i.e. all regions use the same number of eigenvectors in the truncated-SVD regularization, summing over i on both sides of Equation (4.16) yields

$$\begin{aligned}
\hat{\rho}_g = \sum_{i=1} \hat{\rho}_{g,local,i} &= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k) - \frac{k n_s}{n_1 n_2 - n_s k} \sum_{i=1}^{m} \left(r - \sum_{j=1, j\neq i}^{m} \hat{\rho}_{g,local,j} \right) \\
&= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k) - \frac{k n_s}{n_1 n_2 - n_s k} \sum_{i=1}^{m} (\rho - \hat{\rho}_g + \hat{\rho}_{g,local,i}) \\
&= \frac{n_1 n_2}{n_1 n_2 - n_s k} \sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k) + \frac{k n_s m - k n_s}{n_1 n_2 - n_s k} \hat{\rho}_g - \frac{k n_s m \rho}{n_1 n_2 - n_s k}.
\end{aligned} \tag{4.18}$$

Solving for $\hat{\rho}_g$ yields

$$\hat{\rho}_g = \frac{n_1 n_2 \sum_{i=1}^{m} g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k) - k n_s m \rho}{n_1 n_2 - k n_s m}, \tag{4.19}$$

which has variance

$$\mathrm{V}ar[\hat{\rho}_g] = \left(\frac{n_1 n_2}{n_1 n_2 - k n_s m} \right)^2 \sum_{i=1}^{m} \mathrm{V}ar[g(\hat{\boldsymbol{\beta}}_{gwas,i}, \hat{\boldsymbol{\gamma}}_{gwas,i}, k)]. \tag{4.20}$$

Thus, if k is chosen such that $(n_1 n_2 - k n_s m)$ is small (i.e. $\frac{n_1 n_2}{n_1 n_2 - k n_s m}$ large), the estimate of total genetic covariance will have large standard error. To reduce standard error in the estimates (at the cost of some bias), we recommend choosing k such that $\frac{n_1 n_2}{n_1 n_2 - k n_s m}$ is less than 2. When testing for statistical significance, we assume that the estimates of local and genome-wide genetic covariance and correlation follow a normal distribution.

4.2.7 Standardizing local genetic covariance

We estimate the local genetic correlation for the i-th region as

$$\hat{r}_{g,local,i} = \frac{\hat{\rho}_{g,local,i}}{\sqrt{\hat{h}^2_{g\phi,local,i}}\sqrt{\hat{h}^2_{g\psi,local,i}}}, \tag{4.21}$$

where $\hat{h}^2_{g\phi,local,i}$ and $\hat{h}^2_{g\psi,local,i}$ denote the local SNP-heritability of trait ϕ and ψ at the i-th region. In some cases, this estimator of local genetic correlation may yield an estimate with magnitude greater than 1, and we cap the estimate at -1 or 1. In simulations, we show that $\hat{r}_{g,local,i}$ is approximately unbiased when both traits are heritable at the i-th region. In practice, however, the terms, $\hat{h}^2_{g\phi,local,i}$ and $\hat{h}^2_{g\psi,local,i}$, can be close to zero, greatly inflating the standard error of $\hat{r}_{g,local,i}$. Thus, we recommend estimating local genetic correlation only at regions with significant local SNP-heritability. One can also estimate local genetic correlation at a set of regions. For example, to estimate genetic correlation at regions indexed by the index set $\boldsymbol{C}$, one applies the following formula,

$$\hat{r}_{g,\boldsymbol{C}} = \frac{\sum_{i\in\boldsymbol{C}} \hat{\rho}_{g,local,i}}{\sqrt{\sum_{i\in\boldsymbol{C}} \hat{h}^2_{\phi,g,local,i}}\sqrt{\sum_{i\in\boldsymbol{C}} \hat{h}^2_{\psi,g,local,i}}}, \tag{4.22}$$

We estimate standard error of local genetic correlation at a single region through a parametric bootstrap approach [39] and local genetic correlation at a set of regions through jackknife.

4.2.8 Simulation framework

Starting from half (202 individuals) of the EUR reference panel from the 1000 Genomes Project[28], we simulated genotype data for 50,000 individuals at HapMap3[51] SNPs with minor allele frequency (MAF) greater than 5% in 100 randomly selected LD-independent regions defined in ref[12] on chromosome 1 using HAPGEN2[51]. We used the other half of the EUR reference panel (203 individuals) to obtain external reference LD matrices.

We simulated phenotypes from the genotypes according to the linear model $\boldsymbol{\phi} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and

$\boldsymbol{\psi} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\delta}$, where $\boldsymbol{X}$ is the column-standardized genotype matrix. We drew the effects of causal SNPs ($\boldsymbol{\beta}_C$, $\boldsymbol{\gamma}_C$) from the distribution

$$N\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{0}\end{array}\right], \left[\begin{array}{cc}\frac{h^2_{g\phi}}{|C|}\boldsymbol{I} & \frac{\rho_e}{|C|}\boldsymbol{I}\\ \frac{\rho_e}{|C|}\boldsymbol{I} & \frac{h^2_{g\psi}}{|C|}\boldsymbol{I}\end{array}\right]\right), \tag{4.23}$$

where C is the index set of causal SNPs, and set the effects of all other SNPs to be zero. We then drew ($\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$) from the distribution

$$N\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{0}\end{array}\right], \left[\begin{array}{cc}(1-h^2_{g\phi})\boldsymbol{I} & \rho_e\boldsymbol{I}\\ \rho_e\boldsymbol{I} & (1-h^2_{g\psi})\boldsymbol{I}\end{array}\right]\right). \tag{4.24}$$

Finally, we simulated GWAS summary statistics using methods outlined in previous sections. For each $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ drawn from the normal distribution, we simulated 1000 sets of summary statistics by varying $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, and applied ρ-HESS to estimate genetic covariance and genetic correlation for each set of the simulated summary statistics.

4.2.9 Empirical data sets

We obtained GWAS summary data for 36 quantitative complex traits and diseases from 15 GWAS consortia or institutions (see Table 4.1), all of which are based on individuals of European ancestry, and have sample size greater than 20,000. We used approximately independent genomic regions defined in ref[12] to partition the genome, and restricted our analyses on HapMap3 SNPs with minor allele frequency (MAF) greater than 5% in the European population in the 1000 Genomes data [28]. We also removed stand-ambiguous SNPs prior to our analyses. We follow the method outlined in ref[140] to estimate and re-inflate λ_{gc}, and to choose the number of eigenvectors to include in estimating local genetic covariance and SNP-heritability.

4.2.10 Local genetic correlation at regions ascertained for GWAS signals

Recent works leverage the difference in correlations of Z-scores at genomic regions ascertained for GWAS signals specific to each trait to prioritize putative causal models between pairs of complex traits [125, 98]. We evaluated the local genetic correlation at regions harboring GWAS signals specific to each trait across all 298 pairs of traits exhibiting significant genome-wide genetic correlation. We estimate local genetic correlations only for pairs of traits for which the number of loci harboring GWAS hits specific to each trait is greater than 10. The confidence intervals (1.96 times jackknife standard error on each side) of the ascertained local genetic correlations ($\hat{r}_{g,local,trait1}$ and $\hat{r}_{g,local,trait2}$) do not overlap; one of the confidence intervals overlap with 0 and the other does not.

4.3 Results

4.3.1 Local genetic correlation estimation in simulations

We evaluated the performance of our approach (ρ-HESS) through simulations across a wide-range of disease architectures. We included cross-trait LDSC [18], an approach that assumes a random-effect model, in the comparison for completeness purposes. When LD is estimated in-sample, ρ-HESS provides an unbiased estimate of local genetic covariance and nearly unbiased estimates of genetic correlation (i.e. genetic covariance divided with the square root of local SNP heritability, see Methods). Next, we quantified the performance in the more realistic case when in-sample LD is unavailable and needs to be estimated from external reference panels. Although both cross-trait LDSC and ρ-HESS provide accurate estimates of genetic correlation, we observe superior accuracy with higher precision for ρ-HESS (Figure 4.4, S4, S6, S7). We attribute the lower standard error of ρ-HESS to the truncated-SVD regularization of the LD matrix which effectively reduces the degree of freedom of the bi-linear form in Equation (4.7). Different genomic regions vary in their total amount of LD and we observed that the accuracy of genetic correlation estimation decreases with the total

amount of regional LD. This is expected as high LD regions lead to high rank deficiencies in the LD matrix and small eigenvalues, thus increasing the level of statistical noise in the estimation. We also evaluated the performance of local genetic correlation estimation in simulations where we varied the number of causal variants in each region. Overall, we observe that our estimator of genetic covariance and correlation is not sensitive to the underlying polygenicity (i.e. number of causal SNPs) (Figure 4.4 S5, S8, S9). Finally, we also evaluated the performance of the estimator when causal variants are all drawn from DHS regions[157], and observed that the performance is not sensitive to the uneven distribution of causal variants.

4.3.2 Local genetic correlation across 36 quantitative traits

We analyzed GWAS summary data from 36 complex traits to obtain local genetic correlations at 1,703 approximately LD-independent regions in the genome ($\sim$1.6Mb in width on average)[12]. First, as a quality control step, we aggregated the local estimates into genome-wide estimates of genetic correlation (see Methods) and compared to the cross-trait LDSC estimates. Reassuringly, we find a high degree of consistency with genetic correlations estimated by cross-trait LDSC regression ($R = 0.77$, Figure 4.5, S13). Our estimator provides lower standard errors as compared to cross-trait LDSC (likely due to the truncated-SVD regularization procedure), and yields consistently lower estimates for pairs of traits from the same consortium where we conservatively assume full sample overlap (see Discussion). Overall, we identify 298 pairs of traits with significant genome-wide genetic correlation ($p < 0.05/630$). These include previously reported correlations, e.g. body mass index (BMI) and triglyceride (TG), as well as complex traits that have not been studied before using genetic correlation, e.g. red blood cell count (RBC) and fasting insulin (FI) (Figure 4.5).

Next, we searched for genomic regions that disproportionately contribute to the genetic correlation of the 36 analyzed traits; we excluded the HLA region due to complex LD patterns. We identify 25 genomic regions that show both significant local genetic correlation (two-tailed $p < 0.05/1703/630$) as well as significant local SNP-heritability (one-tailed $p < 0.05/1703/36$)

(see Table 4.2). For example, the estimate of local genetic correlation between HDL and TG at chr11:116-117Mb is -0.82 (95% CI [-0.95, -0.69]), suggesting highly shared genetic architecture at this region for HDL and TG. Indeed, the region chr11:116-117M harbors *APOA1* (MIM 107680), which is known to be associated with multiple lipid traits [29]. Interestingly, 4 out of the 25 regions do not contain GWAS significant SNPs ($p < 5e-8$) for either one or both traits and can be viewed as new risk regions for these traits.

Since genetic correlation is an aggregation of local genetic covariance, for pairs of traits with highly positive or negative genetic correlation, we expect the distribution of local genetic covariances to be shifted towards the positive or negative side (see Figure 4.6); whereas for pairs of traits with low genetic correlation, we expect the distribution of local genetic covariances to be centered around zero (see Figure 4.7, 4.8). Indeed, pairs of traits with higher genome-wide genetic correlation tend to harbor more loci with significant local genetic covariance. For instance, only one region exhibits significant local genetic covariance for the pair of traits age at menarche (AM) and height ($r_g = 0.13$, 95% CI [0.10, 0.13]), whereas four loci show significant local genetic covariance for the pair of traits LDL and TG ($r_g = 0.45$, 95% CI [0.42, 0.49]).

4.3.3 Local correlations for pairs of traits with negligible genome-wide correlation

Several pairs of traits show negligible genome-wide genetic correlation although they share GWAS risk regions. For example HDL and LDL share several GWAS risk loci[29] but the genome-wide genetic correlation is negligible (-0.05, 95% CI [-0.09, -0.01]) [18]. The absence of significant genome-wide genetic correlation between these pairs of traits can be attributed to either symmetric distribution of local genetic covariance (positive local genetic covariance cancels out negative local genetic covariance, see Figure 4.1) and/or lack of power to declare significance for genome-wide genetic correlation. Thus, we hypothesize that at the region-specific level, many loci may manifest significant local genetic covariance even if the genome-wide genetic correlation between a pair of traits is not significant. Indeed, 11 genomic regions

show significant local genetic correlation (two-tailed $p < 0.05/1703$) for HDL and LDL (see Figure 4.7). Some of these loci, e.g. chr2:21M-23M, chr11:116M-117M, and chr19:44M-46M, harbor *APOB*, *APOA1*, and *APOE* (MIM 107741), respectively, which are known to be involved in lipid genetics[48, 118, 29]. Across all pairs of traits with non-significant genome-wide correlation, we identify 6 regions across 10 pairs of traits with significant local genetic correlation (two-tailed $p < 0.05/1703/630$) and local SNP-heritability (one-tailed $p < 0.05/1703/36$) (see Table 4.2). For example the region chr6:134-136M harbors *HBS1L* (MIM 612450) [148, 160], and contributes to local genetic covariance across many blood traits (MCH, MCV, RBC, and PLT).

4.3.4 Genetic correlation ascertained for GWAS risk loci

Assessing the correlation in the effects at genomic regions ascertained for trait-specific GWAS regions can be used to prioritize putative causal models between complex traits. We utilized a recently proposed approach[125] to assign putative causal relation to 55 pairs of traits. Restricting to 40 of the 55 pairs of traits that contain at least 10 regions with trait-specific GWAS signals (see Methods), we quantified the local genetic correlation at genomic regions containing GWAS loci specific to each trait (see Table 4.4, Figure 4.9). Overall, the local genetic correlation is highly consistent with the putative causal relationships inferred by correlating the top signals at these loci[125]. For example, when considering body mass index (BMI) and triglyceride levels (TG), the correlation at BMI-specific regions is significantly greater than TG-specific loci ($\hat{r}_{g,local,BM} = 0.47$ 95% CI [0.37, 0.57] vs. $\hat{r}_{g,local,TG} = -0.02$ 95% [-0.14, 0.10]), indicating that loci that increase BMI tend to consistently increase TG, whereas loci that increase TG do not consistently affect BMI, consistent with the putative model that BMI causally increases TG [125, 98]. We also observe correlations consistent with a model in which years of education (EY) consistently decreases hemoglobin level (HB), LDL, TG (see Table 4.4), in line with previous conclusions on the effect of education on health [143, 7]. However, we note that education attainment (or other studied traits) may be confounded by other factors such as social status and that one should exercise caution

when inferring causality from genetic data. Finally, we also report pairs of traits in which the genetic correlation approach attains different results form bi-directional regression on the top signals[125]. For example, when considering body mass index (BMI) and age at menarche (AM), the local correlation approach do not yield different estimates ($r_{g,local,BMI} = -0.49$ 95% CI [-0.63, -0.35] vs $r_{g,local,AM} = -0.47$ 95% CI [-0.59, -0.35]), whereas the approach of ref [125] suggests a putative causal relation. This discrepancy can be due to different model assumptions, e.g., single causal variant versus allelic heterogeneity, with further investigations needed to assign causality from these data.

4.4 Discussion

We have described ρ-HESS, a method to estimate local genetic correlation from GWAS summary association data. Through extensive simulations, we demonstrated that our method is approximately unbiased and provides consistent results irrespective of causal architecture. We analyzed large-scale GWAS summary association data of 36 quantitative traits. Compared with cross-trait LDSC, our methods identified considerably more pairs of traits displaying significant genome-wide genetic correlation likely because of the truncated-SVD regularization of the LD matrix, which decreases the standard error of the estimates. We identify genomic regions that are significantly correlated across pairs of traits regardless of the significance of genome-wide correlation. Finally, we performed bi-directional analyses over the local genetic correlations to identify putative causal relationships, and report local genetic correlations at loci harboring GWAS signal specific to each trait.

We conclude with several limitations highlighting areas for future work. First, our estimator requires phenotype correlation between two traits, as well as the number of shared individuals between the two GWASs. We estimate the phenotype correlation through cross-trait LDSC assuming full sample overlap between GWAS within the same consortium and no sample overlap between GWAS across two consortia. Second, we note that our bi-directional analyses over local genetic correlation can be further extrapolated to infer putative causal

models between complex traits. We refrain from making conclusive causal inferences from the bi-directional analyses because exact inference of causal relations is largely complicated by unobserved confounders such as socioeconomic status, population stratification and/or biological pathways. Furthermore, most of the GWAS summary association data are adjusted for covariates such as age, gender, to increase statistical power [104], and previous works have shown that adjusting for covariates can potentially lead to false positives [4]. Third, in our real data analyses, we made the assumption that the loci are independent of each other. In reality however correlations may exist across adjacent loci due to long range LD, and can lead to biased estimates. Nevertheless, we note that previous works have indicated the effect of LD leakage to be minimal [91, 140], and we conjecture that this statement still hold in estimating local genetic correlation. Lastly, we use truncated-SVD to regularize LD matrix and to reduce standard error in the estimates of local genetic correlation, at the cost of introducing bias. Currently, we use a fixed number of eigenvectors in the truncated-SVD regularization, across all the loci. However, this approach may not be optimal for genomic regions with different LD structure, and leave a principled approach of estimating the number of eigenvectors as future work.

4.5 Appendix

4.5.1 Quantifying shared genetics via covariance of the causal effects

An alternative measure of shared genetics is the covariance of the causal effects ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) of the two traits. Under the fixed-effect model, we define covariance of the causal effects, $\rho_{g,causal}$, as the dot product between the causal effect size vectors of the two traits,

$$\rho_{g,causal} = \boldsymbol{\beta}^{\intercal}\boldsymbol{\gamma}. \tag{4.25}$$

Here, we make the assumption that the average effect size of each SNP is 0.

The definition of covariance of the causal effects in Equation (4.25) coincides with genetic

covariance under the random-effect model. As shown in the supplementary of [18], if one assumes that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have zero mean and

$$\mathrm{V}ar[(\boldsymbol{\beta},\boldsymbol{\gamma})] = \frac{1}{p}\begin{bmatrix} h^2_{g\phi} & \rho_g \\ \rho_g & h^2_{g\psi} \end{bmatrix}, \tag{4.26}$$

then it can be shown that the genetic covariance between two traits is

$$\mathrm{C}ov[\boldsymbol{x}^\intercal\boldsymbol{\beta}, \boldsymbol{x}^\intercal\boldsymbol{\gamma}] = \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}[\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{\beta}_i\boldsymbol{\gamma}_j] = \sum_{i=1}^{p}\mathrm{E}[\boldsymbol{x}_i^2\boldsymbol{\beta}_i\boldsymbol{\gamma}_i] = \sum_{i=1}^{p}\mathrm{E}[\boldsymbol{x}_i^2]\,\mathrm{E}[\boldsymbol{\beta}_i\boldsymbol{\gamma}_i] = \rho_g. \tag{4.27}$$

The random-effect model makes the implicit assumption that many SNPs are causal, which is appropriate for genome-wide analysis but not for local analysis, where few SNPs are likely to be causal.

4.5.2 Estimating covariance of the causal effects from GWAS summary data

For completeness, we derive an estimator for $\rho_{g,causal}$. We assume a linear model for the two traits (see Methods). The effect size estimates from GWAS, $\hat{\boldsymbol{\beta}}_{gwas}$ and $\hat{\boldsymbol{\gamma}}_{gwas}$, follow $\hat{\boldsymbol{\beta}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\beta}, \frac{1-h^2_\phi}{n_1}\boldsymbol{V}\right)$ and $\hat{\boldsymbol{\gamma}}_{gwas} \sim N\left(\boldsymbol{V}\boldsymbol{\gamma}, \frac{1-h^2_\phi}{n_2}\boldsymbol{V}\right)$, with $\mathrm{C}ov[\hat{\boldsymbol{\beta}}_{gwas}, \hat{\boldsymbol{\gamma}}_{gwas}] = \frac{\rho_e n_s}{n_1 n_2}\boldsymbol{V}$, where n_1 and n_2 are the sample size for the two GWASs, and n_s is the number of shared samples (see Methods).

As the sample size, n_1 and n_2, of the two GWASs go to infinity, we have $\boldsymbol{\beta}_{gwas} = \lim_{n\to\infty}\hat{\boldsymbol{\beta}}_{gwas} = \boldsymbol{V}\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_{gwas} = \lim_{n\to\infty}\hat{\boldsymbol{\gamma}}_{gwas} = \boldsymbol{V}\boldsymbol{\gamma}$, which implies $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\boldsymbol{\beta}_{gwas}$ and $\boldsymbol{\gamma} = \boldsymbol{V}^{-1}\boldsymbol{\gamma}_{gwas}$, suggesting the following estimator for covariance of the causal effects,

$$\rho_{g,causal} = \boldsymbol{\beta}^\intercal\boldsymbol{\gamma} = \boldsymbol{\beta}^\intercal_{gwas}\boldsymbol{V}^{-2}\boldsymbol{\gamma}_{gwas}. \tag{4.28}$$

In reality, however, finite sample sizes of GWAS results in noise in the estimates of $\boldsymbol{\beta}$ and

$\boldsymbol{\gamma}$, creating bias in the estimate of $\rho_{g,causal}$. From bilinear form theory, it can be shown that

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{V}^{-2}\hat{\boldsymbol{\gamma}}_{gwas}] = \boldsymbol{\beta}^{\intercal}\boldsymbol{\gamma} + \frac{\rho_e}{n}\,\mathrm{t}r(\boldsymbol{V}^{-2}\boldsymbol{V}) = \boldsymbol{\beta}^{\intercal}\boldsymbol{\gamma} + \frac{\rho_e}{n}\,\mathrm{t}r(\boldsymbol{V}^{-1}), \tag{4.29}$$

suggesting the unbiased estimator of $\rho_{g,causal}$,

$$\hat{\rho}_{g,causal} = \hat{\boldsymbol{\beta}}_{gwas}^{\intercal}\boldsymbol{V}^{-2}\hat{\boldsymbol{\gamma}}_{gwas} - \frac{n_s\rho_e}{n_1 n_2}\,\mathrm{t}r(\boldsymbol{V}^{-1}), \tag{4.30}$$

where the environmental covariance can be estimated through cross-trait LD Score regression [18].

