Title

Methods for the Quantitative Characterization of the Genetic Basis of Human Complex Traits

Author

Burch, Kathryn

Publication Date

2021

UNIVERSITY OF CALIFORNIA

Los Angeles

Methods for the Quantitative Characterization
# of the Genetic Basis of Human Complex Traits

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Kathryn Sakura Burch

2021

ABSTRACT OF THE DISSERTATION

Methods for the Quantitative Characterization of the Genetic Basis of Human Complex Traits

by

Kathryn Sakura Burch

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2021

Professor Bogdan Pasaniuc, Chair

A major finding from the last decade of genome-wide association studies (GWAS) is that variant-phenotype associations are significantly enriched in noncoding regulatory regions of the genome. This result suggests that GWAS associations localize variants that modulate phenotype via gene regulation as opposed to alterations in protein structure/function. However, for most complex traits, most aspects of genetic architecture—the number of causal variants/genes for a trait and the degree to which causal effect sizes are coupled with genomic features such as minor allele frequency (MAF) and linkage disequilibrium (LD)—remain actively debated. In this dissertation, I introduce three new methods to explore and quantitatively characterize complex-trait genetic architecture. First, I derive an unbiased estimator of genome-wide SNP-heritability under a very general random effects model that makes minimal assumptions on the underlying (unknown) genetic architecture of the trait. Second, I introduce a method for estimating the number of causal variants that are shared between two ancestral populations for a given trait, and I discuss the implications of the method and real-data results for improving polygenic risk prediction in ethnic minority populations. Third, I propose methods for partitioning the heritability of individual genes by MAF to identify disease-relevant genes, with the hypothesis that some disease-relevant genes may have relatively large heritability contributions from rare and low-frequency variants while still having low total gene-level heritability.

The dissertation of Kathryn Sakura Burch is approved.

Jason Ernst

Päivi Pajukanta

Sriram Sankararaman

Janet S Sinsheimer

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2021

To my parents and grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Bogdan Pasaniuc, for challenging me as a student and researcher; for his patience and guidance; and for fostering a supportive, collaborative working environment both within the lab and at UCLA. Working with Bogdan taught me not only how to be scientifically rigorous, but also how to define interesting open questions and how to recognize areas in which I can make unique, impactful contributions to the field.

Second, I am grateful to all current and former members of Bogdan's research group. I would like to thank Rob Brown, Nick Mancuso, Huwenbo Shi, and Gleb Kichaev for the many hours they spent answering my questions about linear models, mathematical statistics, GWAS, fine-mapping, heritability… Their mentorship was invaluable, especially during the first half of my PhD. I owe special thanks to Kangcheng Hou, Yi Ding, and Ruth Johnson for the many productive and stimulating discussions we had while working together on various projects. I am grateful to Malika Freund, Claudia Giambartolomei, Megan Roytman, Tommer Schwarz, Arun Majumdar, Igor Mandric, Robert Smith, Megan Major, Valerie Arboleda, Arjun Bhattacharya, Rachel Mester, Jonatan Hervoso, Ella Petter, and Vidhya Venkateswaran for the helpful discussions during lab meetings and journal clubs and for the hilarious, spontaneous conversations we had by the whiteboards and over Slack. I would also like to give special thanks to two honorary members of Bogdan's lab, Steven Gazal and Harold Pimentel, with whom I have spent countless hours talking about everything, science and otherwise. I am so grateful to all of my labmates for the camaraderie and the uncountable interesting discussions, fun, and laughter. It has been a privilege to work with such a talented and kind group of people.

I would like to thank my committee, Jason Ernst, Päivi Pajukanta, Sriram Sankararaman, and Janet Sinsheimer. They have been generous with their time and have provided thoughtful, insightful feedback on my work. I would also like to thank Eleazar Eskin and Van Savage; if they had not encouraged me to apply to PhD programs, I probably would not be writing this today.

The computational genetics community at UCLA has been tremendously helpful and a huge source of support throughout graduate school. I want to give special thanks to Sriram Sankararaman, Ali Pazokitoroudi, Ariel Wu, Alec Chiu, Arun Durvasala, and Boyang Fu; I learned so much from our collaborations and truly enjoyed the hours we spent brainstorming ideas and discussing each other's projects. I am very grateful to Jazlyn Mooney, David Pan, Kristina Garske, Jennifer Zou, Lisa Gai, Marcus Alvarez, Arthur Ko, James Boocock, Nathan LaPierre, Chris Robles, Leah Briscoe, Ashley Cass, and all of those who attended CGSI and the weekly research seminars. While working from home during the pandemic, one of the things I have missed most is running into friends and colleagues in the hallways, at lunch, and at various events and seminars.