4.6 Tables

Table 4.1: **A summary of the 36 GWAS summary data sets analyzed.**

Trait Name	**Abbreviation**	**Consortium**	**# gen corr all consortium**	**# gen corr outside consortium**	**Approx. sample size**
Age at Menarche [124]	AM	REPROGEN	21 (4)	21 (4)	133K
Body Mass Index [90]	BMI	GIANT	27 (17)	23 (14)	231K
Height [165]	HEIGHT	GIANT	17 (2)	13 (1)	241K
Hip Circumference [142]	HIP	GIANT	23 (14)	19 (10)	144K
Waist Circumference [142]	WC	GIANT	26 (18)	22 (15)	153K
Waist-to-hip Ratio [142]	WHR	GIANT	27 (19)	23 (16)	143K
Haemoglobin [160]	HB	HAEMGEN	21 (10)	18 (8)	51K
Mean Cell Haemoglobin [160]	MCH	HAEMGEN	9 (1)	8 (1)	44K
MCH Concentration [160]	MCHC	HAEMGEN	6 (4)	2 (1)	47K
Mean Cell Volume [160]	MCV	HAEMGEN	12 (3)	10 (1)	49K
Packed Cell Volume [160]	PCV	HAEMGEN	18 (11)	14 (8)	45K
Red Blood Cell Count [160]	RBC	HAEMGEN	20 (10)	17 (8)	46K
Number of Platelets [52]	PLT	HAEMGEN	9 (1)	6 (1)	67K
Fasting Glucose [38]	FG	MAGIC	19 (9)	16 (8)	46K
Fasting Insulin [38]	FI	MAGIC	20 (12)	18 (12)	46K
HBA1C [147]	HBA1C	MAGIC	19 (14)	18 (13)	46K
HOMA-B [38]	HOMA-B	MAGIC	17 (11)	15 (11)	46K
HOMA-IR [38]	HOMA-IR	MAGIC	21 (12)	21 (12)	46K
High Density Lipoprotein [29]	HDL	LIPID	23 (12)	21 (11)	96K
Low Density Lipoprotein [29]	LDL	LIPID	19 (6)	17 (4)	91K
Total Cholesterol [29]	TC	LIPID	18 (3)	15 (1)	96K
Triglycerides [29]	TG	LIPID	26 (14)	23 (11)	92K
Forearm BMD [176]	FA	GEFOS	4 (1)	2 (0)	53K
Femoral Neck BMD [176]	FN	GEFOS	4 (2)	2 (0)	53K
Lumbar Spine BMD [176]	LS	GEFOS	7 (1)	5 (0)	53K
Education Years [116]	EY	SSGAC	26 (5)	24 (4)	294K
Neuroticism [115]	NEURO	SSGAC	5 (2)	3 (0)	171K
Subjective Well-being [115]	SWB	SSGAC	4 (1)	2 (0)	298K
Age First Birth [8]	AFB	BIOS	23 (5)	23 (5)	251K
Birth Weight [63]	BW	EGG	13 (1)	13 (1)	68K
Urinary Albumin-to-Creatinine Ratio [155]	UACR	DCCT-EDIC	11 (1)	11 (1)	53K
Rest Heart Rate [42]	HR	EPPINGA	14 (0)	14 (0)	265K
Serum Urate Concentrations [76]	URATE	GUGC	25 (14)	25 (14)	107K
Body Fat [93]	BF	Lu	26 (17)	26 (17)	58K
Extra-Glomerular Filtration Rate of Creatinin [123]	CRN	CKDGEN	10 (1)	10 (1)	133K
Age at Menopause [33]	MP	BCAC	6 (0)	6 (0)	70K

We list the total number of traits with significant non-zero genome-wide genetic correlation (two-tailed $p < 0.05/630$) and the total number of traits outside the consortium with significant non-zero genome-wide genetic correlation in the fourth and fifth column, respectively. Number of traits for which the magnitude of genetic correlation is both significantly non-zero and greater than 0.2 is shown in parentheses.

Table 4.2: **Loci that show significant local genetic covariance (two-tailed $p < 0.05/1703/630$) and local SNP heritability (one-tailed $p < 0.05/1703/36$) for both traits.**

Trait1	Trait2	Locus	$h^2_{g,local,trait1}$	$h^2_{g,local,trait2}$	$r_{g,local}$
AM	HEIGHT	chr9:107-109M	0.15 (0.02)	0.05 (0.01)	0.61 ([0.34,0.87])
BMI	HIP	chr16:53-55M	0.22 (0.02)	0.19 (0.03)	0.99 ([0.76,1.00])
BMI	HIP	chr18:57-59M	0.14 (0.02)	0.13 (0.02)	0.99 ([0.71,1.00])
BMI	WC	chr16:53-55M	0.22 (0.02)	0.21 (0.03)	1.00 ([0.78,1.00])
BMI	WC	chr18:57-59M	0.14 (0.02)	0.13 (0.02)	1.00 ([0.72,1.00])
BW	HEIGHT	chr12:65-67M	0.14 (0.02)	0.23 (0.02)	0.93 ([0.70,1.00])
HDL	TG	chr2:21-23M	0.16 (0.03)	0.22 (0.03)	-0.94 ([-1.00,-0.65])
HDL	TG	chr8:19-20M	0.65 (0.04)	0.82 (0.04)	-1.00 ([-1.00,-0.91])
HDL	TG	chr11:116-117M	0.40 (0.04)	1.27 (0.06)	-0.82 ([-0.95,-0.69])
HDL	TG	chr15:58-59M	1.18 (0.06)	0.18 (0.03)	0.89 ([0.68,1.00])
HEIGHT	HIP	chr16:4-5M	0.06 (0.01)	0.10 (0.02)	0.73 ([0.41,1.00])
HIP	WC	chr16:53-55M	0.19 (0.03)	0.21 (0.03)	0.99 ([0.73,1.00])
HIP	WC	chr18:57-59M	0.13 (0.02)	0.13 (0.02)	1.00 ([0.69,1.00])
LDL	TG	chr1:61-63M	0.14 (0.03)	0.28 (0.03)	0.98 ([0.67,1.00])
LDL	TG	chr2:21-23M	0.84 (0.05)	0.22 (0.03)	0.62 ([0.46,0.78])
LDL	TG	chr8:126-128M	0.16 (0.03)	0.32 (0.04)	0.94 ([0.63,1.00])
LDL	TG	chr19:18-19M	0.18 (0.03)	0.21 (0.03)	0.99 ([0.72,1.00])
PLT	RBC	chr6:134-136M	0.26 (0.05)	0.66 (0.09)	-0.99 ([-1.00,-0.69])
HDL	HEIGHT	chr11:47-49M	0.17 (0.02)	0.07 (0.01)	0.61 ([0.42,0.80])
HDL	LDL	chr2:21-23M	0.16 (0.03)	0.84 (0.05)	-0.56 ([-0.74,-0.39])
HDL	LDL	chr8:9-9M	0.14 (0.02)	0.12 (0.02)	0.99 ([0.70,1.00])
MCH	MCV	chr6:24-25M	0.49 (0.07)	0.37 (0.06)	0.97 ([0.67,1.00])
MCH	MCV	chr6:134-136M	0.86 (0.09)	0.70 (0.08)	0.98 ([0.76,1.00])
MCH	PLT	chr6:134-136M	0.86 (0.09)	0.26 (0.05)	1.00 ([0.72,1.00])
MCH	RBC	chr6:134-136M	0.86 (0.09)	0.66 (0.09)	-0.98 ([-1.00,-0.75])
MCV	PLT	chr6:134-136M	0.70 (0.08)	0.26 (0.05)	1.00 ([0.72,1.00])
MCV	RBC	chr6:134-136M	0.70 (0.08)	0.66 (0.09)	-0.98 ([-1.00,-0.74])
MP	HEIGHT	chr5:175-177M	0.31 (0.04)	0.10 (0.01)	-0.63 ([-0.82,-0.45])
URATE	MCH	chr6:24-25M	0.13 (0.02)	0.53 (0.07)	0.56 ([0.33,0.79])
URATE	MCV	chr6:24-25M	0.13 (0.02)	0.41 (0.06)	0.66 ([0.39,0.92])

We list pairs of traits for which the genome-wide genetic correlation is significant (two-tailed $p < 0.05/630$) and negligible in top and bottom half of this table, respectively. Here, we focus only on the pairs of traits excluding TC (see Table 4.3 for pairs of traits involving TC). Numbers in parentheses represent standard errors for local SNP heritability estimates and 95% confidence intervals for local genetic correlation estimates.

Table 4.3: **Loci that show significant local genetic covariance (two-tailed $p < 0.05/1703/630$) and local SNP heritability (one-tailed $p < 0.05/1703/36$) for both traits.**

Trait1	**Trait2**	**Locus**	$\mathbf{h^2_{g,local,trait1}}$	$\mathbf{h^2_{g,local,trait2}}$	$\mathbf{r_{g,local}}$
HDL	TC	chr2:21-23M	0.16 (0.03)	0.63 (0.04)	-0.50 ([-0.67,-0.33])
HDL	TC	chr8:9-9M	0.14 (0.02)	0.16 (0.02)	0.98 ([0.74,1.00])
HDL	TC	chr9:107-109M	0.28 (0.03)	0.19 (0.03)	0.90 ([0.65,1.00])
HDL	TC	chr11:116-117M	0.40 (0.04)	0.27 (0.03)	-0.41 ([-0.55,-0.27])
HDL	TC	chr15:58-59M	1.18 (0.06)	0.31 (0.03)	0.98 ([0.83,1.00])
LDL	TC	chr1:61-63M	0.14 (0.03)	0.28 (0.03)	0.97 ([0.67,1.00])
LDL	TC	chr1:108-110M	0.74 (0.05)	0.52 (0.04)	1.00 ([0.88,1.00])
LDL	TC	chr2:21-23M	0.84 (0.05)	0.63 (0.04)	1.00 ([0.87,1.00])
LDL	TC	chr2:43-44M	0.31 (0.03)	0.31 (0.03)	1.00 ([0.91,1.00])
LDL	TC	chr5:73-75M	0.28 (0.03)	0.24 (0.03)	1.00 ([0.76,1.00])
LDL	TC	chr5:155-156M	0.11 (0.02)	0.13 (0.02)	0.98 ([0.58,1.00])
LDL	TC	chr8:9-9M	0.12 (0.02)	0.16 (0.02)	1.00 ([0.72,1.00])
LDL	TC	chr8:126-128M	0.16 (0.03)	0.19 (0.03)	0.98 ([0.61,1.00])
LDL	TC	chr16:71-72M	0.19 (0.03)	0.19 (0.03)	0.99 ([0.70,1.00])
LDL	TC	chr19:9-11M	0.49 (0.04)	0.33 (0.03)	1.00 ([0.81,1.00])
LDL	TC	chr19:18-19M	0.18 (0.03)	0.26 (0.03)	0.99 ([0.74,1.00])
LDL	TC	chr19:44-46M	0.77 (0.05)	0.43 (0.04)	1.00 ([0.86,1.00])
TC	TG	chr1:61-63M	0.28 (0.03)	0.28 (0.03)	0.99 ([0.77,1.00])
TC	TG	chr2:21-23M	0.63 (0.04)	0.22 (0.03)	0.60 ([0.43,0.76])
TC	TG	chr8:126-128M	0.19 (0.03)	0.32 (0.04)	0.96 ([0.69,1.00])
TC	TG	chr11:116-117M	0.27 (0.03)	1.27 (0.06)	0.89 ([0.73,1.00])
TC	TG	chr15:58-59M	0.31 (0.03)	0.18 (0.03)	0.97 ([0.69,1.00])
TC	TG	chr19:18-19M	0.26 (0.03)	0.21 (0.03)	0.98 ([0.75,1.00])

Here, we focus only on the pairs of traits involving TC. The genome-wide genetic correlation of each pair of traits is also significant (two-tailed $p < 0.05/630$). Numbers in parentheses represent standard errors for local SNP heritability estimates and 95% confidence intervals for local genetic correlation estimates.

Table 4.4: **Bi-directional analysis of local genetic correlation identifies 40 pairs of traits for which one is likely a causal factor of the other.**

Trait1	$\hat{\mathbf{r}}_{\mathbf{g,local,trait1}}$	**No. loci**	**Trait2**	$\hat{\mathbf{r}}_{\mathbf{g,local,trait2}}$	**No. loci**	**Direction**	**Ratio**
AM	-0.47 (0.06)	54	BMI	-0.49 (0.07)	51	BMI ↓ AM	0.00e+00
AM	0.26 (0.05)	39	HEIGHT	0.09 (0.02)	429	AM ↑ HEIGHT	0.00e+00
AM	-0.18 (0.05)	60	HIP	-0.26 (0.08)	36	HIP ↓ AM	7.00e-05
AM	-0.23 (0.05)	58	WC	-0.36 (0.09)	28	WC ↓ AM	1.09e-04
BMI	-0.25 (0.06)	60	EY	-0.35 (0.04)	133	EY ↓ BMI	0.00e+00
BMI	-0.47 (0.05)	57	HDL	-0.18 (0.04)	81	BMI ↓ HDL	0.00e+00
BMI	-0.02 (0.04)	39	HEIGHT	-0.16 (0.02)	432	HEIGHT ↓ BMI	0.00e+00
BMI	0.95 (0.02)	32	HIP	0.77 (0.11)	11	BMI ↑ HIP	0.00e+00
BMI	0.47 (0.05)	59	TG	-0.02 (0.06)	60	BMI ↑ TG	0.00e+00
URATE	0.07 (0.08)	28	BMI	0.55 (0.05)	64	BMI ↓ URATE	1.80e-05
BMI	0.69 (0.03)	58	WHR	0.13 (0.13)	22	BMI ↑ WHR	0.00e+00
BW	-0.22 (0.05)	41	URATE	-0.08 (0.09)	28	URATE ↓ BW	0.00e+00
URATE	-0.13 (0.05)	22	CRN	-0.36 (0.08)	36	URATE ↓ CRN	1.00e-06
CRN	0.04 (0.06)	41	WHR	0.07 (0.09)	27	CRN ↓ WHR	2.06e-04
EY	-0.19 (0.05)	138	HB	-0.05 (0.10)	15	EY ↓ HB	1.00e-06
EY	0.22 (0.03)	134	HDL	0.08 (0.03)	85	EY ↑ HDL	3.91e-04
EY	0.16 (0.03)	100	HEIGHT	0.16 (0.02)	420	HEIGHT ↑ EY	0.00e+00
EY	-0.24 (0.04)	133	NEURO	-0.14 (0.11)	11	EY ↓ NEURO	0.00e+00
EY	-0.20 (0.03)	134	TG	-0.05 (0.05)	62	EY ↓ TG	0.00e+00
EY	-0.30 (0.03)	134	WC	-0.25 (0.08)	34	EY ↓ WC	0.00e+00
EY	-0.34 (0.03)	136	WHR	-0.17 (0.06)	27	EY ↓ WHR	0.00e+00
HDL	-0.12 (0.04)	81	HIP	-0.50 (0.05)	36	HIP ↓ HDL	2.00e-06
HDL	-0.51 (0.07)	52	TG	-0.48 (0.10)	29	HDL ↓ TG	1.00e-06
HDL	-0.25 (0.04)	82	WC	-0.64 (0.05)	31	WC ↓ HDL	0.00e+00
HDL	-0.27 (0.06)	79	WHR	-0.59 (0.12)	19	WHR ↓ HDL	6.82e-04
HEIGHT	0.44 (0.01)	432	HIP	0.25 (0.06)	18	HEIGHT ↑ HIP	0.00e+00
HEIGHT	-0.09 (0.02)	420	LDL	-0.01 (0.04)	30	HEIGHT ↓ LDL	0.00e+00
HEIGHT	-0.14 (0.02)	446	PLT	-0.08 (0.13)	17	HEIGHT ↓ PLT	0.00e+00
HEIGHT	-0.12 (0.02)	415	TC	-0.05 (0.05)	40	HEIGHT ↓ TC	0.00e+00
HEIGHT	-0.08 (0.02)	429	TG	-0.05 (0.05)	37	HEIGHT ↓ TG	1.00e-06
HEIGHT	0.30 (0.02)	443	WC	0.17 (0.05)	23	HEIGHT ↑ WC	0.00e+00
HIP	0.26 (0.07)	41	TG	-0.09 (0.06)	63	HIP ↑ TG	4.59e-04
HIP	0.44 (0.08)	37	WHR	-0.14 (0.09)	22	HIP ↑ WHR	7.00e-06
LDL	0.93 (0.03)	11	TC	0.80 (0.08)	26	TC ↑ LDL	0.00e+00
URATE	0.05 (0.05)	29	MP	-0.02 (0.05)	62	URATE ↑ MP	1.33e-04
NEURO	0.08 (0.11)	17	PLT	-0.04 (0.07)	30	NEURO ↑ PLT	4.00e-06
TG	0.09 (0.05)	62	WC	0.56 (0.06)	34	WC ↑ TG	0.00e+00
TG	0.28 (0.05)	59	WHR	0.57 (0.10)	22	WHR ↑ TG	6.67e-04
URATE	0.07 (0.05)	30	WC	0.69 (0.06)	39	WC ↓ URATE	9.70e-05
WC	0.80 (0.02)	29	WHR	0.60 (0.07)	20	WC ↑ WHR	0.00e+00

Here, "Trait 1" and "Trait 2" refer to the trait for which the GWAS hit loci were ascertained in the bi-directional analysis. Traits that are likely a causal factor of the other are marked with stars. Numbers in parentheses represent standard errors of the local genetic correlation estimates.

4.7 Figures

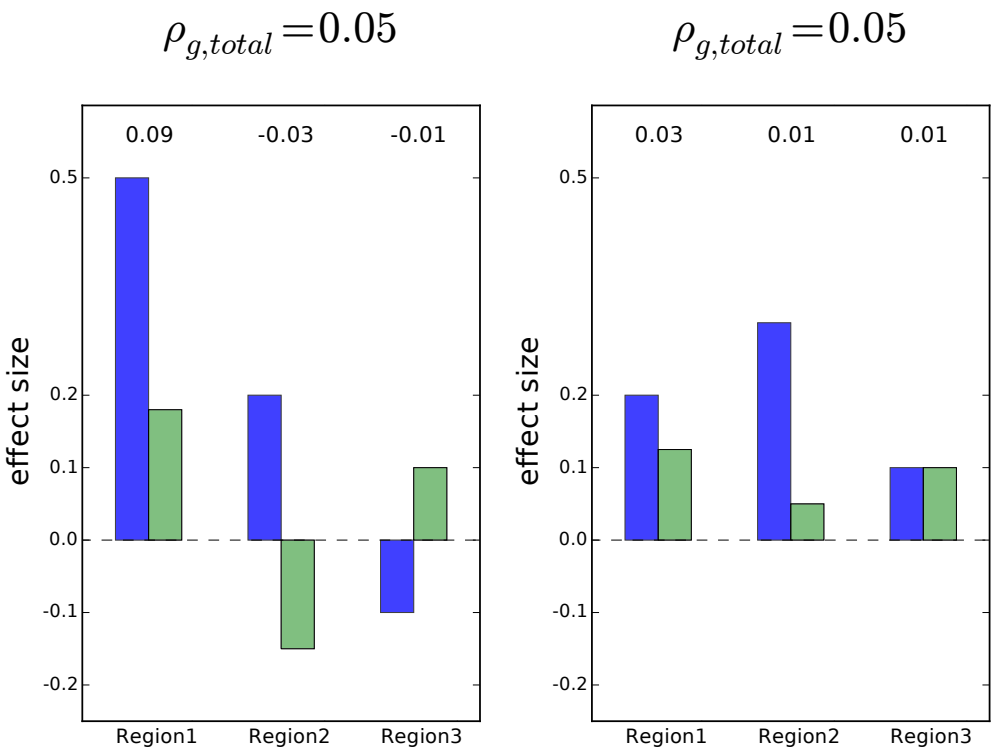


Figure 4.1: **Examples of two different distributions of local genetic covariances (shown at the top of each bar) that result in the same total genetic covariance ($\rho_{g,total}$ = 0.05).** In the left example, the total genetic covariance is a summation of a large positive local genetic covariance at Region1 and two smaller negative local genetic covariances at Region2 and Region3 (e.g, Regions 2 and 3 impact traits through a different pathway than Region1). In the right example, the total genetic covariance is a summation of small positive local genetic covariances (e.g., all three regions impact both traits through the same pathway). Positive local genetic covariance can be interpreted as a locus driving a pathway that regulates two traits in the same direction, and negative local genetic covariance the opposite direction.