I am so grateful to my friends and family, who encouraged me through the most challenging parts of graduate school and helped me recognize/celebrate my accomplishments, both big and small: Dustin Roberts, Kiku Koyano, Tev Dincer, Grace Hancock, Katrina Hill, Nivedita Kar, Yazeed Kesbeh, Bridgette Hassett, Andreya Van Buren, Andrea Limones, Dorit Stein, Peter Debbaneh, Aya Okada, Rachel Fukuyama, Kamil Khan, Eric Viquez, Tony Carbayo, Anjol Mendoza, Stevie and Angie Gassner, Irine Tyutereva, and many, many others. Finally, I am so incredibly thankful for my parents; they selflessly provided me with everything I needed to graduate (twice!) from UCLA. I owe my successes and accomplishments to them.

VITA

2010-2014	B.S., Computational and Systems Biology University of California, Los Angeles | Los Angeles, CA
2013-2014	Undergraduate Student Researcher, Department of Biomathematics University of California, Los Angeles | Los Angeles, CA
2016-2021	Graduate Student Researcher, Department of Computational Medicine University of California, Los Angeles | Los Angeles, CA
Winter 2019	Teaching Assistant, Life Sciences 107 – Genetics University of California, Los Angeles | Los Angeles, CA

PUBLICATIONS

* Denotes equal contributions or joint supervision

1. **Burch KS***, Hou K*, Ding Y, Wang Y, Gazal S, Shi H, Pasaniuc B. Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. In Review. Preprint: https://doi.org/10.1101/2021.08.17.456722

2. Hou K, Bhattacharya A, Mester R, **Burch KS**, Pasaniuc B. On powerful GWAS in admixed populations. *Nature Genetics*, In Press (2021).

3. Ding Y*, Hou K*, **Burch KS,** Lapinska S, Privé F, Vilhjálmsson B, Sankararaman S, Pasaniuc B. Large uncertainty in individual PRS estimation impacts PRS-based risk stratification. *Nature Genetics,* In Press (2021). Preprint: https://doi.org/10.1101/2020.11.30.403188

4. Pazokitoroudi A, Chiu AM, **Burch KS**, Pasaniuc B, Sankararaman S. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *The American Journal of Human Genetics* (2021). https://doi.org/10.1016/j.ajhg.2021.03.018

5. Majumdar A, **Burch KS**, Haldar T, Sankararaman S, Pasaniuc B, Gauderman WJ, Witte JS. A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. *Bioinformatics* (2021). https://doi.org/10.1093/bioinformatics/btaa1083

6. Shi H*, **Burch KS***, Johnson R, Freund MK, Kichaev G, Mancuso N, Manuel AM, Dong N, Pasaniuc B. Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *The American Journal of Human Genetics* (2020). https://doi.org/10.1016/j.ajhg.2020.04.012

7. Pazokitoroudi A, Wu Y, **Burch KS**, Hou K, Zhou A, Pasaniuc B, Sankararaman S. Efficient variance components analysis across millions of genomes. *Nature Communications* (2020). https://doi.org/10.1038/s41467-020-17576-9

8. Johnson R, **Burch KS**, Hou K, Paciuc M, Pasaniuc B, Sankararaman S. A Scalable Method for Estimating the Regional Polygenicity of Complex Traits. *RECOMB* (2020). https://doi.org/10.1007/978-3-030-45257-5_26

9. Hou K*, **Burch KS***, Majumdar A, Shi H, Mancuso N, Wu Y, Sankararaman S, and Pasaniuc B. Accurate estimation of SNP-heritability irrespective of genetic architecture. *Nature Genetics* (2019). https://doi.org/10.1038/s41588-019-0465-0

10. Major M, Freund MK, **Burch KS**, Mancuso N, Ng M, Furniss D, Pasaniuc B*, and Ophoff RA*. Integrative analysis of Dupuytren's disease identifies novel risk locus and reveals a shared genetic etiology with BMI. *Genetic Epidemiology* (2019). https://doi.org/10.1002/gepi.22209