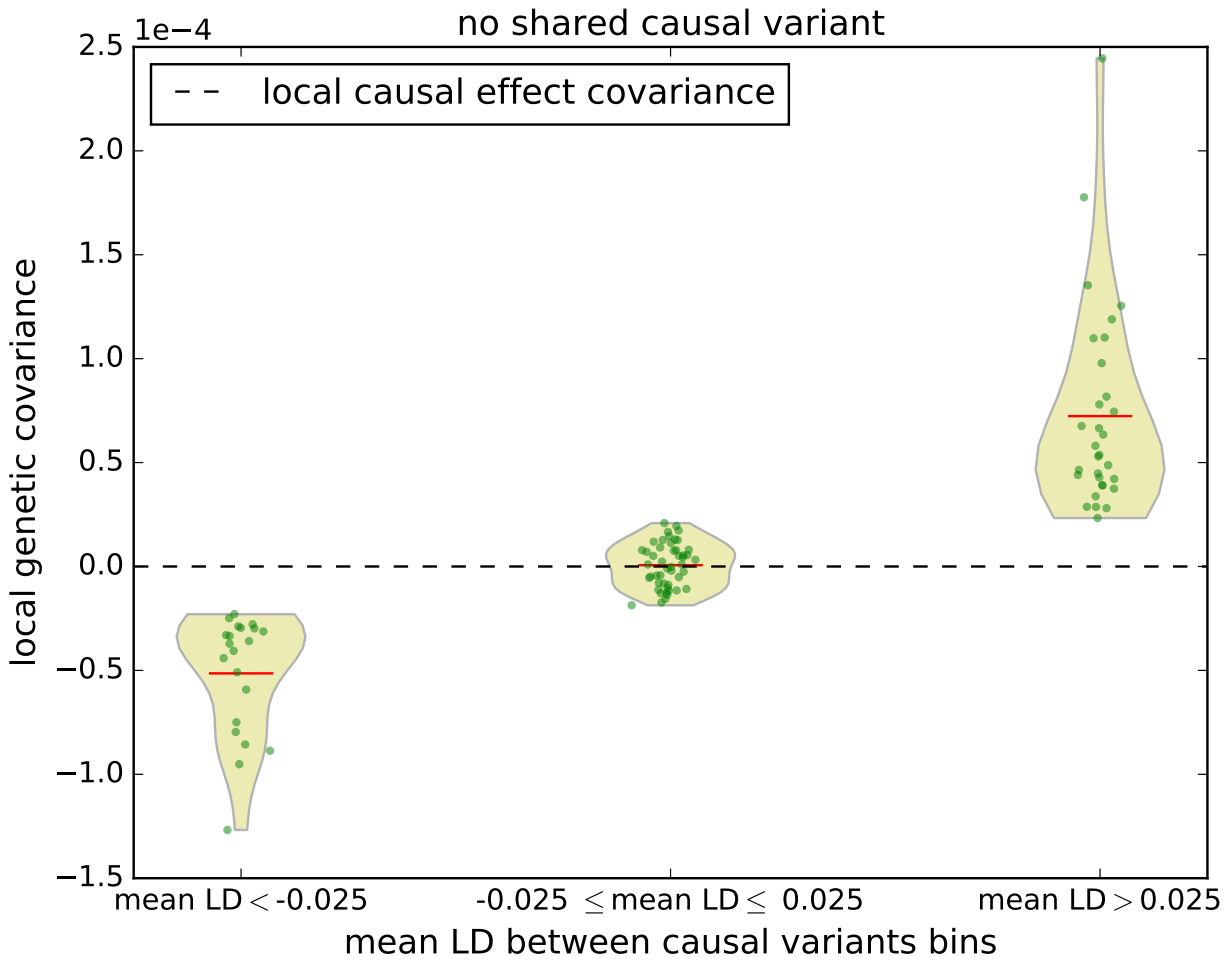


Figure 4.2: **Distribution of simulated genetic covariance and causal effect covariance across 100 LD-independent regions on chromosome 1 binned by average LD between causal variants.** The red lines represent the average local genetic covariance in each bin. For each region, we simulated 2 traits, each with 3 causal variants with effect sizes set to 0.01, and with no shared causal variants (see Figure 4.3 for the case where the two traits share causal variants). Genetic covariance varies with respect to LD whereas causal effect covariance is always 0 (horizontal dotted line). Since genetic covariance can be thought as an upper bound of prediction accuracy using causal effects from one trait to another, a positive genetic covariance indicates that non-zero prediction accuracy could be attained by virtue of LD tagging.

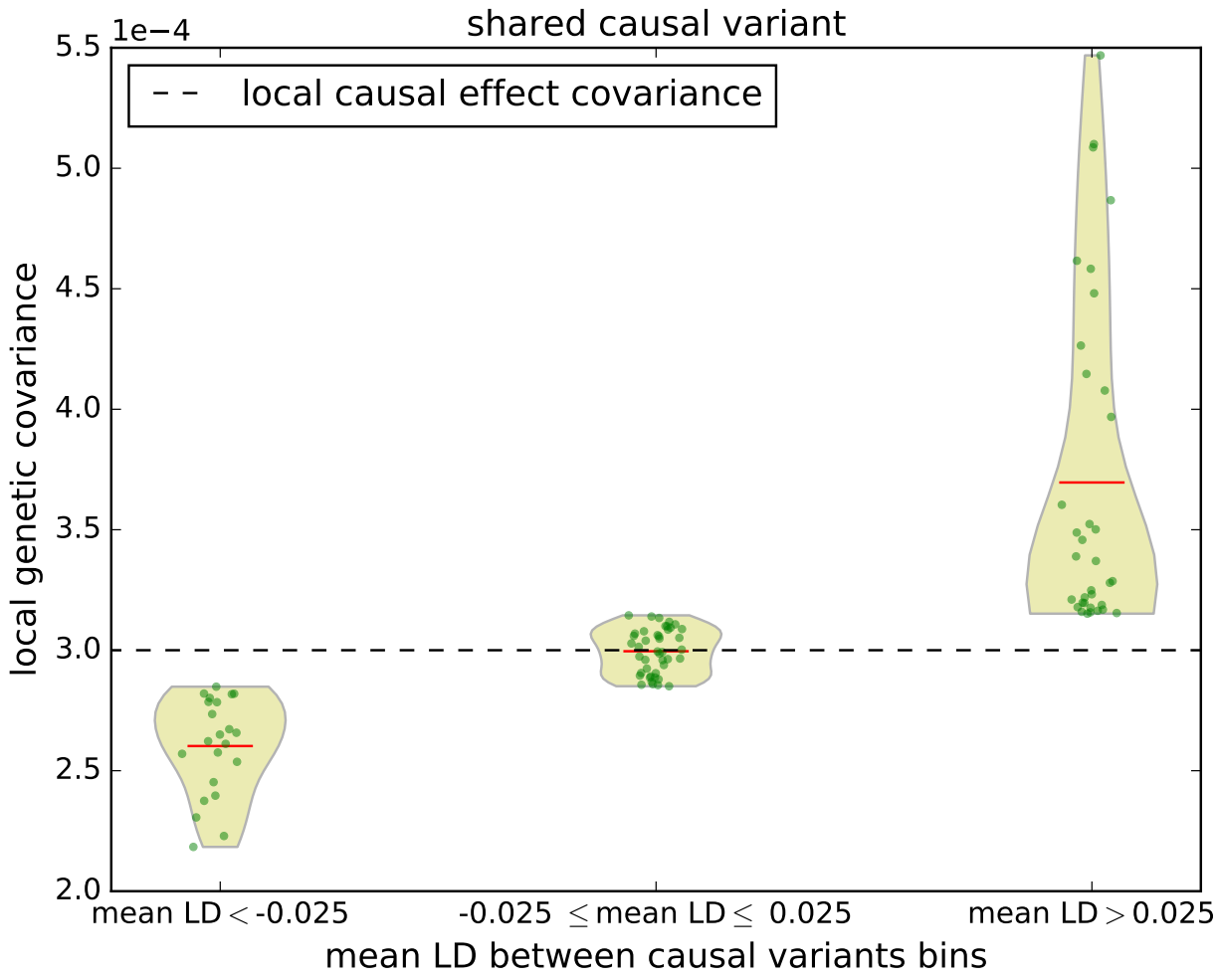


Figure 4.3: **Distribution of simulated local genetic covariance and causal effect covariance across 100 LD-independent regions on chromosome 1 binned by average LD between causal variants.** The red lines represent the average local genetic covariance in each bin. Both traits each have 3 causal variants with effect size set to 0.01, and share all the causal variants. Here, local genetic covariance varies with respect to LD whereas local causal effect covariance is fixed at 0.0003.

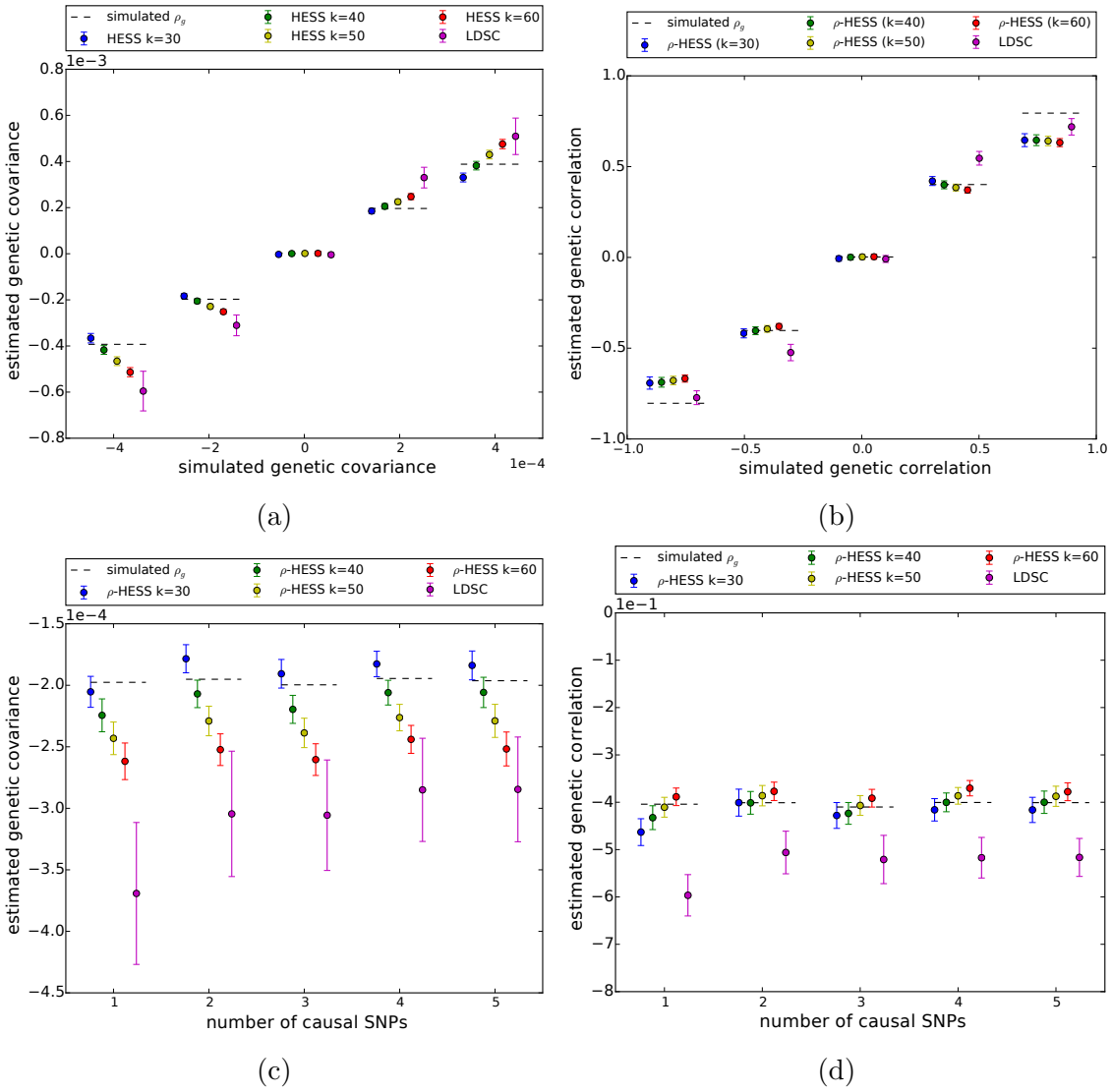


Figure 4.4: **Performance of ρ-HESS and cross-trait LDSC using external reference LD across 100 LD-independent regions, with each region having 1000 simulations.** Here, each dot represents the mean (over 100 regions) of the average performance (over 1000 simulations per region), with error bars representing 1.96 times the standard error on both sides. Overall, ρ-HESS provides approximately unbiased estimates of local genetic covariance (see Figure 4.4a) and correlation (see Figure 4.4b), and is not sensitive to the underlying genetic architectures (see Figure 4.4c for covariance and 4.4d for correlation). We also observe that ρ-HESS is less biased, more consistent, and has smaller standard error than cross-trait LDSC.

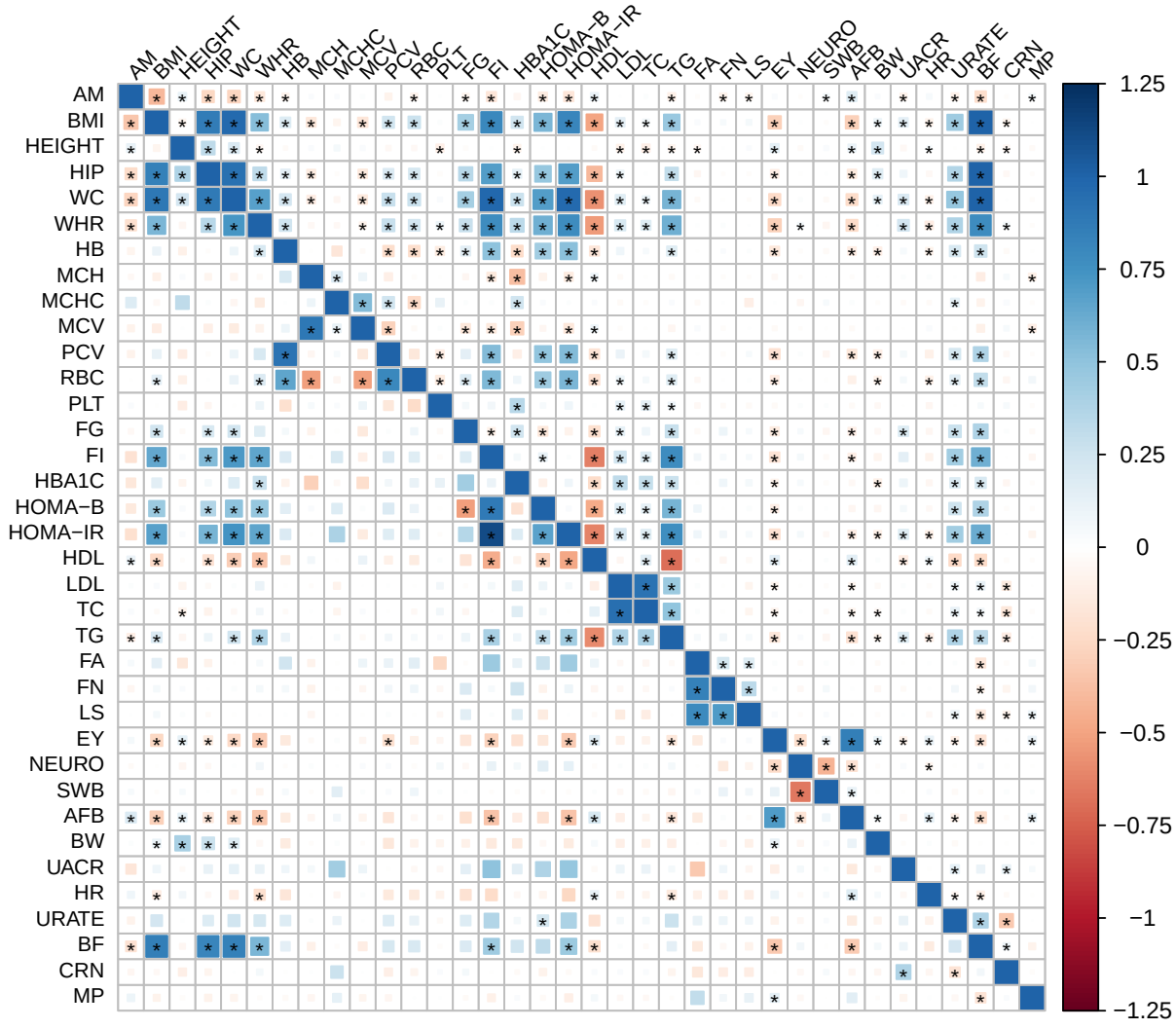


Figure 4.5: **Genetic correlation across the 36 complex traits obtained by ρ-HESS (top half) and cross-trait LDSC [18] (bottom half).** The magnitude of the correlation is represented by the color and the size of the square. Among the 630 pairs of traits, ρ-HESS (cross-trait LDSC) identified 298 (115) pairs showing significant genetic correlation (marked with dots)

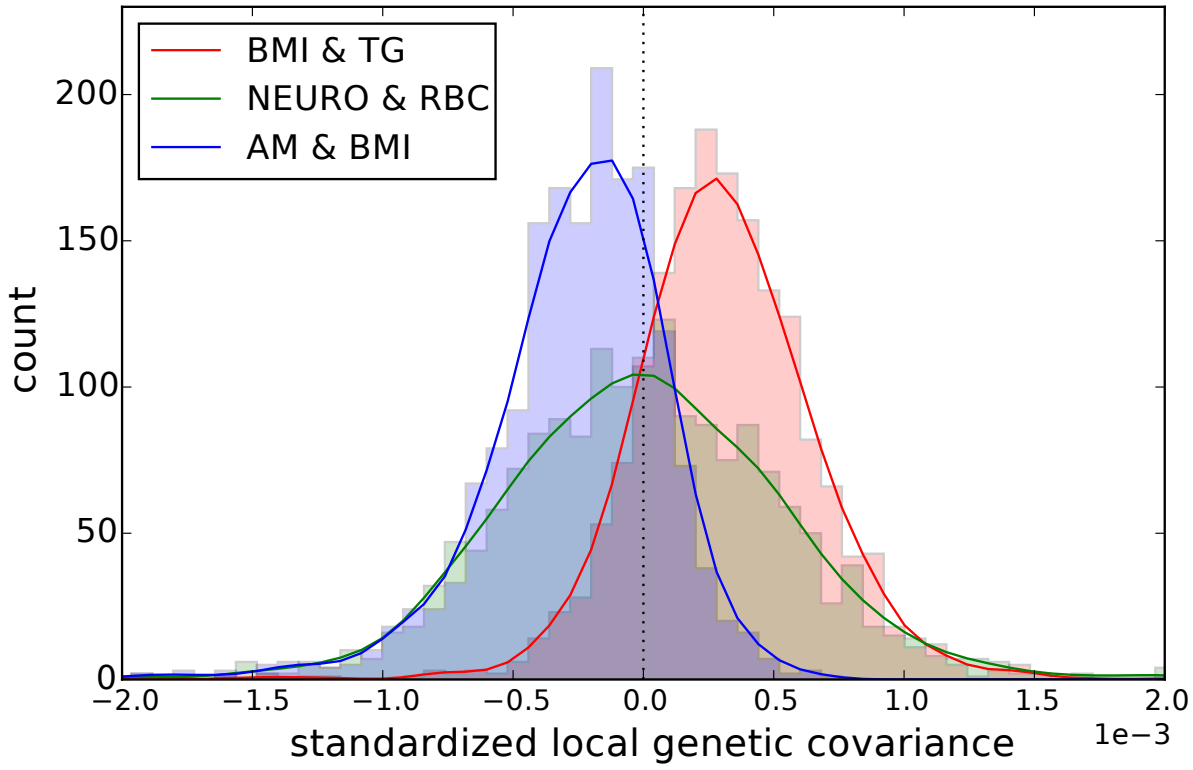


Figure 4.6: **Distribution of standardized local genetic covariance (local genetic covariance standardized by the square roots of total SNP-heritability of two traits) for the pairs of traits BMI and TG, NEURO and RBC, AM and BMI.** Pairs of traits with positive (negative) genome-wide genetic correlation show a shift in the distribution of standardized local genetic covariance away from 0.

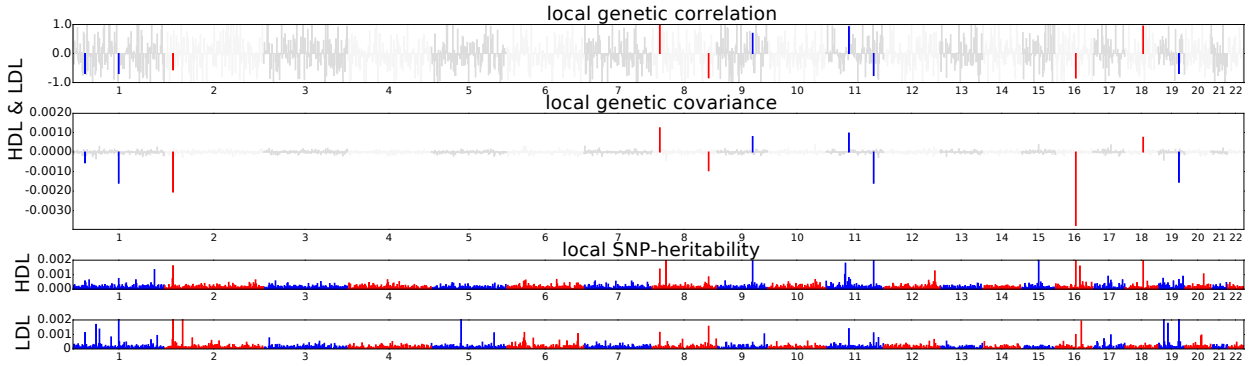


Figure 4.7: **Manhattan-style plots showing the estimates of local genetic covariance for the pairs of traits HDL and LDL.** Although the genome-wide genetic correlation between HDL and LDL does not reach the significance level ($p < 0.05/630$), 11 loci exhibit significant local genetic covariance.