11. Kichaev G, Bhatia G, Loh PR, Gazal S, **Burch KS**, Freund MK, Schoech A, Pasaniuc B* and Price AL*. Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics* (2019). https://doi.org/10.1016/j.ajhg.2018.11.008

12. Freund MK, **Burch KS**, Shi H, Mancuso N, Kichaev G, Garske KM, Pan DZ, Miao Z, Mohlke KL, Laasko M, Pajukanta P, Pasaniuc B*, and Arboleda VA*. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *The American Journal of Human Genetics* (2018). https://doi.org/10.1016/j.ajhg.2018.08.017

1 Introduction

Complex traits are phenotypes that are influenced by multiple genetic and environmental factors. In humans, examples of complex traits include quantitative traits such as height and cholesterol levels and many common diseases such as cancer, cardiovascular disease, and neuropsychiatric disorders. In contrast to monogenic traits, which are typically driven by rare genetic variants in a single gene, complex traits tend to be *polygenic*—that is, regulated by many genes that each have small individual effects on phenotype. The ability to collect genetic data and quantify various genetic factors contributing to complex traits is critical for many applications, including identification of potential therapeutic targets[1]; polygenic risk prediction for early disease detection or assessment of drug safety/efficacy[2,3]; and better understanding natural selection and human demographic history[4,5].

The sheer size of the human genome (~3 billion base pairs and ~17K protein-coding genes) and the complexity of the biology underlying complex traits create significant obstacles to identifying specific causal genes, pathways, and mechanisms. In the last twenty years, however, massive reductions in the costs of genotyping and sequencing technologies have enabled the genome-wide association study (GWAS), a powerful, cost-effective way to screen the genome for alleles that are associated (correlated) with a trait of interest[1,6]. Performing a GWAS essentially involves collecting genetic and phenotypic measurements in tens or hundreds of thousands of individuals and then testing for associations between each genetic variant in the genome (typically in the millions) and the phenotype or disease risk.

The statistical power of GWAS comes from its cost-effective design based on genotyping arrays, which enable collection of genetic data at a large scale within a reasonable budget. Genotyping arrays leverage a phenomenon called linkage disequilibrium (LD)—population-level correlations between alleles at different sites—to reduce the number of genetic variants that need to be directly measured while still capturing most of the common genetic variation in a population. Given genotype array data, one can impute the genotypes at other variants using estimates of LD obtained from a *reference panel*, which are typically whole genomes measured via whole-genome sequencing (WGS) in a set of individuals sampled from the population[7–10]. Thus, with the advent of genotyping arrays, the availability of reference genomes, and the establishment of large-scale biobanks in several countries[11–17], the number of published GWASs has grown exponentially in the last decade[18–21], and the largest meta-analyses have sample sizes of well over 100,000 individuals[11,22–25].

A major finding from the last decade is that, while the vast majority of GWAS associations lie in noncoding regions of the genome, GWAS signal is significantly enriched in regulatory regions[1,20,26]. This result, which has been replicated across a wide range of traits, suggests that GWAS associations localize variants that modulate phenotype via gene regulation as opposed to alterations in protein structure/function. However, the same LD that helps to reduce the costs of large-scale genetic studies also creates significant computational and statistical challenges in the analysis and interpretation of GWAS results. A genetic variant that is significantly associated with a trait is not necessarily causal as it can be "tagging" the effect of a nearby causal variant via LD. Identifying the causal variant in a region identified by GWAS ("GWAS risk region") is a nontrivial problem: if two variants are in "high LD" with one another—that is, the genotypes at the two loci are highly correlated in the population—it can be impossible to definitely elucidate the true causal variant without additional information[27–29].

Due to varying selective pressures on different traits in different populations, the *genetic architecture* of a complex trait—which is broadly defined here to mean the number of causal variants/genes for a trait and the degree to which causal effect sizes are coupled with genomic features such as allele frequency and local LD—can vary significantly across populations[30]. Many quantitative genetic models rely on strong assumptions about genetic architecture to estimate critical parameters such as *heritability*, which is the proportion of phenotypic variance in the population explained by additive genetic variation[31]. In particular, in a linear model relating a given set of variants to phenotype, *SNP-heritability*, the heritability explained by the set of variants in the model, is the theoretical upper bound on polygenic risk prediction accuracy that is attainable from this model[32]. Several recent works have demonstrated that methods for estimating heritability make various assumptions on genetic architecture that can yield different estimates, even when applied to the same GWAS data[33–36]. This discrepancy has been the source of much recent debate in the literature, particularly as it applies to estimating *enrichments* of heritability in certain genomic regions of interest[34,36].