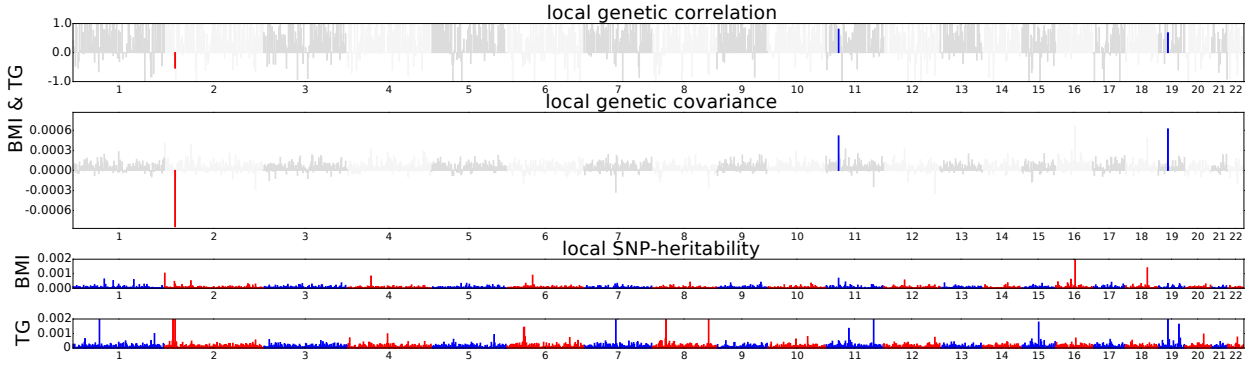


Figure 4.8: **Manhattan-style plots showing the estimates of local genetic covariance for the pairs of traits BMI and TG.** That the local genetic covariance between BMI and TG is mostly one-sided implies plausible causal relationship between the two traits

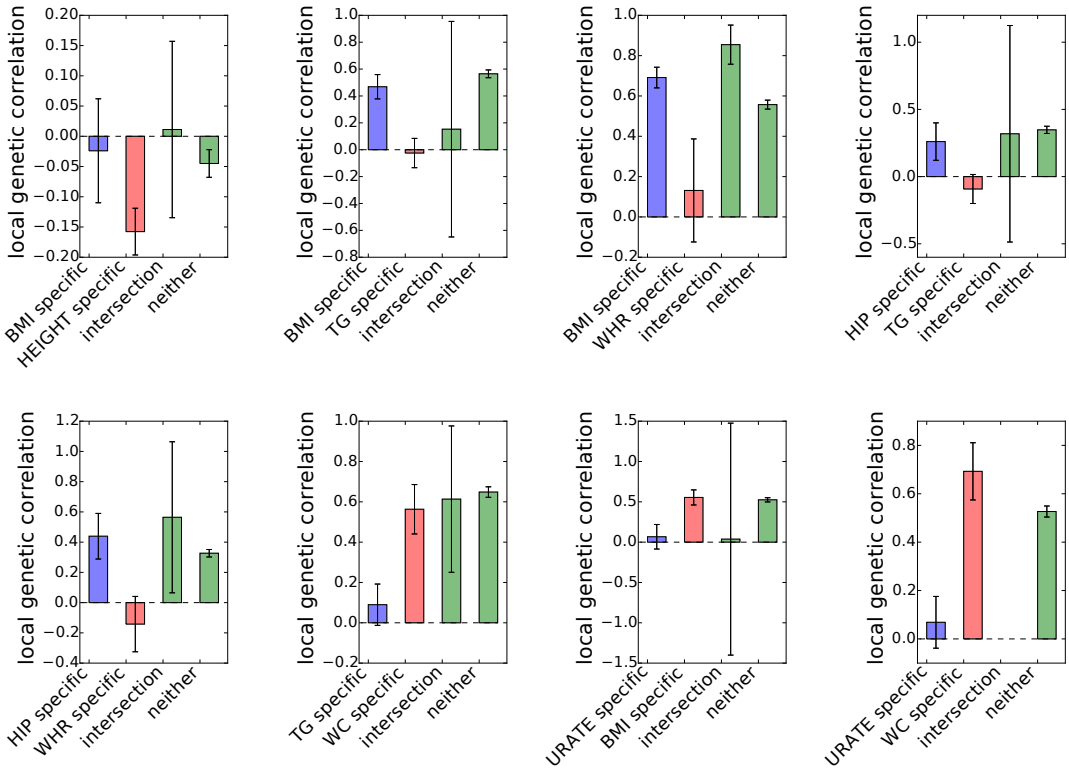


Figure 4.9: **Estimates of local genetic correlation at loci ascertained for GWAS risk variants for 8 examples pairs of traits that show plausible causal relationship.** We obtained standard error using a jackknife approach. Error bars represent 1.96 times the standard error on each side.

CHAPTER 5

Dissecting genetic architectures of complex traits specific to and shared by East Asian and European populations

5.1 Introduction

Most genome-wide association studies (GWASs) to date are based on samples of European descent [127, 135, 161, 120]. The lack of GWAS in other continental populations, such as African and Asians, limits the transferability of GWAS findings in Europeans into other populations due to factors including heterogeneity in genetic architectures, linkage disequilibrium (LD), minor allele frequencies, and environmental background [100, 102, 135, 144, 15]. The recent increase in the number of GWASs in non-European populations creates immense opportunities for trans-ethnic genetic studies [114, 21, 87, 71]. Indeed, analyzing GWAS results obtained from different continental populations has been shown to greatly improve power of disease mapping [108, 87, 114], resolution of fine-mapping [167, 5, 72, 173], and accuracy of risk prediction [101]. A fundamental quantity of interest in trans-ethnic genetic studies is the similarity of genetic architectures of a complex trait in two continental populations, and has been measured through trans-ethnic genetic correlation [170, 97, 15]. Methods for estimating trans-ethnic genetic correlation typically rely on the infinitesimal model, assuming every genetic variant contributes a small effect to the complex trait in both populations [170, 15], and thus do not explicitly model polygenicity of the complex trait. While previous study suggests that most common causal variants are shared across continental populations

[100], what the proportion of shared causal variants is and which genetic variants [44, 43, 56] are shared / population-specific remain unclear.

Here, we introduce POSC, a method to dissect genetic variants that are causal only in a single continental population (i.e. population-specific) and those that are causal in both continental populations (i.e. shared) from GWAS summary statistics data. POSC employs a Bayesian approach to explicitly model polygenicity of a complex trait and linkage disequilibrium in both continental populations. POSC yields as output estimates of genome-wide proportions of population-specific and shared causal variants. These estimates constitute prior probabilities in an empirical Bayes framework to quantify posterior probability of each SNP to be population-specific or shared. Further, we define enrichment of population-specific / shared causal variants in a functional category as the ratio between the posterior expectation and prior expectation of the number of population-specific / shared causal variants in the functional category.

Through extensive simulations, we show that POSC yields accurate estimates of proportion of population-specific / shared causal variants and well-calibrated statistics for testing enrichment using either in-sample or external reference LD matrices. We applied POSC on summary associations statistics of 18 large-scale GWASs of 9 complex traits and diseases in East Asian and European populations (average $N_{\mathrm{EAS}} = 94,621$ $N_{\mathrm{EUR}} = 103,507$) using 1000 Genomes Project [28] as the external reference panel. First, we estimated genome-wide proportion of population-specific and shared causal variants. Next, we quantified posterior probability for each SNP to be population-specific / shared, and estimated expected number of population-specific / shared causal variants at GWAS risk regions. Finally, we estimated enrichment of population-specific / shared causal variants in specifically expressed genes in 53 GTEx tissues [43]. We found that all the traits analyzed were highly polygenic in both populations, that while a large proportion of causal variants of these traits and diseases were shared by both populations, each population also possessed a substantial proportion of population-specific causal variants, and that regions of genes expressed in trait-relevant tissues harbor both population-specific and shared causal variants.

5.2 Methods

5.2.1 Overview of methods

We introduce POSC, a method for dissecting population-specific and shared causal variants from GWAS summary statistics data, while accounting for linkage disequilibrium in two continental populations. POSC explicitly models the SNP causal status vectors in two continental populations (see Figure 5.1), and imposes a mixture of zero and normal prior on SNP effect sizes [73, 10, 64]. POSC yields estimates of numbers of population-specific and shared causal variants using a expectation maximization (EM) algorithm [37] coupled with Markov Chain Monte Carlo (MCMC) sampling. These estimates are subsequently used to quantify posterior probability of each SNP to be population-specific or shared in an empirical Bayes framework. We also provides method to quantify enrichment of population-specific / shared causal variants in a functional category, analogous to but conceptually different from definitions of enrichment of SNP-heritability [44, 149].

5.2.2 The multivariate Bernoulli (MVB) distribution

The multivariate Bernoulli (MVB) is a generalization of the Bernoulli for modeling distribution of binary vectors of arbitrary size [30, 141]. Let $\boldsymbol{B} \in \{0,1\}^p$ represent a random binary vector of size p, then the distribution of $\boldsymbol{B}$ under MVB can be described by 2^p probabilities, namely $\Pr(\boldsymbol{B} = 0, \cdots, 0) \cdots \Pr(\boldsymbol{B} = 1, \cdots, 1)$, one for each of the 2^p possible realizations of $\boldsymbol{B}$ [30, 141]. Alternatively, one can adopt an index set representation of the binary vector $\boldsymbol{B}$, $\boldsymbol{A} = \{i : \boldsymbol{B}_i = 1\}$, a set of indices of 1's in $\boldsymbol{B}$, and represent the distribution of $\boldsymbol{B}$ as the ratio,

$$\Pr(\boldsymbol{B}) = \Pr(\boldsymbol{A}) = \frac{\exp\left(\sum_{\boldsymbol{C} \subseteq \boldsymbol{A}} f_{\boldsymbol{C}}\right)}{\sum_{\boldsymbol{D}} \exp\left(\sum_{\boldsymbol{C} \subseteq \boldsymbol{D}} f_{\boldsymbol{C}}\right)} = \frac{\exp\left(S_{\boldsymbol{A}}\right)}{\sum_{\boldsymbol{D}} \exp\left(S_{\boldsymbol{D}}\right)}, \tag{5.1}$$

where $f_{\boldsymbol{C}}$'s are the natural parameters of the MVB [30, 141], and $S_{\boldsymbol{A}} = \sum_{\boldsymbol{C} \subseteq \boldsymbol{A}} f_{\boldsymbol{C}}$.

We use the convention that the right-most bit in the binary vector is the first bit, and the

left-most bit is the last bit. For the sake of convenience, we use binary string and index set representation of binary vectors interchangeably (e.g. both the binary string 011 and the index set $\{1, 2\}$ represent the binary vector $(0, 1, 1)$).

As a concrete example, consider binary vectors of size 2. The probabilities of each possible realization of binary vectors of size 2 under the MVB are

$$\begin{aligned}
\Pr(00) = \Pr(\phi) &= \frac{\exp(f_{00})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(01) = \Pr(\{1\}) &= \frac{\exp(f_{00}+f_{01})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(10) = \Pr(\{2\}) &= \frac{\exp(f_{00}+f_{10})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(11) = \Pr(\{1,2\}) &= \frac{\exp(f_{00}+f_{01}+f_{10}+f_{11})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})}
\end{aligned} . \tag{5.2}$$

5.2.3 Modeling GWAS summary statistics in two ancestral populations

5.2.3.1 MVB prior for SNP causal status in two ancestral populations

We use binary vector of size 2, $\boldsymbol{C}_i = (\boldsymbol{c}_{i1}, \boldsymbol{c}_{i2})$, to model the causal statuses of SNP i in two ancestral populations. In total, there are 4 possible binary vectors of size 2: (1) $\boldsymbol{C}_i = 00$, the SNP is causal for none of the population; (2) $\boldsymbol{C}_i = 01$, the SNP is causal in population 1; (3) $\boldsymbol{C}_i = 10$, the SNP is causal in population 2; (4) $\boldsymbol{C}_i = 11$, the SNP is causal in both populations.

$\boldsymbol{C}_i$ can be modeled using a multinomial distribution, $\text{Mult}(p_{00}, p_{01}, p_{10}, p_{11})$, where p_{00}, p_{01}, p_{10}, and p_{11} represent the probability of each possible binary vector of size 2. Equivalently, one can model $\boldsymbol{C}_i$ through the MVB as,

$$\begin{aligned}
\Pr(\boldsymbol{C}_i = 00) = \frac{\exp(f_{00})}{\eta},\ \Pr(\boldsymbol{C}_i = 01) &= \frac{\exp(f_{01}+f_{00})}{\eta} \\
\Pr(\boldsymbol{C}_i = 10) = \frac{\exp(f_{10}+f_{00})}{\eta},\ \Pr(\boldsymbol{C}_i = 11) &= \frac{\exp(f_{11}+f_{10}+f_{01}+f_{00})}{\eta},
\end{aligned} \tag{5.3}$$

where $\eta = \exp(f_{00}) + \exp(f_{01} + f_{00}) + \exp(f_{10} + f_{00}) + \exp(f_{11} + f_{10} + f_{01} + f_{00})$ is the normalization constant, and $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$ parameters of the MVB (see Equation (5.2)).

The MVB distribution is invariant with respect to the parameter f_{00}, and we enforce f_{00} to be 0 as a convention [30]. The parameters, f_{01} and f_{10}, govern the probability of a SNP being specific to a population, and f_{11}, the dependence of causal status between two populations – a zero f_{11} indicates independence, and a non-zero f_{11}, dependence [30, 141]. Each MVB parameter is a real number (i.e. $\boldsymbol{f} \in \mathbb{R}^4$), one can apply unconstrained optimization to estimate the MVB parameters.

5.2.3.2 Joint distribution of GWAS summary statistics in two ancestral populations

We model phenotypes in two ancestral populations, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, using the linear model, $\boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$ and $\boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, where $\boldsymbol{Y}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{n_2}$ are the phenotype measurements of the phenotype in the two populations, with sample size n_1 and n_2, $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times p}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times p}$ column-standardized genotype matrix at p SNPs, $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ standardized effect size of SNPs on the phenotype in two populations, and $\boldsymbol{\epsilon}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{n_2}$ environmental effect. We further assume that each row of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is drawn from a distribution where the covariance structure is $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, the LD matrix of each population, respectively, and that $\boldsymbol{\epsilon}_{1i} \sim N(0, \sigma_{e1}^2)$, $\boldsymbol{\beta}_{2i} \sim N(0, \sigma_{e2}^2)$, where σ_{e1}^2 and σ_{e2}^2 represent variance of the environmental effects.

In typical GWASs, one obtains association statistics (Z-scores) of every SNP as

$$\begin{aligned} \boldsymbol{Z}_1 &= \frac{1}{\sqrt{n_1}} \boldsymbol{X}_1^\intercal \boldsymbol{Y}_1, \\ \boldsymbol{Z}_2 &= \frac{1}{\sqrt{n_2}} \boldsymbol{X}_2^\intercal \boldsymbol{Y}_2, \end{aligned} \tag{5.4}$$

and have been shown to follow a multivariate normal distribution [140],

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{\beta}_1 &\sim N\left(\sqrt{n_1}\boldsymbol{V}_1\boldsymbol{\beta}_1, \sigma_{e1}^2\boldsymbol{V}_1\right), \\ \boldsymbol{Z}_2|\boldsymbol{\beta}_2 &\sim N\left(\sqrt{n_2}\boldsymbol{V}_1\boldsymbol{\beta}_2, \sigma_{e2}^2\boldsymbol{V}\right). \end{aligned} \tag{5.5}$$

Further, given causal status vectors, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, of every SNP in each population, one obtains the conditional distribution $\boldsymbol{Z}_1|\boldsymbol{\beta}_1, \boldsymbol{c}_1$ and $\boldsymbol{Z}_2|\boldsymbol{\beta}_2, \boldsymbol{c}_2$ as

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{\beta}_1, \boldsymbol{c}_1 &\sim N\left(\sqrt{n_1}\boldsymbol{V}_1(\boldsymbol{\beta}_1 \circ \boldsymbol{c}_1), \sigma_{e1}^2\boldsymbol{V}_1\right), \\ \boldsymbol{Z}_2|\boldsymbol{\beta}_2, \boldsymbol{c}_2 &\sim N\left(\sqrt{n_2}\boldsymbol{V}_2(\boldsymbol{\beta}_2 \circ \boldsymbol{c}_2), \sigma_{e2}^2\boldsymbol{V}_2\right), \end{aligned} \tag{5.6}$$

where $\circ$ denotes the Hadamard product [74].

Following Equation (5.6), one can evaluate the likelihood of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ given the true causal effect size vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. However, in reality the true causal effect size vectors are not given, and estimating these parameters from data will likely lead to over-fit. Instead, we impose a normal prior on the causal SNPs in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$,

$$\begin{aligned} \boldsymbol{\beta}_1|\boldsymbol{c}_1 &\sim N\left(\boldsymbol{0}, \frac{h_{g1}^2}{|\boldsymbol{c}_1|}\,\mathrm{d}iag(\boldsymbol{c}_1)\right), \\ \boldsymbol{\beta}_2|\boldsymbol{c}_2 &\sim N\left(\boldsymbol{0}, \frac{h_{g2}^2}{|\boldsymbol{c}_2|}\,\mathrm{d}iag(\boldsymbol{c}_2)\right), \end{aligned} \tag{5.7}$$

where h_{g1}^2, h_{g2}^2 are the SNP-heritability of the phenotype in the two populations, and $|\boldsymbol{c}_1|$, $|\boldsymbol{c}_2|$ denote the number of 1's (i.e. number of causal SNPs) in the binary vectors [73, 10, 64]. With the normal prior on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the conditional distribution, $\boldsymbol{Z}_1|\boldsymbol{c}_1$ and $\boldsymbol{Z}_2|\boldsymbol{c}_2$, is then

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{c}_1 &\sim N\left(\boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2\boldsymbol{V}_1\,\mathrm{d}iag(\boldsymbol{c}_1)\boldsymbol{V}_1\right), \\ \boldsymbol{Z}_2|\boldsymbol{c}_2 &\sim N\left(\boldsymbol{0}, \boldsymbol{V}_2 + \sigma_2^2\boldsymbol{V}_2\,\mathrm{d}iag(\boldsymbol{c}_2)\boldsymbol{V}_2\right), \end{aligned} \tag{5.8}$$

where $\sigma_1^2 = \frac{n_1 h_{g1}^2}{|\boldsymbol{c}_1|}$ and $\sigma_2^2 = \frac{n_2 h_{g2}^2}{|\boldsymbol{c}_2|}$.

Incorporating the MVB prior on the causal status vectors, one obtains the distribution of

$\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, parameterized by the MVB parameters, $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$,

$$\begin{aligned}\Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}) &= \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{c}_1, \boldsymbol{c}_2; \boldsymbol{f}) = \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \Pr(\boldsymbol{Z}_1|\boldsymbol{c}_1)\Pr(\boldsymbol{Z}_2|\boldsymbol{c}_2)\Pr(\boldsymbol{c}_1, \boldsymbol{c}_2; \boldsymbol{f}) \\ &= \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \left[\begin{array}{c} N(\boldsymbol{Z}_1; \boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2 \boldsymbol{V}_1 \, \mathrm{d}iag(\boldsymbol{c}_1)\boldsymbol{V}_1)\times \\ N(\boldsymbol{Z}_2; \boldsymbol{0}, \boldsymbol{V}_2 + \sigma_2^2 \boldsymbol{V}_2 \, \mathrm{d}iag(\boldsymbol{c}_2)\boldsymbol{V}_2) \times \prod_{i=1}^{p} \frac{\exp(S_{\boldsymbol{C}_i})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right] \end{aligned} \tag{5.9}$$

To model joint distribution of GWAS summary statistics across L LD-independent loci, we take the product of the probability of Z-scores at each loci,

$$\begin{aligned}&\Pr(\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}}; \boldsymbol{f}) = \prod_{l=1}^{L} \Pr(\boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}; \boldsymbol{f}) \\ &= \prod_{l=1}^{L} \left\{ \sum_{\boldsymbol{c}_{1l}}\sum_{\boldsymbol{c}_{2l}} \left[\begin{array}{c} N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \, \mathrm{d}iag(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l})\times \\ N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \, \mathrm{d}iag(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \times \prod_{i=1}^{p_l} \frac{\exp(S_{\boldsymbol{C}_{li}})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right] \right\}. \end{aligned} \tag{5.10}$$

5.2.4 Model fitting using Expectation Maximization

5.2.4.1 Expectation step

We use expectation-maximization (EM) to estimate the model parameters $\boldsymbol{f}$. First, we derive the complete log-likelihood of the data

$$\begin{aligned}&\ell\left(\boldsymbol{f}|\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}}, \boldsymbol{c}_{1\{1,\cdots,L\}}, \boldsymbol{c}_{2\{1,\cdots,L\}}\right) \\ &= \log\left\{ \prod_{l=1}^{L} \left[\begin{array}{c} N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \, \mathrm{d}iag(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l})\times \\ N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \, \mathrm{d}iag(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \times \prod_{i=1}^{p_l} \frac{\exp(S_{\boldsymbol{C}_{li}})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right] \right\} \\ &= \sum_{l=1}^{L} \left[\log N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \, \mathrm{d}iag(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l}) + \log N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \, \mathrm{d}iag(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l})\right] \\ &\quad + \sum_{l=1}^{L}\sum_{i=1}^{p_l} S_{C_{li}} - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l. \end{aligned} \tag{5.11}$$

In the expectation step of the EM algorithm, one finds the expectation of the log likelihood with respect to the causal status vectors $\boldsymbol{c}_{1\{1,\cdots,L\}}$, $\boldsymbol{c}_{2\{1,\cdots,L\}}$, conditioned on the current estimate of the model parameter $\boldsymbol{f}^{(t)}$,

$$\begin{aligned} Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right) &= \mathrm{E}\left[\ell\left(\boldsymbol{f}|\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}}, \boldsymbol{c}_{1\{1,\cdots,L\}}, \boldsymbol{c}_{2\{1,\cdots,L\}}\right)\right] \\ &= \sum_{l=1}^{L}\sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left[\begin{array}{c} \log N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l}\, \mathrm{d}iag(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l}) \\ + \log N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l}\, \mathrm{d}iag(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \end{array}\right] \\ &\quad + \sum_{l=1}^{L}\sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right) - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l, \end{aligned} \tag{5.12}$$

where $\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right)$ can be found as,

$$\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) = \frac{\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}|\boldsymbol{f}^{(t)}\right)}{\sum_{\boldsymbol{b}_{1l},\boldsymbol{b}_{2l}} \Pr\left(\boldsymbol{b}_{1l}, \boldsymbol{b}_{2l}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}|\boldsymbol{f}^{(t)}\right)}. \tag{5.13}$$

5.2.4.2 Maximization step

In the maximization step, one finds

$$\boldsymbol{f}^{(t+1)} = \mathrm{a}rgmax_{\boldsymbol{f}}\, Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right) = \mathrm{a}rgmax_{\boldsymbol{f}}\, g(\boldsymbol{f}), \tag{5.14}$$

where

$$g(\boldsymbol{f}) = \sum_{l=1}^{L}\sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right) - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l, \tag{5.15}$$

removing the irrelevant constant in $Q(\boldsymbol{f}|\boldsymbol{f}^{(t)})$.