In this dissertation, I introduce new methods to explore and quantitatively characterize the genetic architecture of complex traits. Motivated by recent debate on the topic[34–36], in **Chapter 2**, I investigate whether it is possible to obtain unbiased estimates of SNP-heritability without making assumptions on the underlying (unknown) genetic architecture. I derive an unbiased estimator of genome-wide SNP-heritability under a random effects model that is a generalization of the random effects models assumed by other state-of-the-art-methods. I refer to this model as a "generalized random effects" model; the resulting estimator is referred to as the "GRE estimator" to emphasize that it was derived under a "GRE model."

Critically, the GRE estimator depends on having individual-level GWAS data at sample sizes larger than the number of genotyped SNPs on an array. With the availability of individual-level genotype and phenotype data in >300K "white British" individuals the UK Biobank[11], we were able to implement the GRE estimator and compare it to a range of existing methods, each of which makes assumptions that can be subsumed under the GRE model. This work is published in *Nature Genetics*[37]. My contributions to this work were the mathematical derivations; the design of simulation experiments; the design and execution of analyses in real data; the interpretation of results, including statistical analyses; and writing the paper.

As a result of diverging human migration histories and geodemographic events taking place over thousands of years, allele frequencies and LD patterns vary across global populations[38–41]. For many complex traits, a substantial number of GWAS associations have been replicated in multiple ancestries, suggesting that at least some amount of genetic risk is shared between ancestral populations[42,43]. However, polygenic risk scores (PRSs) have repeatedly been shown to perform poorly if applied in individuals whose ancestries differ from that of the GWAS participants used to construct the PRS[44–48]. Taken together with estimates of transethnic correlations < 1 reported in the literature for many complex traits[49,50], the population-specificity of risk prediction models suggests that causal variants and their effect sizes ("causal effect sizes") may differ between populations[43].

In **Chapter 3**, motivated by the open question of whether disease risk is modulated by the same variants in different ancestral groups, I introduce a method for estimating the numbers of causal variants that are unique to versus shared between two ancestral populations from GWAS summary statistics. The method can be applied both genome-wide and within any genomic annotation of interest (e.g., a set of genes). We applied the method to 9 complex traits and diseases for which summary-level GWAS data were available in European and East Asian ancestries and found that, on average across traits, ~80% of common causal variants (minor allele frequency > 5% in each population separately) are shared between individuals of European and East Asian ancestry. This work, including a detailed discussion on a number of important caveats, is published in *The American Journal of Human Genetics*[51]. My contributions were the statistical analysis of simulation results; the design and interpretation of analyses in real data; and writing the manuscript.

One of the caveats discussed at the end of Chapter 3 is that the method may be biased towards identifying common genetic variants. Since common variants are the dominant contributors to SNP-heritability[52], such a bias would be acceptable for certain downstream applications—for example, improving transethnic portability of PRSs. For the purpose of understanding disease etiology, however, whether SNP-heritability enrichments—which are dominated by common-variant heritability—can be used to identify the most critical genes for a trait is unclear[53,54]. Specifically, individual rare variants with large per-allele effects contribute very little to population-level phenotypic variance likely because selection acts on high-effect alleles, thus keeping them at low frequencies in the population. A critical implication of this is that the most important genes for a trait may not be in GWAS risk regions or regions enriched with common-variant heritability[54,55].

In **Chapter 4**, I introduce a quantity called gene-level heritability, defined as the proportion of phenotypic variance explained by the additive effects of a given set of variants assigned to a gene of interest. I propose an approach for partitioning gene-level heritability by allele-frequency classes with the goal of finding genes whose total gene-level heritability is explained exclusively by rare causal variants (0.5% ≤ MAF < 1%). We analyze ~17K protein-coding genes and 25 quantitative traits in the UK Biobank (N=290K) and find that among genes with nonzero heritability, only ~0.8% (on average across traits) have heritability explained exclusively by rare variants. Total and rare-variant gene-level heritability exhibit starkly different trends that, taken together, provide a more comprehensive picture of complex-trait genetic architecture. Our findings are consistent with the hypotheses that (i) selection "flattens" heritability to be more evenly distributed among common variants[54] and (ii) complex traits may be modulated in part by dysregulation of genes that—if completely disrupted—cause phenotypically similar Mendelian disorders[56]. This work is available as a preprint[57] and is currently undergoing peer-review.