Evaluating $g(\boldsymbol{f})$ involves a summation over all possible causal status vectors, which has time

complexity on the order of $O(2^{2p_l})$ and is intractable. Instead we recognize that

$$\begin{aligned} g(\boldsymbol{f}) &= \sum_{l=1}^{L} \sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \mathrm{E}\left[\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right] - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l \\ &\approx h(\boldsymbol{f}) = \sum_{l=1}^{L}\left[\frac{1}{J}\sum_{j=1}^{J}\left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}^{(j)}}\right)\right] - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l, \end{aligned} \tag{5.16}$$

where $\boldsymbol{C}_{li}^{(j)} = \left(\boldsymbol{c}_{1i}^{(j)}, \boldsymbol{c}_{2i}^{(j)}\right)$ represents the causal status of the i-th SNP at locus l in the two populations, from the causal status vectors, $\boldsymbol{c}_1^{(j)}$, $\boldsymbol{c}_2^{(j)}$, sampled from the posterior distribution $\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l} | \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}, \boldsymbol{f}^*\right)$. We use Gibbs sampling to efficiently sample causal status vectors from the posterior (see Section 5.2.5).

It can be shown that the following parameter updates maximizes $h(\boldsymbol{f})$,

$$\begin{aligned} \boldsymbol{f}_{00}^{(t+1)} &= 0, \\ \boldsymbol{f}_{01}^{(t+1)} &= \log \bar{q}_{01} - \log \bar{q}_{00}, \\ \boldsymbol{f}_{10}^{(t+1)} &= \log \bar{q}_{10} - \log \bar{q}_{00}, \\ \boldsymbol{f}_{11}^{(t+1)} &= \log \bar{q}_{11} - \log \bar{q}_{01} - \log \bar{q}_{10} + \log \bar{q}_{00}, \end{aligned} \tag{5.17}$$

where $\bar{q}_{00}$, $\bar{q}_{01}$, $\bar{q}_{10}$, and $\bar{q}_{11}$ represent the average count of 01, 10, and 11 causal status at a single SNP in two ancestral populations across MCMC samples from the Gibbs sampler (see Section 5.2.5).

5.2.5 Sampling causal status vectors from posterior distribution

We use Gibbs sampling to sample $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$ from the posterior distribution,

$$\boldsymbol{C} \sim \Pr\left(\boldsymbol{C} | \boldsymbol{f}, \boldsymbol{Z}_1, \boldsymbol{Z}_2\right) \propto \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C} | \boldsymbol{f}\right). \tag{5.18}$$

For notational simplicity, here, we drop the index l, representing different loci. To advance the Markov chain from step j to step $j+1$ in Gibbs sampling, at step j we select SNP k and

evaluate the probability of the four possible cross-population causal configurations at that SNP,

$$\begin{aligned}&\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 00, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \ \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 01, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \\ &\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 10, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \ \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 11, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right),\end{aligned} \tag{5.19}$$

where $\boldsymbol{C}_{\neg j}^{(j)}$ denotes the rest of the causal configurations excluding that of SNP k in the j-th step. Then we sample $\boldsymbol{C}^{(j+1)}$ based on the following probability

$$\Pr\left(\boldsymbol{C}^{(t+1)} = \left(\boldsymbol{C}_k = \boldsymbol{b}', \boldsymbol{C}_{\neg j}^{(j)}\right)\right) = \frac{\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = \boldsymbol{b}', \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right)}{\sum_{\boldsymbol{b}} \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = \boldsymbol{b}, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right)}. \tag{5.20}$$

To evaluate $\Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{c}_1, \boldsymbol{c}_2 | \boldsymbol{f}) = \Pr(\boldsymbol{Z}_1 | \boldsymbol{c}_1) \Pr(\boldsymbol{Z}_2 | \boldsymbol{c}_2) \Pr(\boldsymbol{c}_1, \boldsymbol{c}_2 | \boldsymbol{f})$, we note that previous work has shown that

$$\Pr(\boldsymbol{Z}_1 | \boldsymbol{c}_1) = N\left(\boldsymbol{Z}_1 | \boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2 \boldsymbol{V}_1^2\right) \propto \frac{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2\right)}{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1}\right)}, \tag{5.21}$$

where $BF_1 = \frac{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2\right)}{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1}\right)}$ is the Bayes factor at only the causal SNPs, reducing the time complexity of evaluating the probability from p^3 to p_{causal}^3. Let $\boldsymbol{V}_{1\boldsymbol{c}_1} = \sum_{i=1}^{p_{\text{causal}}} w_i \boldsymbol{u}_i \boldsymbol{u}_i^\intercal$ be the eigenvalue decomposition of $\boldsymbol{V}_{1\boldsymbol{c}_1}$, where w_i and $\boldsymbol{u}_i$ are the eigenvalues and eigenvectors of $\boldsymbol{V}_{1\boldsymbol{c}_1}$, we further note that BF_1 can be expressed as

$$\begin{aligned}BF_1 &= \frac{\det(\boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal (\boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2)^{-1} \boldsymbol{Z}_{1\boldsymbol{c}_1}\right]}{\det(\boldsymbol{V}_{1\boldsymbol{c}_1})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal \boldsymbol{V}_{1\boldsymbol{c}_1}^{-1} \boldsymbol{Z}_{1\boldsymbol{c}_1}\right)} \\ &\propto \left(\prod_{i=1}^{p_{\text{causal}}} \frac{1}{1 + \sigma_1^2 w_i}\right)^{\frac{1}{2}} \exp\left[\frac{1}{2} \sum_{i=1}^{p_{\text{causal}}} \frac{\sigma_1^2}{1 + \sigma_1^2 w_i} \left(\boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal \boldsymbol{u}_i\right)^2\right],\end{aligned} \tag{5.22}$$

avoiding numerical instability introduced by small eigenvalues. The Bayes factor for $\boldsymbol{Z}_{2\boldsymbol{c}_2}$ can be obtained using the same approach.

5.2.6 Posterior probability of each SNP to be ancestry-specific or shared

We use posterior probability of each SNP to be ancestry-specific or shared to assess evidence of specific or shared genetic architecture at single-SNP resolution. Specifically, we evaluate

$$\Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) \tag{5.23}$$

for $\boldsymbol{b} \in \{01, 10, 11\}$ at each SNP i, where $\boldsymbol{f}^*$ denotes the estimated MVB parameter. We show below that the per SNP posterior probability in Equation (5.23) can be evaluated using the Gibbs sampling procedure outlined in Section 5.2.5. First, we note that

$$\begin{aligned}
&\Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) = \sum_{\boldsymbol{C}_{\neg i}} \Pr(\boldsymbol{C}_i = \boldsymbol{b}, \boldsymbol{C}_{\neg i}|\boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*) \\
&= \sum_{\boldsymbol{C}_{\neg i}} \Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*) \Pr(\boldsymbol{C}_{\neg i}|\boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*) \\
&= \mathrm{E}\left[\Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*)\right] = \mathrm{E}\left[\mathrm{E}[\mathbb{1}_{\{\boldsymbol{C}_i=\boldsymbol{b}\}}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*]\right] \\
&= \mathrm{E}[\mathbb{1}_{\{\boldsymbol{C}_i=\boldsymbol{b}\}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2\boldsymbol{f}^*] \approx \frac{\sum_{j=1}^{J} \mathbb{1}_{\{\boldsymbol{C}_i^{(j)}=\boldsymbol{b}\}}}{J},
\end{aligned} \tag{5.24}$$

where $\boldsymbol{C}^{(j)}$ is the j-th causal status vectors (out of a total of J samples) sampled from the posterior distribution $\Pr(\boldsymbol{C}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)$ (see Section 5.2.5). To ensure stable estimates of the posterior probability, we run the Gibbs sampling procedure 20 times and report the average posterior probability.

5.2.7 Defining approximately independent LD blocks in both ancestral populations

We adapted LDetect [12] to define blocks of SNPs that are approximately independent in both East Asian and European populations. Briefly, LDetect is a method to define approximately independent blocks of SNPs in a single population [12]. It involves estimating a regularized LD matrix of a single population and identifying block structures in the LD

matrix, which constitute the approximately independent LD blocks[12].

To define approximately independent blocks of SNPs for two populations, we first compute regularized LD matrices of both populations ($\boldsymbol{V}_{\text{EAS}}$ and $\boldsymbol{V}_{\text{EUR}}$), following the LDetect procedure[12]. Then we construct a new matrix ($\boldsymbol{V}_{\text{trans}}$) by taking the maximum LD in East Asian and European LD matrices for each pair of SNPs,

$$\boldsymbol{V}_{\text{trasn},ij} = \begin{cases} \boldsymbol{V}_{\text{EAS},ij} & \text{if} |\boldsymbol{V}_{\text{EAS},ij}| > |\boldsymbol{V}_{\text{EUR},ij}| \\ \boldsymbol{V}_{\text{EUR},ij} & \text{if} |\boldsymbol{V}_{\text{EUR},ij}| > |\boldsymbol{V}_{\text{EAS},ij}| \end{cases}. \tag{5.25}$$

The matrix $\boldsymbol{V}_{\text{trans}}$ is block diagonal due to shared recombination hot spots in both populations. We then applied LDetect procedure [12] to identify block structure in $\boldsymbol{V}_{\text{trans}}$.

Using the above procedure, we identified 1,368 approximately independent LD blocks (2Mb wide on average) in both East Asian and European populations.

5.2.8 Simulation framework

We used genotype data of chromosome 22 from CONVERGE [21] and UK Biobank [151] to simulate GWAS summary statistics for East Asian and European populations. We used genotype data from 1000 Genomes Project [2] as the reference panel. Since SNPs in perfect LD have identical Z-scores, we performed minimal LD pruning (at R^2 threshold of 0.95) on reference LD matrix using PLINK 1.9 [132], to remove perfectly correlated SNPs. We also removed strand-ambiguous SNPs and SNPs with minor allele frequency (MAF) less than 1% in either population, resulting in a total of 8,599 SNPs on chromosome 22.

We simulated phenotypes based on the linear model $\boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, where effects of causal SNPs, $\boldsymbol{\beta}_{1\boldsymbol{c}_1}$ and $\boldsymbol{\beta}_{2\boldsymbol{c}_1}$, in each population, were drawn from

$$\boldsymbol{\beta}_{1\boldsymbol{c}_1} \sim N\left(\boldsymbol{0}, \frac{h^2_{g1}}{|\boldsymbol{c}_1|}\boldsymbol{I}\right), \boldsymbol{\beta}_{2\boldsymbol{c}_2} \sim N\left(\boldsymbol{0}, \frac{h^2_{g2}}{|\boldsymbol{c}_2|}\boldsymbol{I}\right), \tag{5.26}$$

and effects of non-causal SNPs were set to 0. Here, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are the index set of causal SNPs in each population. We simulated environmental effect of each individual i, $\boldsymbol{\epsilon}_{1i}$ and $\boldsymbol{\epsilon}_{2i}$, from $\boldsymbol{\epsilon}_{1i} \sim N(0, 1 - h^2_{g1})$ and $\boldsymbol{\epsilon}_{2i} \sim N(0, 1 - h^2_{g2})$. We then simulated Z-scores for the entire chromosome 22.

5.2.8.1 Specifically expressed genes annotation

We obtained SNP annotations for genes specifically expressed in a tissue across 53 GTEx tissues from Finucane et al. [43]

5.3 Results

5.3.1 Performance of POSC in simulations

We assessed the performance of POSC through extensive simulations. First, we evaluated the computation efficiency of POSC. The EM-MCMC algorithm for estimating the number of population-specific / shared causal variants typically converged in 200 iterations (see Figure 5.2), with run time increasing with total number of causal SNPs (see Figure 5.2). For example, in simulations where we randomly drew 20 causal variants for each population, POSC terminated in 90 minutes on average, increasing to 360 minutes in simulations involving 100 causal variants. We note, however, that the EM-MCMC procedure can be parallelized to decrease run time. Evaluating per-SNP posterior probability to be population-specific or shared using the estimated prior took on average 5 minutes in simulations where 20 causal variants were drawn for each population, and 28 minutes in simulations involving 100 causal variants (see Figure 5.3).

Next, we evaluated the accuracy of POSC in estimating the number of population-specific and shared causal variants. When in-sample LD was used, POSC yielded approximately unbiased estimates of the number of population-specific and shared causal variants (see

Figure 5.4). And when external reference LD obtained from 1000 Genomes Project [28] was used, POSC yielded slightly upwardly biased estimates (see Figure 5.4 bottom panel). For example, in simulations where we randomly chose (out of a total of 8,599 SNPs) 50 East Asian-specific, 50 European-specific, and 50 shared causal variants, POSC yielded an estimates of 37.8 (S.E. 4.5), 40.3 (S.E. 4.9), and 64.9 (S.E. 6.3), respectively, when in-sample LD was used, and an estimates of 48.0 (S.E. 5.9), 53.7 (S.E. 7.44), and 78.8 (S.E. 7.6), respectively, when external reference LD was used.

We also assessed the effect of SNP-heritability of the trait and sample size of the GWAS on the estimates of POSC. We saw a slight decrease in accuracy as the product between SNP-heritability and sample size decreases (see Figure 5.5). This is not unexpected since the likelihood of GWAS summary statistics is a function of the product between SNP-heritability of the trait and sample size of the GWAS – as the product decreases to zero, Z-scores give little information regarding the causal status, leading to inaccurate estimates.

Finally, we obtained statistics testing enrichment of population-specific and shared causal variants in annotations defined by specifically expressed genes in 53 GTEx tissues [43] in simulations. Since we drew causal variants at random, the simulations constituted null simulations with no enrichment in any functional annotation. Overall, the statistics were conservative with either in-sample LD or external reference LD (see Figure 5.10), at different levels of polygenicity, or with different power (product between SNP-heritability and sample size) of the GWAS.

5.3.2 Number of population-specific and shared causal variants in complex traits

We analyzed 18 publicly available summary association statistics of GWAS of 9 complex traits in East Asian (EAS) and European (EUR) populations (see Table 5.1). For computational efficiency, we first estimated number of population-specific and shared causal variants on each chromosome separately in parallel, with 500 EM iterations for each chromosome

to ensure convergence. The EM algorithm converged after 200 iterates for all traits except BMI, which didn't converge until after 300 iteration, likely due to its high polygenicity.

Next, we aggregated the chromosomal estimates to obtain genome-wide number of population-specific and shared causal variants. The complex traits we analyzed displayed a wide range of degrees of polygenicity, with number of causal variants ranging from 877 (S.E. 8) and 1,228 (S.E. 11) in EAS and EUR for rheumatoid arthritis (RA), to 25,296 (S.E. 85) and 26,206 (S.E. 66) in EAS and EUR for BMI (see Table 5.2). Notably, we highlight that our estimated proportion of causal variants for BMI in EUR, 10.1%, was consistent with an estimate obtained by a recent method using individual-level data [175].

As expected, in all analyzed traits, a large fraction of causal variants in each population were shared with the other population (see Table 5.2), consistent with similar conclusions reached a by previous study [100]. For example, we estimated that among the 25,296 (S.E. 85) and 26,206 (S.E. 66) (EUR) causal variants for BMI in EAS and EUR, 22,664 (S.E. 141) were shared by both populations, comprising 90% and 86% of the total causal variants of BMI in each population, respectively. However, for some complex traits, each population also possessed a substantial proportion of population-specific causal variants. For example, out of the estimated total of 6,356 (S.E. 19) and 5,892 (S.E. 27) causal variants for total cholesterol (TC) in East Asians and Europeans, 2,467 (S.E. 22) and 2,003 (S.E. 16) were specific to each population, comprising 39% and 34% of the estimated total number of causal variants in each population, respectively (see Table 5.2).

5.3.3 Causal variants of complex traits are spread across the entire genome

We divided the estimated number of population-specific / shared causal variants by the total number of SNPs in each GWAS study to obtain proportion estimates, and used them as prior probabilities in an empirical Bayes framework to evaluate posterior probability of each SNP to be population-specific / shared (see Figure 5.7). We aggregated the posterior probabilities of SNPs in each defined approximately LD-independent regions to obtain expected number

of population-specific / shared causal variants at each region. For most analyzed traits, both population-specific and shared causal variants were widely spread across the entire genome (see Figure 5.8). As an example, mean corpuscular hemoglobin (MCH) harbored 0.68 (S.D. 0.42) EAS-specific, 0.53 (S.D. 0.40) EUR-specific, and 2.19 (S.D. 1.46) shared causal variants, per LD-independent region (see Figure 5.8). Interestingly, we found that for rheumatoid arthritis (RA), nearly all the population-specific causal variants concentrated in the MHC (chr6:25M–35M) region. Indeed, selection at MCH region may give rise to different causal variants in each population [105], although we caution that complex LD structures around the MHC region might introduced bias our estimates. Next, we aggregated the posterior probabilities by chromosome to obtain the expected number of population-specific / shared causal variants on each chromosome. The expected number of both population-specific and shared causal variants was highly proportional to the size of the chromosome, recapitulating similar findings using local SNP-heritability [91, 140].

5.3.4 GWAS risk regions contain multiple causal variants in both populations

We investigated whether genomic regions harboring significant associations (GWAS risk regions) in only one population contained causal variants that are shared by both populations. We quantified the expected number of shared as well as population-specific causal variants in regions that contained GWAS-significant associations in only the East Asian (EAS) GWAS or European (EUR) GWAS. First, regions with GWAS-significant associations in both EAS and EUR GWAS harbored multiple causal variants that were shared across the two populations (see Figure 5.9), recapitulating previous findings on allelic heterogeneity of complex traits [65, 140, 54]. Second, regions with GWAS-significant associations only in EAS or EUR GWAS contained causal variants shared by both populations (see Figure 5.9). For example, regions with GWAS-significant associations for mean corpuscular volume (MCV) only in the EAS / EUR GWAS harbored 3.0 (S.D. 1.7) / 3.3 (S.D. 1.5) shared causal variants on average, respectively (see Figure 5.9), consistent with previous study suggesting that the lack of shared GWAS associations in two continental populations is likely in part due

to heterogeneity in LD structures of the two populations [100]. In addition, we didn't observe a noticeable difference in the expected number of EAS-specific and EUR-specific causal variants at regions harboring GWAS-significant associations only in East Asian / European GWAS (see Figure 5.9). Finally, regions that contained no GWAS-significant association in either population also harbored multiple causal variants shared (see Figure 5.9), suggesting that a large fraction of causal variants of complex traits resided in sub-GWAS regions.

5.3.5 Enrichment analysis of population-specific and shared causal variants

Recent work has found enrichment of SNP-heritability in regions of specifically expressed genes (SEG) in trait-relevant tissues and cell types [43, 44]. Here, we investigated whether genetic architectures of complex traits were consistent in SEG in trait-relevant tissues across East Asians (EAS) and Europeans (EUR). First, we estimated enrichments of population-specific and shared causal variants for each analyzed trait in SEG across 53 GTEx tissues using SNP annotations described in [43], and tested for significance with a stringent Bonferroni corrected threshold of 0.05/53 (see Figure 5.10). All analyzed traits except major depressive disorder (MDD) exhibited significant enrichment of shared causal variants in at least one SEG in a trait-relevant tissue (see Figure 5.10), suggesting that in regions of genes that are expressed in trait-relevant tissues, there were more shared causal variants across EAS and EUR than the rest of the genome.

Next, we investigated whether there was heterogeneity in genetic architectures in SEG across EAS and EUR. Out of the 9 analyzed traits, 7 traits showed significant enrichment of population-specific causal variants in at least one SEG (see Figure 5.10). We highlight the two hematological traits, mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV), which displayed significant enrichments of both population-specific and shared causal variants in SEG in multiple tissues (see Figure 5.10). For example, MCV showed a 1.3x (S.E. 0.0041, p=$5.3 \times 10-14$) enrichment of shared causal variants in SEG in blood, and also a 1.1x (S.E. 0.0031, p=2.0×10^{-4}) and 1.3x (S.E. 0.0062, p=3.6×10^{-6}) enrichment of EAS-specific and EUR-specific causal variants in SEG in whole blood (see Figure 5.11),

respectively, suggesting that regions of genes that are expressed in whole blood harbor both more population-specific and shared causal variants across EAS and EUR than the rest of the genome.

5.4 Discussion

We have presented POSC, a method to dissect population-specific and shared causal variants of complex traits across two continental populations from GWAS summary association statistics data. POSC employs a Bayesian approach to explicitly model polygenicity of a trait and LD structures in both populations. In extensive simulations using either in-sample or external reference LD, POSC yielded accurate and robust estimates of proportion of population-specific / shared causal variants and well-calibrated statistics for testing enrichment in functional annotations. We applied POSC on 18 summary association statistics data of 9 complex traits obtained from samples of East Asian (EAS) and European (EUR) descent to glean insights into the underlying genetic architectures of complex traits in both populations.

First, we showed that while East Asian and European populations shared a large proportion of causal variants for multiple complex traits, each population also possessed a substantial proportion of population-specific causal variants. Second, our results suggested that regions that harbor GWAS risk variants for one population was enriched for causal variants in the other population, indicating that the lack of GWAS signal was likely attributable to differences in LD structure and power of GWAS. Third, our analysis of enrichment of population-specific / shared causal variants in SEG annotations [43] demonstrated that regions of genes expressed in trait-relevant tissues harbored an excess of both shared and population-specific causal variants for multiple complex traits. Overall, our analysis provides valuable insights into the underlying genetic architectures of complex traits in different populations, and highlights the importance of performing GWAS in non-European populations.

We conclude by highlighting caveats and limitations of our analysis. First, we note that the

estimates of proportion of population-specific and shared causal variants can be influenced by gene-environment interactions, and that one should exercise caution when interpreting these results. For example, if a SNP has effect on a trait only under the East Asian environment, then POSC will interpret that SNP as an EAS-specific causal variant, even though the SNP may still be biologically causal in Europeans. Second, our analysis only included genetic variants with minor allele frequency (MAF) greater than 1% in both populations for the sake of numerical stability. Therefore, our estimates of proportion of population-specific and shared causal variants will be an underestimate if there is a substantial proportion of rare variants contributing to the trait. We note however that a large fraction of trait variance can be attributable to common variants [168, 100]. Third, POSC relies on LD blocks that are approximately independent in both populations for computational efficiency, and will result in biased estimates in case of LD leakage. Thus, we recommend that LD blocks are specifically defined for each pair of population one analyzes. Fourth, we observed that performance of POSC decreases as the product between trait SNP-heritability and GWAS sample size decreases. Therefore, we recommend that POSC is applied only on highly heritable traits and (or) GWAS with large sample size. In light of the global efforts on developing biobanks [151, 111], we anticipate that future GWASs will have sufficient power to study traits with wide range of heritability. Fifth, POSC doesn't explicitly model correlation of the effect sizes of shared causal variants across two populations for computational efficiency. And we conjecture that explicitly modeling correlation of the trans-ethnic SNP effect sizes can further improve accuracy of POSC.

5.5 Tables

Trait name	Population	SNP-heritability (S.E.) %	Sample size	Reference
Body Mass Index (BMI)	EAS	19.8 (0.67)	224,698	[90]
	EUR	20.6 (0.91)	158,284	[1]
Mean Corpuscular Hemoglobin (MCH)	EAS	18.6 (2.2)	108,054	[71]
	EUR	22.7 (3.2)	172,332	[6]
Mean Corpuscular Volume (MCV)	EAS	21.0 (2.13)	108,256	[71]
	EUR	23.6 (3.1)	172,433	[6]
High Density Lipoprotein (HDL)	EAS	20.7 (3.03)	70,657	[71]
	EUR	16.4 (2.2)	89,614	[154]
Low Density Lipoprotein (LDL)	EAS	9.5 (1.3)	72,866	[71]
	EUR	13.6 (1.93)	85,491	[154]
Total Cholesterol (TC)	EAS	8.1 (0.84)	128,305	[71]
	EUR	22.5 (2.1)	89,865	[154]
Triglyceride (TG)	EAS	13.5 (3.3)	105,597	[71]
	EUR	13.6 (2.2)	86,502	[154]
Major Depressive Disorder (MDD)	EAS	35.6 (3.4)	10,640	[21]
	EUR	19.0 (1.8)	18,759	[166]
Rheumatoid Arthritis (RA)	EAS	28.9 (18.3)	22,515	[114]
	EUR	9.5 (1.9)	58,284	[114]

Table 5.1: **A list of GWAS summary statistics data set analyzed.** We obtain genome-wide SNP-heritability estimates of these traits using LD score regression [19], with intercept term constrained to 1.

Trait name	Total no. of SNPs	Estimated no. of EAS-specific causal SNPs (S.E.)	Estimated no. of EUR-specific causal SNPs (S.E.)	Estimated no. of shared causal SNPs (S.E.)	Estimated total no. of causal SNPs in EAS (S.E.)	Estimated total no. of causal SNPs in EUR (S.E.)
BMI	258,536	2,632 (58)	3,542 (79)	22,664 (141)	25,296 (85)	26,206 (66)
MCH	481,402	993 (13)	784 (11)	2,805 (14)	3,799 (22)	3,589 (12)
MCV	481,396	933 (10)	739 (5)	3,055 (14)	3,989 (16)	3,795 (16)
HDL	268,673	4,016 (19)	1,309 (39)	4,099 (16)	8,115 (14)	5,408 (30)
LDL	268,676	1,434 (19)	927 (23)	2,681 (22)	4,116 (23)	3,608 (13)
TC	268,672	2,467 (22)	2,003 (16)	3,889 (36)	6,356 (19)	5,892 (27)
TG	268,673	2,689 (10)	756 (12)	3,193 (11)	5,882 (12)	3,949 (12)
MDD	96,863	324 (15)	4,897 (24)	5,519 (51)	5,844 (41)	10,417 (53)
RA	529,404	187 (3)	539 (8)	689 (7)	877 (8)	1,228 (11)

Table 5.2: **Total number of SNPs, estimated number of population-specific and shared causal variants for BMI, MCH, and MCV.** We estimated the standard errors of the numbers of population-specific and shared causal variants using the last 25 iterations of the EM-MCMC algorithm for estimating the prior proportion of population-specific and shared causal variants.

5.6 Figures

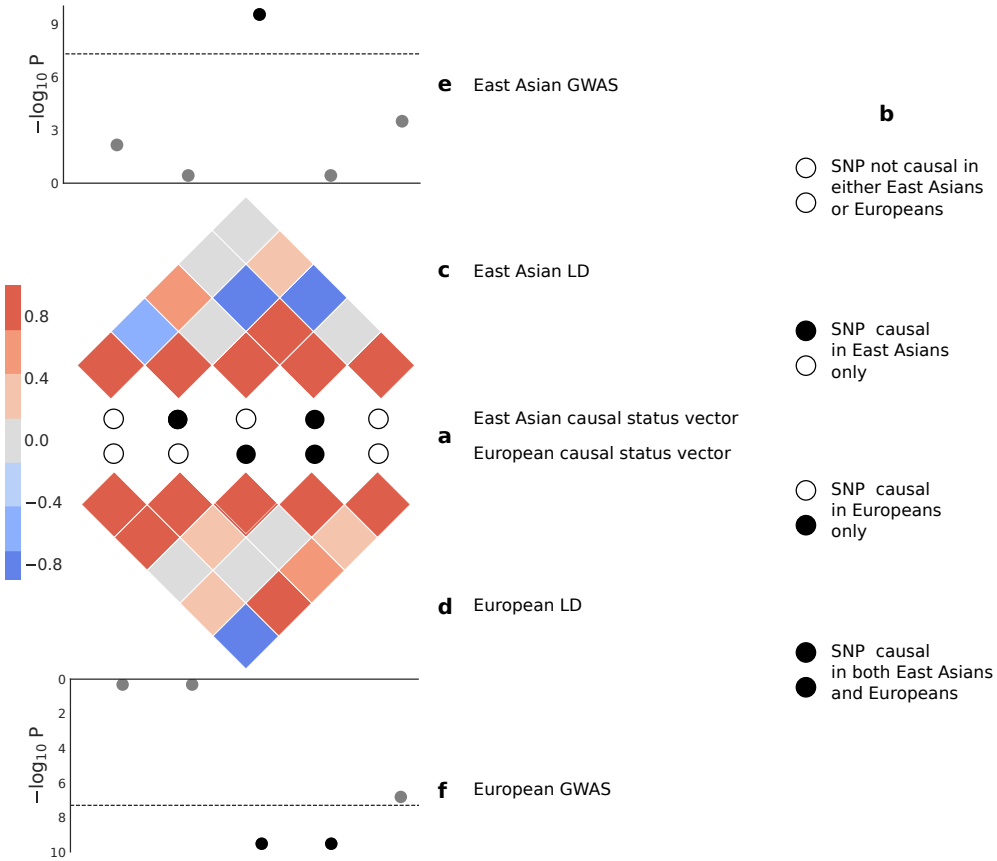


Figure 5.1: **Example of how differences in genetic architectures and LD pattern between East Asians and Europeans affect observed GWAS associations.** a) We use filled and unfilled circles to represent SNPs causal and not causal in each ancestral population. b) Four possible causal statuses of a SNP in the two ancestral populations. Namely, the SNP is not causal in either ancestral populations; the SNP is only causal in East Asians; the SNP is only causal in Europeans; the SNP is causal in both ancestral populations. c) and d) LD pattern in East Asian and European population, respectively. e) and f) Manhattan plots of GWASs in East Asians and Europeans, respectively. SNPs passing the significance threshold are marked in black.

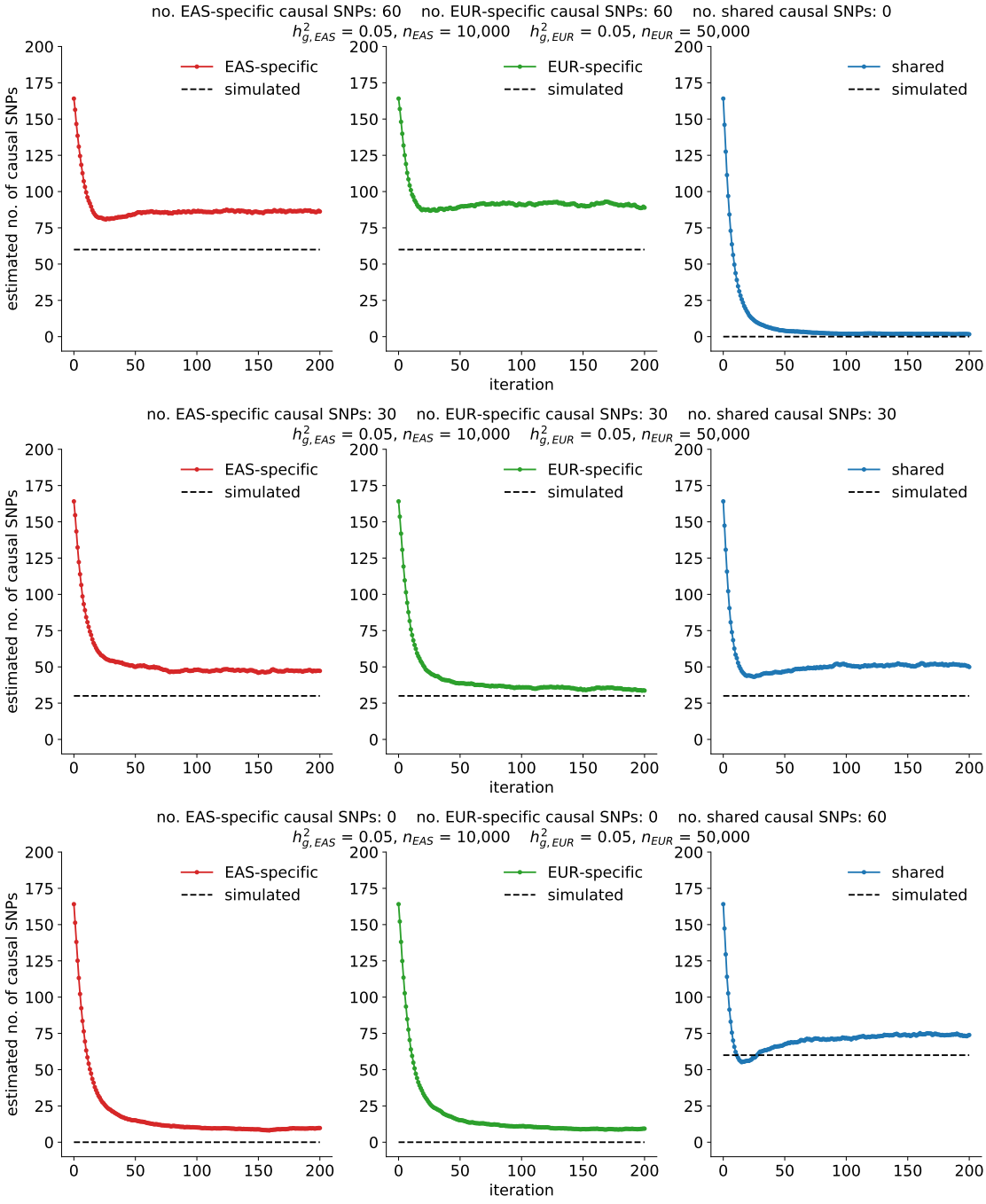


Figure 5.2: **Estimated number of population-specific and shared causal variants across iterations of the EM algorithm.** We randomly selected 60 causal SNPs (out of 8,599) in both populations, and set the product between SNP-heritability and GWAS sample size in both populations to 500. Each curve represents the average across 25 simulations.

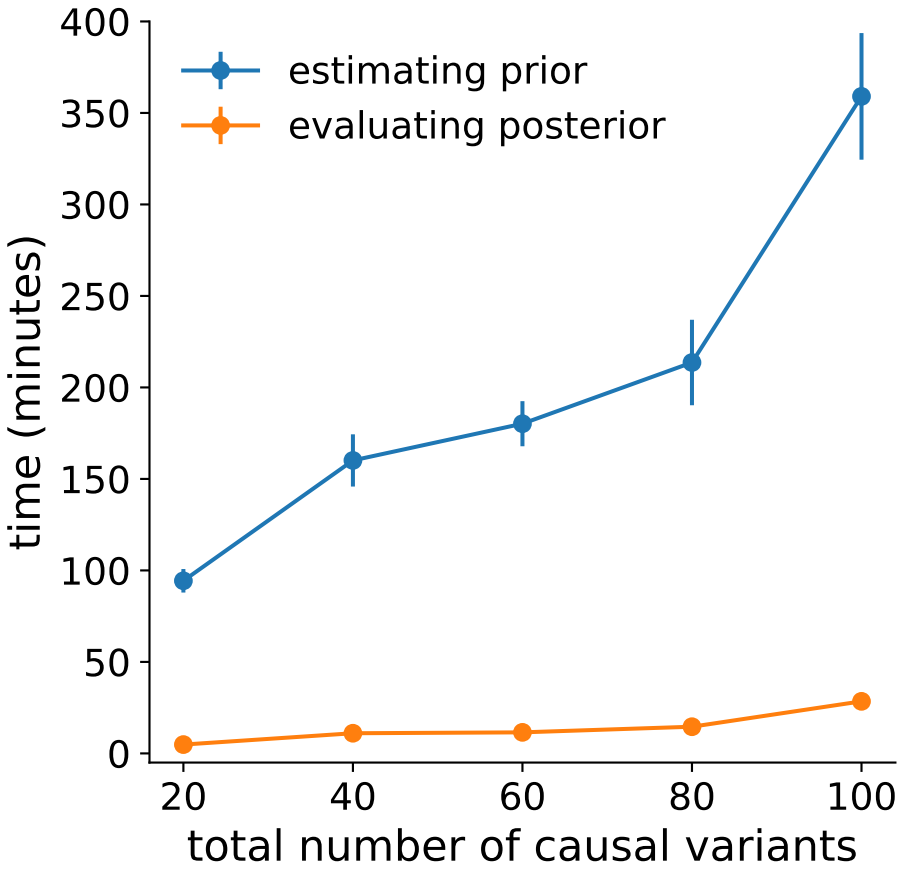


Figure 5.3: **Average run time for estimating the prior (MVB parameters) and evaluating per-SNP posterior probability to be population-specific and shared.** Each dot represents the average run time across all simulations with total causal variants in each population specified on the x-axis. Error bars represent 1.96 times the standard error on each side.

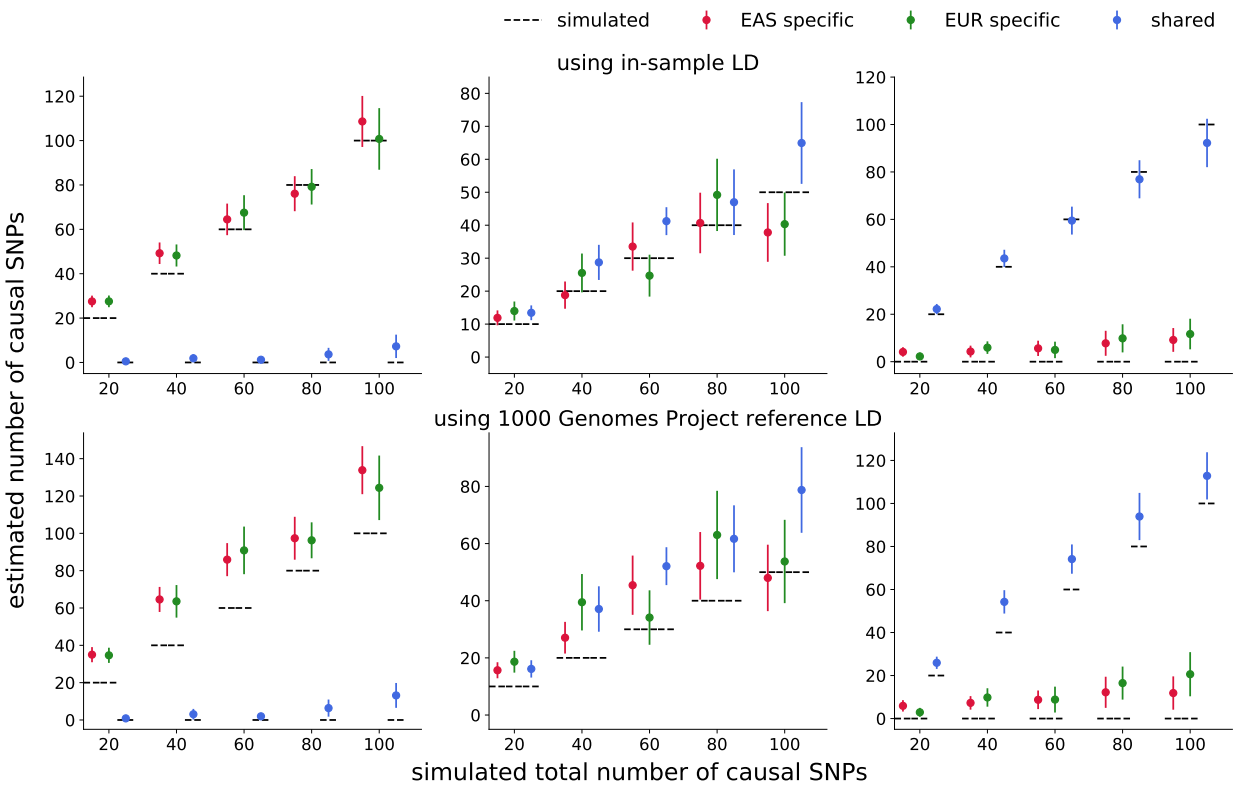


Figure 5.4: **Performance of POSC in simulations.** POSC yielded approximately unbiased estimates of the number of population-specific and shared causal variants in simulations when in-sample LD was used (top panel), and slightly upwardly biased estimates when external reference LD was used (bottom panel). We set the product of SNP-heritability of the trait and sample size of the GWAS to 500 in both populations. Mean and standard error were obtained across 25 simulations. Error bars represent 1.96 times the standard error on each side.

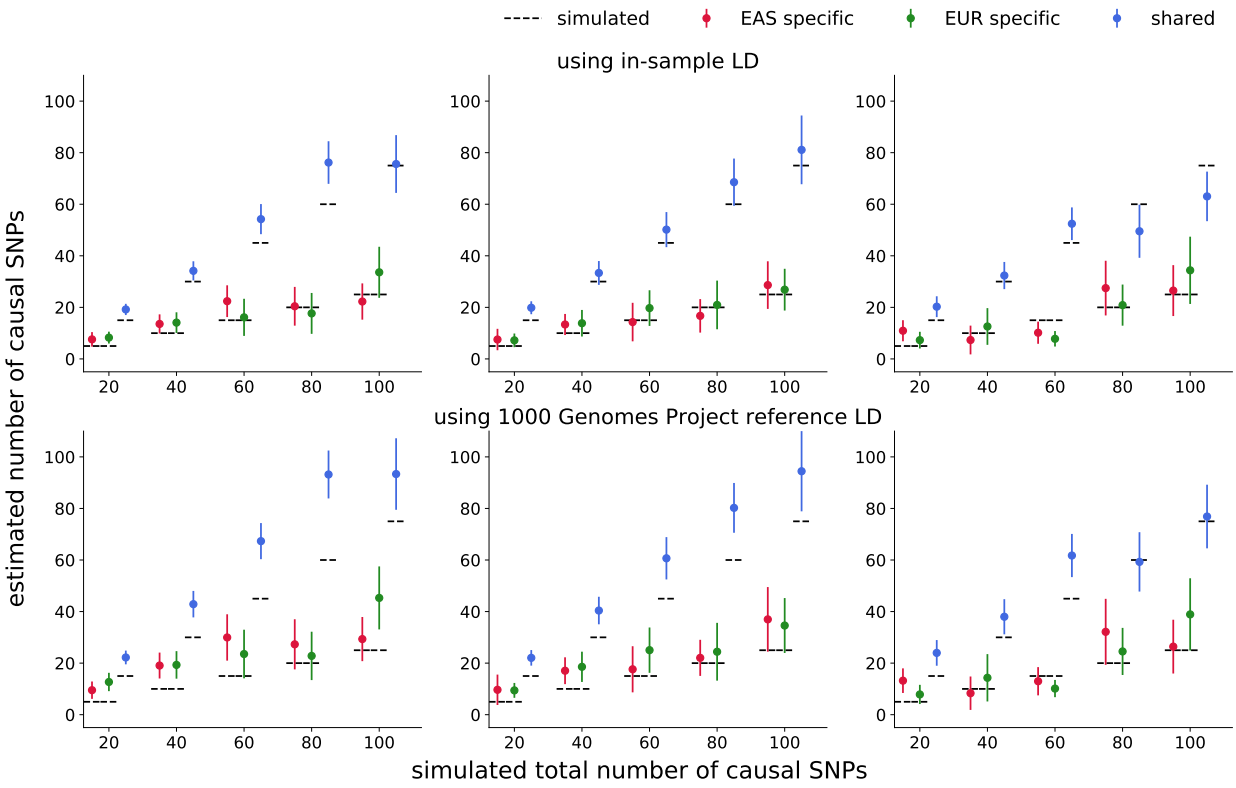


Figure 5.5: **Performance of POSC in simulations.** We simulated 20 to 100 causal variants for each population, where 75% of these causal variants were shared by both populations. We set the product between SNP-heritability of the trait and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column). Each dot represents the mean across 25 simulations, and errorbars represent 1.96 times the standard error on each side.

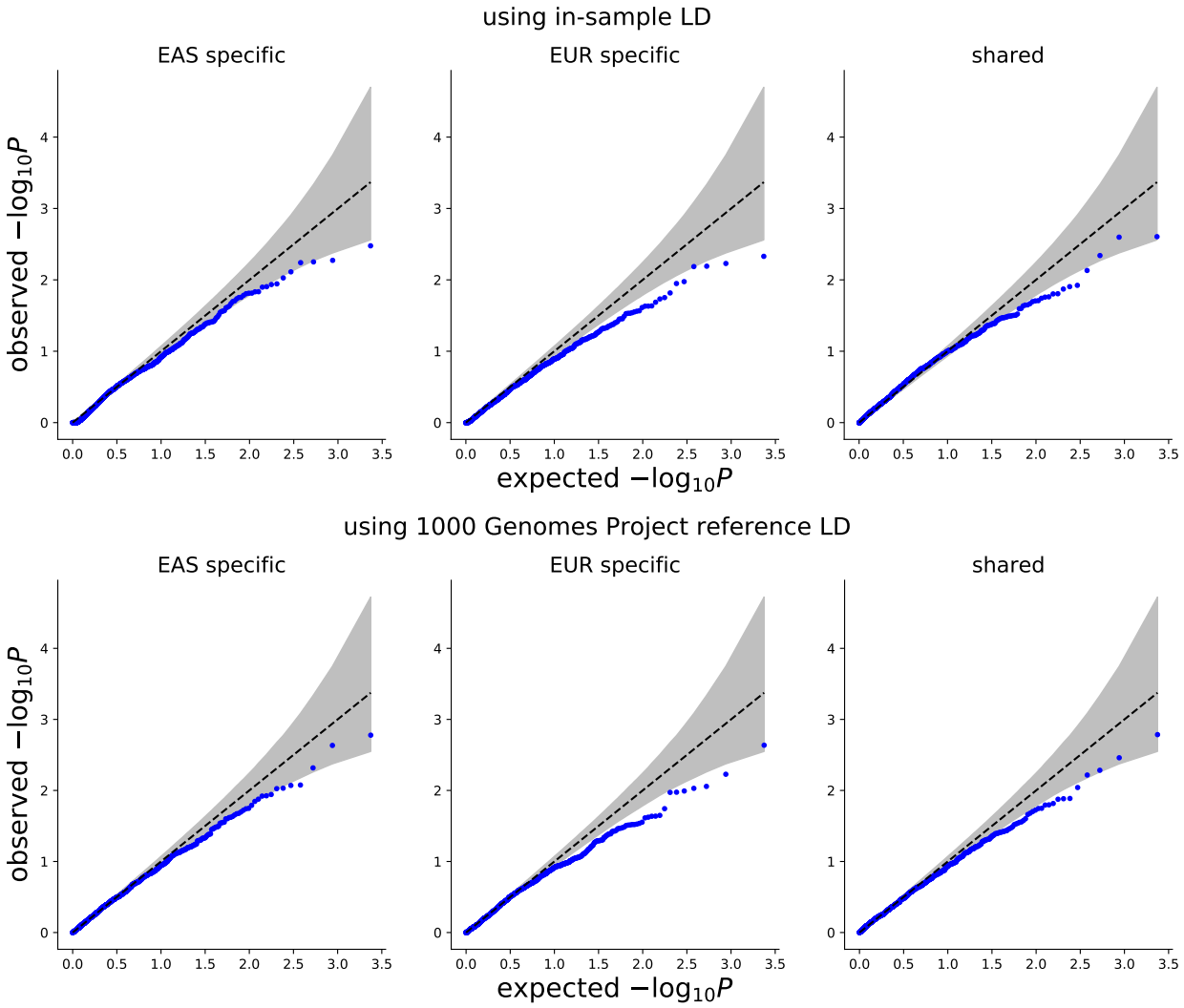


Figure 5.6: **Q-Q plot for p-values testing enrichment of population-specific and shared causal variants in SEG annotations [43].** We obtained of p-values for SEG annotations across 53 GTEx tissues from 25 null simulation, where we drew 25 EAS-specific, 25 EUR-specific, and 75 shared causal variants at random. In all simulations, we set the product of SNP-heritability of the trait and sample size of the GWAS to 500 in both populations. The top and bottom three figures represent results obtained using in-sample and 1000 Genomes Project reference LD matrix, respectively.

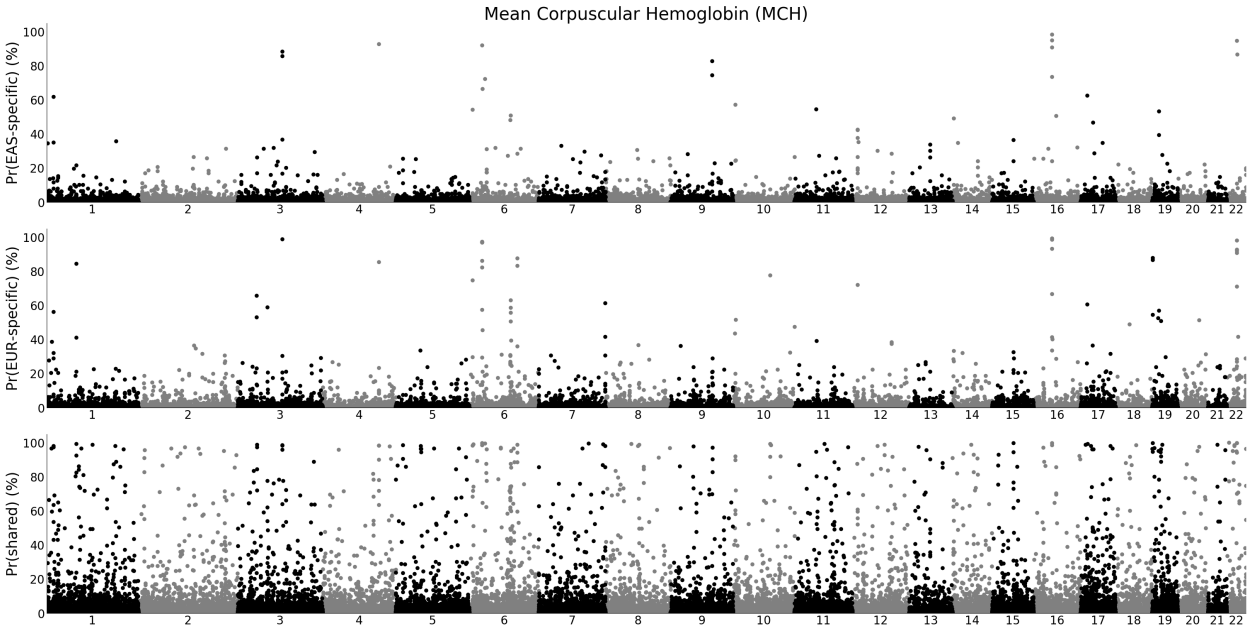


Figure 5.7: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for MCH.**

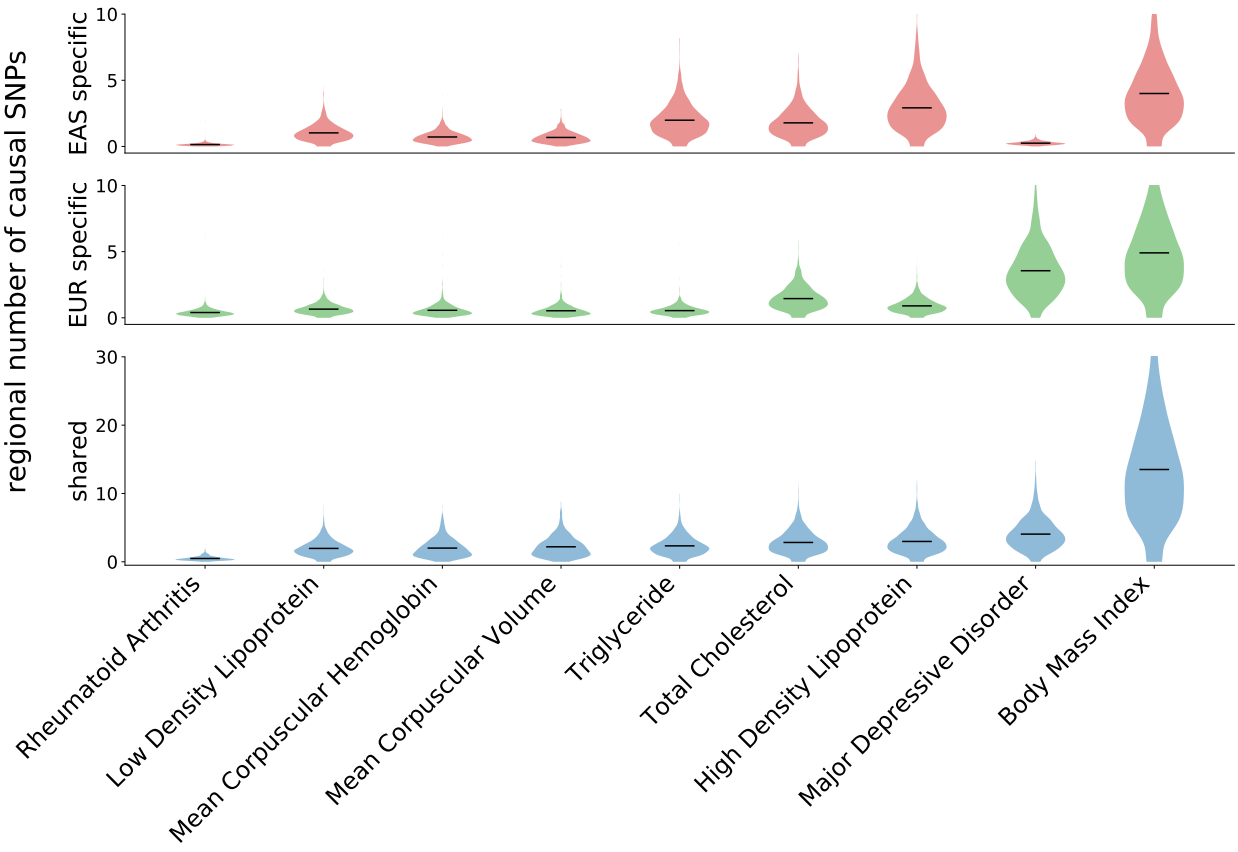


Figure 5.8: **Distribution of number of population-specific and shared causal variants per region.** Each violin plot shows the distribution of population-specific and shared causal variants across the genome, where the dark line represents the mean of the distribution. We sort the traits based on the average regional number of shared causal variants.

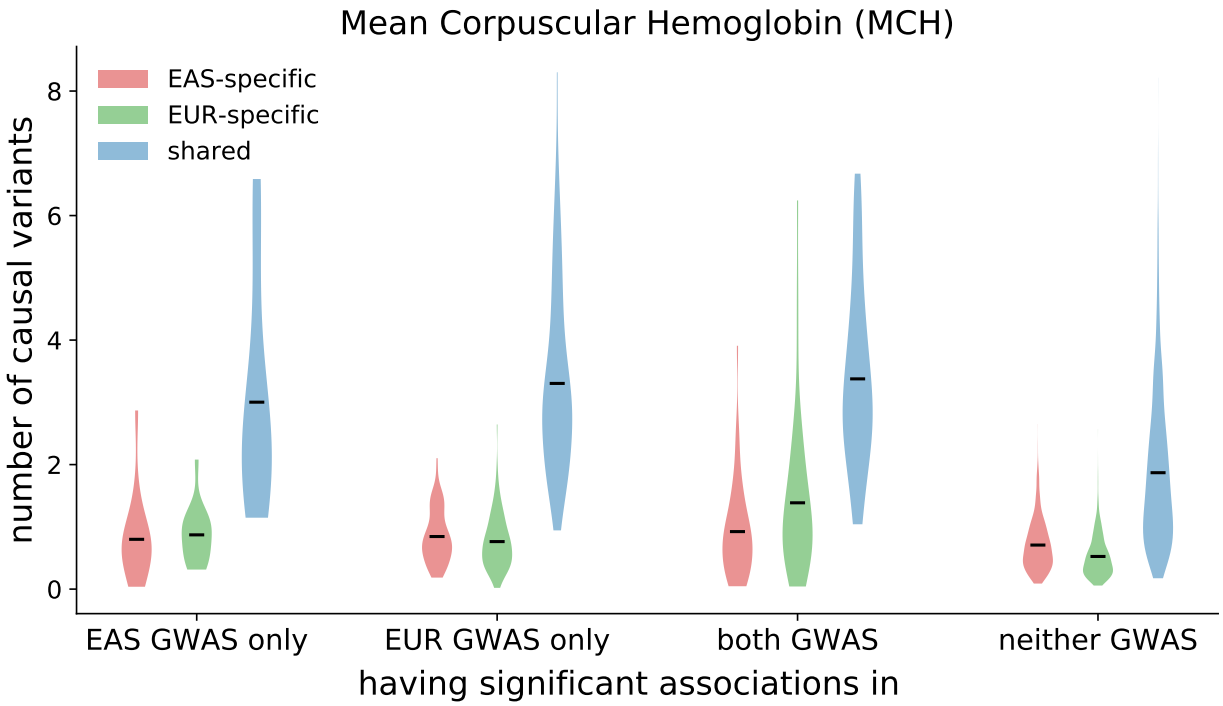


Figure 5.9: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.

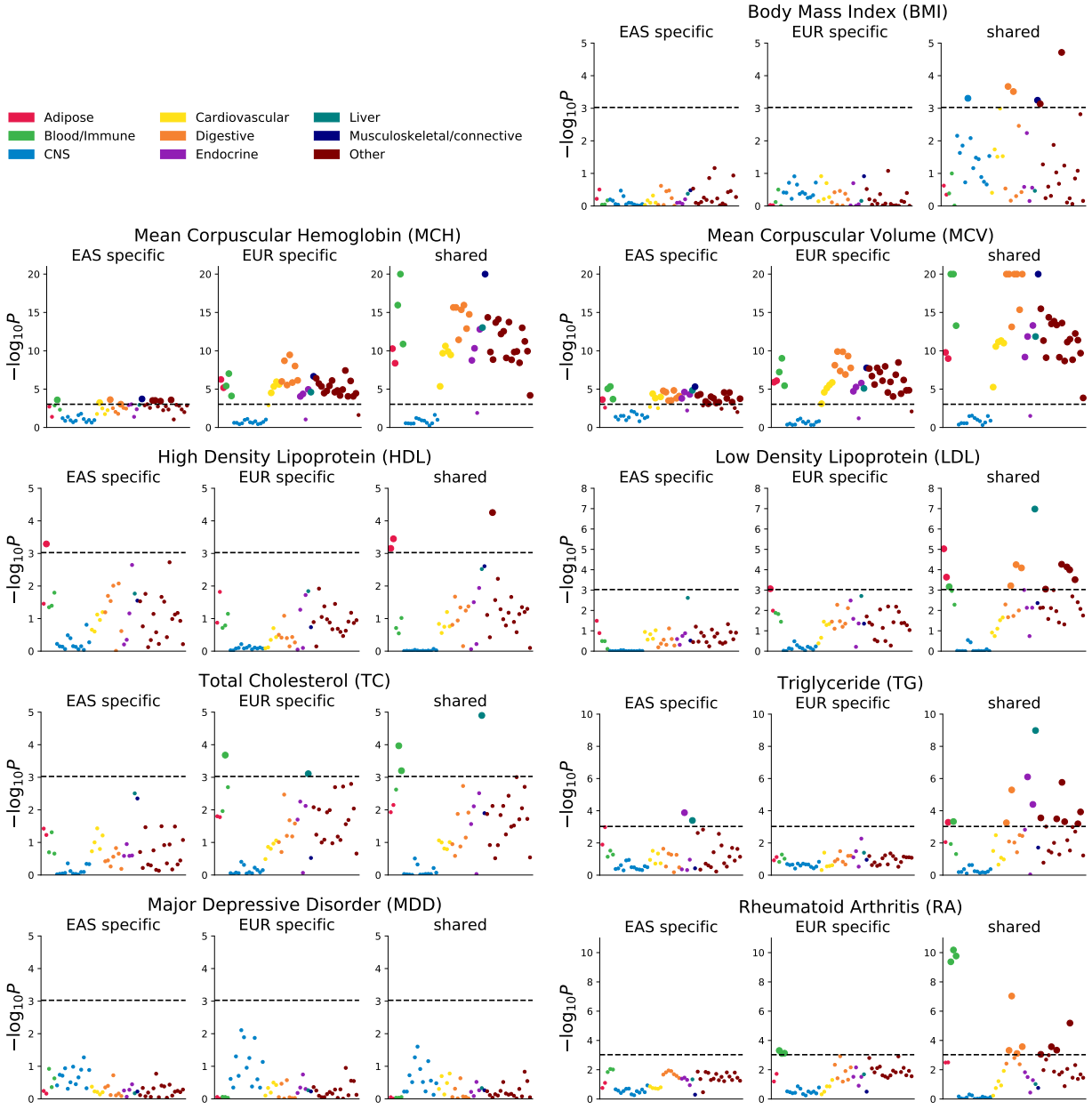


Figure 5.10: **Enrichment of population-specific and shared causal variants for BMI, MCH, and MCV in specifically expressed genes (SEG) annotations across 53 GTEx tissues.** We used a consistent significance threshold of $0.05 / 53$ ($-\log_{10} P = -3.03$) as represented by the dotted line to test for enrichment across all traits. We represent annotations passing the significance threshold using larger dots.

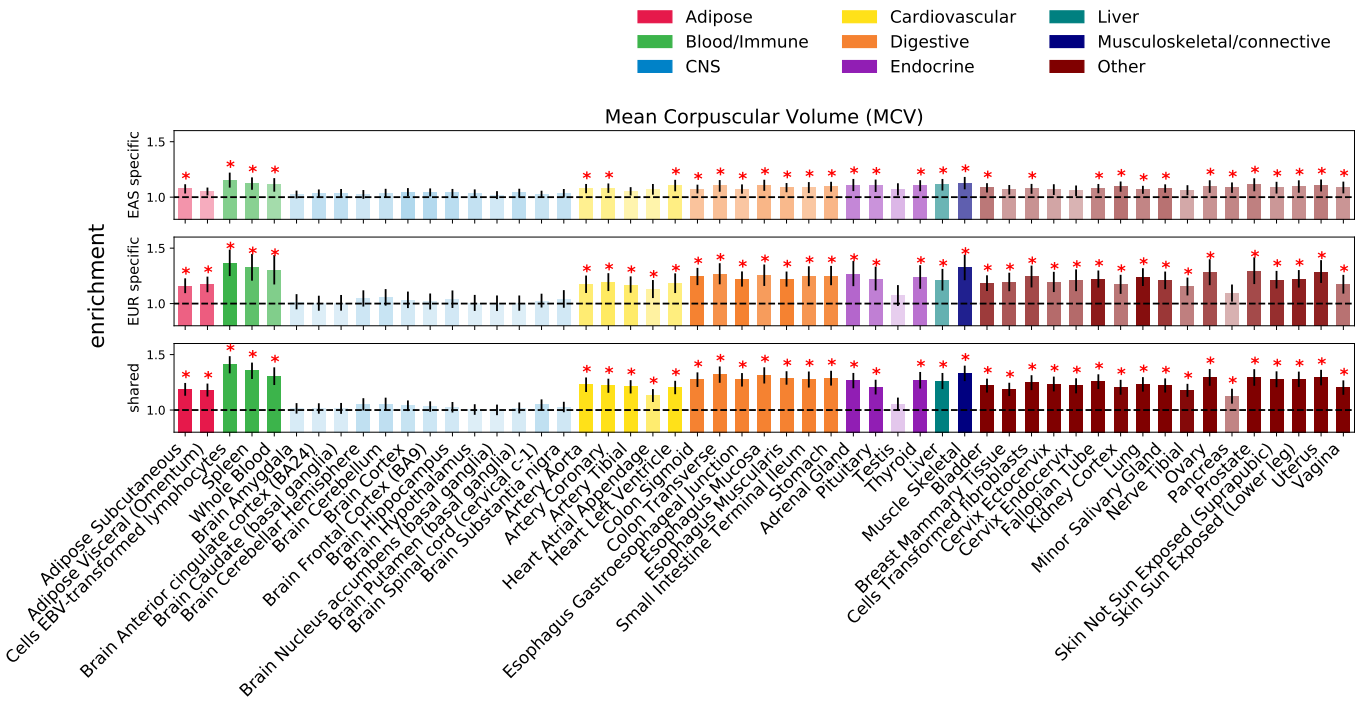


Figure 5.11: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than 0.05/53 with a star.

REFERENCES

[1] Masato Akiyama, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genome-wide association study identifies 112 new loci for body mass index in the japanese population. *Nature genetics*, 49(10):1458, 2017.

[2] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415, 2013.

[3] NK Arden, J Baker, C Hogg, K Baan, and TD Spector. The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *Journal of Bone and Mineral Research*, 11(4):530–534, 1996.

[4] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, 96(2):329–339, 2015.

[5] Jennifer L Asimit, Konstantinos Hatzikotoulas, Mark McCarthy, Andrew P Morris, and Eleftheria Zeggini. Trans-ethnic study design approaches for fine-mapping. *European Journal of Human Genetics*, 24(9):1330, 2016.

[6] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429, 2016.

[7] David P Baker, Juan Leon, Emily G Smith Greenaway, John Collins, and Marcela Movit. The education effect on population health: a reassessment. *Population and development review*, 37(2):307–332, 2011.

[8] Nicola Barban, Rick Jansen, Ronald de Vlaming, Ahmad Vaez, Jornt J Mandemakers, Felix C Tropf, Xia Shen, James F Wilson, Daniel I Chasman, Ilja M Nolte, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature genetics*, 2016.

[9] Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

[10] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[11] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, page btv546, 2015.

[12] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics (Oxford, England)*, 32(2):283, 2016.

[13] Dorret Boomsma, Andreas Busjahn, and Leena Peltonen. Classical twin studies and beyond. *Nature reviews genetics*, 3(11):872–882, 2002.

[14] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.

[15] Brielin C Brown, Chun Jimmie Ye, Alkes L Price, Noah Zaitlen, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, et al. Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics*, 99(1):76–88, 2016.

[16] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

[17] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[18] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 2015.

[19] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[20] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[21] Na Cai, Tim B Bigdeli, Warren Kretzschmar, Yihan Li, Jieqin Liang, Li Song, Jingchu Hu, Qibin Li, Wei Jin, Zhenfei Hu, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*, 523(7562):588, 2015.

[22] Gregory Carey. Inference about genetic correlations. *Behavior genetics*, 18(3):329–338, 1988.

[23] Zhao Chen, Hua Tang, Rehan Qayyum, Ursula M Schick, Michael A Nalls, Robert Handsaker, Jin Li, Yingchang Lu, Lisa R Yanek, Brendan Keating, et al. Genome-wide association analysis of red blood cell traits in african americans: the cogent network. *Human molecular genetics*, 22(12):2529–2538, 2013.

[24] Charles C Chung, Peter A Kanetsky, Zhaoming Wang, Michelle A T Hildebrandt, Roelof Koster, Rolf I Skotheim, Christian P Kratz, Clare Turnbull, Victoria K Cortessis, Anne C Bakken, D. Timothy Bishop, Michael B Cook, R. Loren Erickson, Sophie D Foss, Kevin B Jacobs, Larissa A Korde, Sigrid M Kraggerud, Ragnhild A Lothe, Jennifer T Loud, Nazneen Rahman, Eila C Skinner, Duncan C Thomas, Xifeng Wu, Meredith Yeager, Fredrick R Schumacher, Mark H Greene, Stephen M Schwartz, Katherine A McGlynn, Stephen J Chanock, and Katherine L Nathanson. Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nature Genetics*, 45(6):680–685, Jun 2013.

[25] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015.

[26] Robin Z Cohen, Mary V Seeman, Andrew Gotowiec, and Lili Kopala. Earlier puberty as a predictor of later onset of schizophrenia in women. *American Journal of Psychiatry*, 1999.

[27] 1000 Genomes Project Consortium, Gonalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[28] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[29] Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 2013.

[30] Bin Dai, Shilin Ding, Grace Wahba, et al. Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 2013.

[31] Mark J. Daly, John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson, and Eric S. Lande. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(6):229–232, Jun 2001.

[32] George Davey Smith and Shah Ebrahim. mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

[33] Felix R Day, Katherine S Ruth, Deborah J Thompson, Kathryn L Lunetta, Natalia

Pervjakova, Daniel I Chasman, Lisette Stolk, Hilary K Finucane, Patrick Sulem, Brendan Bulik-Sullivan, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and brca1-mediated dna repair. *Nature genetics*, 47(11):1294–1303, 2015.

[34] Gustavo de los Campos, Ana I Vazquez, Rohan Fernando, Yann C Klimentidis, and Daniel Sorensen. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS genetics*, 9(7):e1003608, 2013.

[35] SV de Miranda Chagas, S Kanaan, H Chung Kang, M Cagy, RE de Abreu, LA da Silva, RC Garcia, and ML Garcia Rosa. Environmental factors, familial aggregation and heritability of total cholesterol, low density lipoprotein-cholesterol and high density lipoprotein-cholesterol in a brazilian population assisted by the family doctor program. *public health*, 125(6):329–337, 2011.

[36] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, et al. Dnase [thinsp] i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[37] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[38] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116, 2010.

[39] Bradley Efron. Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, 6(4):1971, 2012.

[40] SC Elbein and SJ Hasstedt. Quantitative trait linkage analysis of lipid-related traits in familial type 2 diabetes evidence for linkage of triglyceride levels to chromosome 19q. *Diabetes*, 51(2):528–535, 2002.

[41] Richard S Elman, Nikita Karpenko, and Alexander Merkurjev. *The algebraic and geometric theory of quadratic forms*, volume 56. American Mathematical Soc., 2008.

[42] Ruben N Eppinga, Yanick Hagemeijer, Stephen Burgess, David A Hinds, Kari Stefansson, Daniel F Gudbjartsson, Dirk J van Veldhuisen, Patricia B Munroe, Niek Verweij, and Pim van der Harst. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nature Genetics*, 48(12):1557–1563, 2016.

[43] Hilary Finucane, Yakir Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Giulio Genovese, Arpiar Saunders, et al.

Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv*, page 103069, 2017.

[44] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

[45] Eric R Gamazon, Nancy J Cox, and Lea K Davis. Structural architecture of snp effects on complex traits. *The American Journal of Human Genetics*, 95(5):477–489, 2014.

[46] Santhi K Ganesh, Neil A Zakai, Frank JA van Rooij, Nicole Soranzo, Albert V Smith, Michael A Nalls, Ming-Huei Chen, Anna Kottgen, Nicole L Glazer, Abbas Dehghan, et al. Multiple loci influence erythrocyte phenotypes in the charge consortium. *Nature genetics*, 41(11):1191–1198, 2009.

[47] C Garner, T Tatu, JE Reittie, T Littlewood, J Darley, S Cervino, M Farrall, P Kelly, TD Spector, and SL Thein. Genetic influences on f cells and other hematologic variables: a twin heritability study. *Blood*, 95(1):342–346, 2000.

[48] Godfrey S Getz and Catherine A Reardon. Apoprotein e as a lipid transport and signaling protein in the blood, liver, and artery wall. *Journal of lipid research*, 50(Supplement):S156–S161, 2009.

[49] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383, 2014.

[50] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, et al. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003.

[51] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[52] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H Goodall, Yann Labrune, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, 2011.

[53] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[54] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya

Raychaudhuri, et al. Quantifying missing heritability at known gwas loci. *Plos Genetics*, 2013.

[55] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 2016.

[56] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[57] Per Christian Hansen. The truncatedsvd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[58] SMJ Harney, C Vilariño-Güell, IE Adamopoulos, A-M Sims, RW Lawrence, LR Cardon, JL Newton, C Meisel, JJ Pointon, C Darke, et al. Fine mapping of the mhc class iii region demonstrates association of aif1 and rheumatoid arthritis. *Rheumatology*, 47(12):1761–1767, 2008.

[59] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[60] Joseph P Hegmann and Bernard Possidente. Estimating genetic correlations from inbred strains. *Behavior genetics*, 11(2):103–114, 1981.

[61] Gibran Hemani, Jian Yang, Anna Vinkhuyzen, Joseph E Powell, Gonneke Willemsen, Jouke-Jan Hottenga, Abdel Abdellaoui, Massimo Mangino, Ana M Valdes, Sarah E Medland, et al. Inference of the genetic architecture underlying bmi and height with the use of 20,240 sibling pairs. *The American Journal of Human Genetics*, 93(5):865–875, 2013.

[62] Jesse D Hinckley, Diana Abbott, Trudy L Burns, Meadow Heiman, Amy D Shapiro, Kai Wang, and Jorge Di Paola. Quantitative trait locus linkage analysis in a large amish pedigree identifies novel candidate loci for erythrocyte traits. *Molecular genetics & genomic medicine*, 1(3):131–141, 2013.

[63] Momoko Horikoshi, Robin N Beaumont, Felix R Day, Nicole M Warrington, Marjolein N Kooijman, Juan Fernandez-Tajes, Bjarke Feenstra, Natalie R van Zuydam, Kyle J Gaulton, Niels Grarup, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*, 538(7624):248–252, 2016.

[64] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[65] Farhad Hormozdiari, Anthony Zhu, Gleb Kichaev, Chelsea J-T Ju, Ayellet V Segre, Jong Wha J Joo, Hyejung Won, Sriram Sankararaman, Bogdan Pasaniuc, Sagiv Shifman, et al. Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789–802, 2017.

[66] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 2012.

[67] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959, Aug 2012.

[68] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, Jun 2009.

[69] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

[70] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[71] Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, 50(3):390, 2018.

[72] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271, 2015.

[73] Gleb Kichaev, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindstroem, Peter Kraft, and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2):248–255, 2017.

[74] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *Plos Genetics*, 2014.

[75] Daniel L Koller, Hou-Feng Zheng, David Karasik, Laura Yerges-Armstrong, Ching-Ti Liu, Fiona McGuigan, John P Kemp, Sylvie Giroux, Dongbing Lai, Howard J Edenberg, et al. Meta-analysis of genome-wide studies identifies wnt16 and esr1 snps associated with bone mineral density in premenopausal women. *Journal of Bone and Mineral Research*, 28(3):547–558, 2013.

[76] Anna Köttgen, Eva Albrecht, Alexander Teumer, Veronique Vitart, Jan Krumsiek, Claudia Hundertmark, Giorgio Pistis, Daniela Ruggiero, Conall M O'Seaghdha, Toomas Haller, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature genetics*, 45(2):145–154, 2013.

[77] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22(2):139–144, Jun 1999.

[78] K. Lange. *Applied Probability*. Springer Texts in Statistics. Springer New York, 2010.

[79] K. Lange. *Optimization*. Springer Texts in Statistics. Springer, 2013.

[80] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[81] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, 2012.

[82] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[83] Sang Hong Lee, Jian Yang, Michael E Goddard, Peter M Visscher, and Naomi R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

[84] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[85] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[86] Yun Li, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonalo R. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–834, Dec 2010.

[87] Zhiqiang Li, Jianhua Chen, Hao Yu, Lin He, Yifeng Xu, Dai Zhang, Qizhong Yi, Changgui Li, Xingwang Li, Jiawei Shen, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature genetics*, 49(11):1576, 2017.

[88] Jing-Ping Lin, Christopher J O'Donnell, Li Jin, Caroline Fox, Qiong Yang, and L Adri-

enne Cupples. Evidence for linkage of red blood cell size and count: Genome-wide scans in the framingham heart study. *American journal of hematology*, 82(7):605–610, 2007.

[89] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, 2015.

[90] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[91] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 2015.

[92] Kirk E. Lohmueller, Carlos D. Bustamante, and Andrew G. Clark. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, 182(1):217–231, May 2009.

[93] Yingchang Lu, Felix R Day, Stefan Gustafsson, Martin L Buchkovich, Jianbo Na, Veronique Bataille, Diana L Cousminer, Zari Dastani, Alexander W Drong, Tõnu Esko, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications*, 7, 2016.

[94] Pedro Madrigal and Paweł Krajewski. Current bioinformatic approaches to identify dnase i hypersensitive sites and genomic footprints from dnase-seq data. *Frontiers in genetics*, 3, 2012.

[95] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21, 2008.

[96] Nicholas Mancuso, Nadin Rohland, Kristin A Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, Heng Li, Alex Stram, Xin Sheng, et al. The contribution of rare variation to prostate cancer heritability. *Nature genetics*, 2015.

[97] Nicholas Mancuso, Nadin Rohland, Kristin A Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, Heng Li, Alex Stram, Xin Sheng, et al. The contribution of rare variation to prostate cancer heritability. *Nature genetics*, 48(1):30, 2016.

[98] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

[99] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, Jul 2007.

[100] Urko M Marigorta and Arcadi Navarro. High trans-ethnic replicability of gwas results implies common causal variants. *PLoS genetics*, 9(6):e1003566, 2013.

[101] Carla Márquez-Luna, Po-Ru Loh, and Alkes L Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, 41(8):811–823, 2017.

[102] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[103] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Richard Durbin, Goncalo Abecasis, and Jonathan Marchini. A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv*, page 035170, 2015.

[104] Joel Mefford and John S Witte. The covariate's dilemma. *PLoS Genet*, 8(11):e1003096, 2012.

[105] Diogo Meyer and Glenys Thomson. How selection shapes variation of the human major histocompatibility complex: a review. *Annals of human genetics*, 65(1):1–26, 2001.

[106] GW Mills, PJ Avery, MI McCarthy, AT Hattersley, JC Levy, GA Hitman, M Sampson, and M Walker. Heritability estimates for beta cell function and features of the insulin resistance syndrome in uk families with an increased susceptibility to type 2 diabetes. *Diabetologia*, 47(4):732–738, 2004.

[107] Andrew P Morris. A flexible bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *The American Journal of Human Genetics*, 79(4):679–694, 2006.

[108] Andrew P Morris. Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology*, 35(8):809–822, 2011.

[109] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[110] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, et al.

From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, 2010.

[111] Akiko Nagai, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, Akiko Tamakoshi, Zentaro Yamagata, Taisei Mushiroda, et al. Overview of the biobank japan project: study design and profile. *Journal of epidemiology*, 27(3):S2–S8, 2017.

[112] Michael Neale and Lon Cardon. *Methodology for genetic studies of twins and families*, volume 67. Springer Science & Business Media, 1992.

[113] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[114] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[115] Aysu Okbay, Bart ML Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G Nivard, Mark Alan Fontana, S Fleur W Meddens, Richard Karlsson Linnér, Cornelius A Rietveld, Jaime Derringer, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics*, 2016.

[116] Aysu Okbay, Jonathan P Beauchamp, Mark Alan Fontana, James J Lee, Tune H Pers, Cornelius A Rietveld, Patrick Turley, Guo-Bo Chen, Valur Emilsson, S Fleur W Meddens, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, 2016.

[117] Luigi Palla and Frank Dudbridge. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics*, 97(2):250–259, 2015.

[118] C Pallaud, R Gueguen, C Sass, M Grow, S Cheng, G Siest, and S Visvikis. Genetic influences on lipid metabolism trait variability within the stanislas cohort. *Journal of lipid research*, 42(11):1879–1890, 2001.

[119] Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*, advance online publication, Nov 2016. Review.

[120] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117, 2017.

[121] Bogdan Pasaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference

of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, Jun 2009.

[122] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, page btu416, 2014.

[123] Cristian Pattaro, Alexander Teumer, Mathias Gorski, Audrey Y Chu, Man Li, Vladan Mijatovic, Maija Garnaas, Adrienne Tin, Rossella Sorice, Yong Li, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nature communications*, 7, 2016.

[124] John RB Perry, Felix Day, Cathy E Elks, Patrick Sulem, Deborah J Thompson, Teresa Ferreira, Chunyan He, Daniel I Chasman, Tõnu Esko, Gudmar Thorleifsson, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, 2014.

[125] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics*, 2016.

[126] John E. Pool, Ines Hellmann, Jeffrey D. Jensen, and Rasmus Nielsen. Population genetic inference from genomic sequence variation. *Genome Res*, 20(3):291–300, Mar 2010.

[127] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161, 2016.

[128] P Poulsen, K Ohm Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, 1999.

[129] Alkes L Price, Chris CA Spencer, and Peter Donnelly. Progress and promise in understanding the genetic basis of common diseases. In *Proc. R. Soc. B*, volume 282, page 20151684. The Royal Society, 2015.

[130] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, Jun 2009.

[131] Alkes L Price, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, Jerome I Rotter, Esther Torres, Kent D Taylor, et al. Long-range ld can confound genome scans in admixed populations. *American journal of human genetics*, 83(1):132, 2008.

[132] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[133] LJ Rasmussen-Torvik, JS Pankow, DR Jacobs, LM Steffen, AM Moran, J Steinberger, and AR Sinaiko. Heritability and genetic correlations of insulin sensitivity measured by the euglycaemic clamp. *Diabetic Medicine*, 24(11):1286–1289, 2007.

[134] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tonu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, 340(6139):1467–1471, 2013.

[135] Noah A Rosenberg, Lucy Huang, Ethan M Jewett, Zachary A Szpiech, Ivana Jankovic, and Michael Boehnke. Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356, 2010.

[136] Sharon A Savage, Lisa Mirabello, Zhaoming Wang, Julie M Gastier-Foster, Richard Gorlick, Chand Khanna, Adrienne M Flanagan, Roberto Tirabosco, Irene L Andrulis, Jay S Wunder, Nalan Gokgoz, Ana Patio-Garcia, Luis Sierrasesmaga, Fernando Lecanda, Nilgn Kurucu, Inci Ergurhan Ilhan, Neriman Sari, Massimo Serra, Claudia Hattinger, Piero Picci, Logan G Spector, Donald A Barkauskas, Neyssa Marina, Silvia Regina Caminada de Toledo, Antonio S Petrilli, Maria Fernanda Amary, Dina Halai, David M Thomas, Chester Douglass, Paul S Meltzer, Kevin Jacobs, Charles C Chung, Sonja I Berndt, Mark P Purdue, Neil E Caporaso, Margaret Tucker, Nathaniel Rothman, Maria Teresa Landi, Debra T Silverman, Peter Kraft, David J Hunter, Nuria Malats, Manolis Kogevinas, Sholom Wacholder, Rebecca Troisi, Lee Helman, Joseph F Fraumeni, Meredith Yeager, Robert N Hoover, and Stephen J Chanock. Genome-wide association study identifies two susceptibility loci for osteosarcoma. *Nature Genetics*, 45(7):799–803, Jul 2013.

[137] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[138] Shayle R Searle. *Linear models*, page 65. John Wiley & Sons, Inc., 1971.

[139] Nuala A Sheehan, Vanessa Didelez, Paul R Burton, and Martin D Tobin. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med*, 5(8):e177, 2008.

[140] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *bioRxiv*, page 035907, 2016.

[141] Huwenbo Shi, Bogdan Pasaniuc, and Kenneth L Lange. A multivariate bernoulli

model to predict dnasei hypersensitivity status from haplotype data. *Bioinformatics*, 31(21):3514–3521, 2015.

[142] Dmitry Shungin, Thomas W Winkler, Damien C Croteau-Chonka, Teresa Ferreira, Adam E Locke, Reedik Mägi, Rona J Strawbridge, Tune H Pers, Krista Fischer, Anne E Justice, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015.

[143] Mary A Silles. The causal effect of education on health: Evidence from the united kingdom. *Economics of Education review*, 28(1):122–128, 2009.

[144] Xueling Sim, Rick Twee-Hee Ong, Chen Suo, Wan-Ting Tay, Jianjun Liu, Daniel Peng-Keat Ng, Michael Boehnke, Kee-Seng Chia, Tien-Yin Wong, Mark Seielstad, et al. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from southeast asia. *PLoS genetics*, 7(4):e1001363, 2011.

[145] Annemarie Simonis-Bik, Elisabeth MW Eekhoff, Michaela Diamant, Dorret I Boomsma, Rob J Heine, Jacqueline M Dekker, Gonneke Willemsen, Marieke Van Leeuwen, and Eco JC De Geus. The heritability of hba1c and fasting blood glucose in different measurement settings. *Twin Research and Human Genetics*, 11(06):597–602, 2008.

[146] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

[147] Nicole Soranzo, Serena Sanna, Eleanor Wheeler, Christian Gieger, Dörte Radke, Josée Dupuis, Nabila Bouatia-Naji, Claudia Langenberg, Inga Prokopenko, Elliot Stolerman, et al. Common variants at 10 genomic loci influence hemoglobin a1c levels via glycemic and nonglycemic pathways. *Diabetes*, 59(12):3229–3239, 2010.

[148] Nicole Soranzo, Tim D Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the haemgen consortium. *Nature genetics*, 41(11):1182–1190, 2009.

[149] Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986, 2017.

[150] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011.

[151] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank:

an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[152] Patrick F Sullivan, Kenneth S Kendler, and Michael C Neale. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry*, 60(12):1187–1192, 2003.

[153] Alan R Templeton. Haplotype trees and modern human origins. *American journal of physical anthropology*, 128(S41):33–59, 2005.

[154] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707, 2010.

[155] Alexander Teumer, Adrienne Tin, Rossella Sorice, Mathias Gorski, Nan Cher Yeo, Audrey Y Chu, Man Li, Yong Li, Vladan Mijatovic, Yi-An Ko, et al. Genome-wide association studies identify genetic loci associated with albuminuria in diabetes. *Diabetes*, page db151313, 2015.

[156] B Towne. Heritability of age at menarche in girls from the fels longitudinal study. *American journal of physical anthropology*, 128(1):210, 2005.

[157] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–130, 2013.

[158] Stephen Turner, Loren L Armstrong, Yuki Bradford, Christopher S Carlson, Dana C Crawford, Andrew T Crenshaw, Mariza Andrade, Kimberly F Doheny, Jonathan L Haines, Geoffrey Hayes, et al. Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, pages 1–19, 2011.

[159] C Tysk, E Lindberg, G Järnerot, and B Floderus-Myrhed. Ulcerative colitis and crohn's disease in an unselected population of monozygotic and dizygotic twins. a study of heritability and the influence of smoking. *Gut*, 29(7):990–996, 1988.

[160] Pim van der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–375, 2012.

[161] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[162] Benjamin F Voight, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580, 2012.

[163] Jeffrey D Wall and Jonathan K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597, 2003.

[164] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[165] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[166] Naomi R Wray, Patrick F Sullivan, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv*, page 167577, 2017.

[167] Ying Wu, Lindsay L Waite, Anne U Jackson, Wayne HH Sheu, Steven Buyske, Devin Absher, Donna K Arnett, Eric Boerwinkle, Lori L Bonnycastle, Cara L Carty, et al. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS genetics*, 9(3):e1003379, 2013.

[168] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[169] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[170] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[171] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

[172] Wen-Yun Yang, Farhad Hormozdiari, Eleazar Eskin, and Bogdan Pasaniuc. A spatial-aware haplotype copying model with applications to genotype imputation. In *Research in Computational Molecular Biology*, pages 371–384. Springer, 2014.

[173] Noah Zaitlen, Bogdan Paşaniuc, Tom Gur, Elad Ziv, and Eran Halperin. Leveraging genetic variability across populations for the identification of causal variants. *The American Journal of Human Genetics*, 86(1):23–33, 2010.

[174] Noah Zaitlen, Bogdan Pasaniuc, Sriram Sankararaman, Gaurav Bhatia, Jianqi Zhang, Alexander Gusev, Taylor Young, Arti Tandon, Samuela Pollack, Bjarni J Vilhjálmsson, et al. Leveraging population admixture to characterize the heritability of complex traits. *Nature genetics*, 46(12):1356–1362, 2014.

[175] Jian Zeng, Ronald Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, page 1, 2018.

[176] Hou-Feng Zheng, Vincenzo Forgetta, Yi-Hsiang Hsu, Karol Estrada, Alberto Rosello-Diez, Paul J Leo, Chitra L Dahia, Kyung Hyun Park-Min, Jonathan H Tobias, Charles Kooperberg, et al. Whole-genome sequencing identifies en1 as a determinant of bone density and fracture. *Nature*, 526(7571):112–117, 2015.

[177] Hou-Feng Zheng, Jon H Tobias, Emma Duncan, David M Evans, Joel Eriksson, Lavinia Paternoster, Laura M Yerges-Armstrong, Terho Lehtimäki, Ulrica Bergström, Mika Kähönen, et al. Wnt16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *Plos Genetics*, 2012.

[178] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.