2 Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture

2.1 Introduction

SNP-heritability, the proportion of phenotypic variance attributable to the additive effects of a given set of SNPs, is a fundamental quantity in genetics[31]; it provides an upper bound on risk prediction from a linear model[58] and, when defined as a function of all SNPs on an array, yields insights into the "missing heritability" of complex traits[1,6,52]. Traditionally, SNP-heritability is estimated by fitting variance components models with REML[35,52,59–61]. With some exceptions[61], REML-based methods are not scalable to biobanks that assay hundreds of thousands of individuals (e.g., UK Biobank[11]). SNP-heritability can also be estimated by assessing the deviation in marginal association statistics as a function of LD scores[34,62–64]; such methods can scale to millions of individuals. More recently, a randomized extension of Haseman-Elston regression[65] was shown to estimate a single genetic variance component from individual-level data as accurately as REML methods but in a fraction of the run-time[66].

To facilitate inference, all existing methods for genome-wide SNP-heritability inference make assumptions on genetic architecture, which is typically parametrized by *polygenicity* (the number of variants with effects larger than some small constant δ) and *MAF/LD-dependence* (the

This chapter is published in Hou*, Burch*, et al. *Nature Genetics* (2019). Citations have been updated and renumbered. Supplementary material is freely available at www.ncbi.nlm.nih.gov/pmc/articles/PMC6686906 coupling of effects with minor allele frequency (MAF), local linkage disequilibrium (LD), or other functional annotations)[30]. Since the true genetic architecture of any given trait is unknown, existing methods are susceptible to bias and often yield vastly different estimates even when applied to the same data[34–36,67]. Although multi-component methods that stratify SNPs by MAF/LD ameliorate some of these robustness issues, fitting multiple variance components to biobank-scale data with REML is highly resource-intensive[61] and it is unclear whether multi-component methods based on summary statistics produce accurate estimates of total SNP-heritability. Alternate methods that explicitly model MAF/LD-dependency[34,35,59] are also sensitive to model misspecification[36,67]. In addition, genetic architecture varies across traits and populations due to, for example, variable degrees of negative selection acting on different traits in different populations[44,68–70]. Methods that jointly infer SNP-heritability and parameters such as the strength of negative selection or polygenicity[71,72] are computationally intensive and/or sensitive to LD-dependency. Thus, it remains unclear which estimates of SNP-heritability computed from biobank-scale data are reliable.

In this work, we investigate whether genome-wide SNP-heritability can be accurately estimated under a generalized random effects (GRE) model that makes minimal assumptions on genetic architecture. Under this model, every causal effect has an arbitrary SNP-specific variance, and SNP-heritability is defined as the sum of the SNP-specific variances (Methods). To the best of our knowledge, all existing methods make additional assumptions on top of the GRE model (Table 2.1). For example, GREML[52] (and several other methods[61,62,66]) imposes an inverse relationship between MAF and allelic effect size whereas LDAK assumes that each SNP-specific variance is inversely proportional to both MAF and LD tagging[34,35,59]. We derive a closed-form estimator for SNP-heritability as a function of marginal association statistics and in-sample LD and show that this estimator is consistent (approaches the true SNP-heritability as sample size increases) and unbiased (its expectation is equal to the true SNP-heritability) when the number of individuals exceeds the number of SNPs. Most importantly, the accuracy of this estimator is invariant to genetic architecture. While the GRE estimator is similar in form to previously proposed fixed-effect estimators[73,74], our approach differs from previous work in two main ways. First, SNP-heritability defined under a fixed effect model is different from the estimand of interest here (Ch. 2.4.1 Methods). Second, previous work applied the estimator locally to identify regions contributing disproportionately to the genome-wide signal[73,74]; here we define a different genome-wide estimator (Equation 2.1) that requires large-scale genotype data. In addition, previous work applied an SVD-based regularization to account for errors in LD estimation from reference panels[74], which was unnecessary in this work (Ch. 2.4.2 Methods).

Through extensive simulations across a range of MAF/LD-dependent architectures starting from real genotypes from the UK Biobank[11] (337K individuals, 593K SNPs), we find that the GRE estimator is nearly unbiased across all architectures whereas existing methods are sensitive to model misspecification. For example, across 126 distinct architectures, the maximum bias of the GRE estimator is 2% of the simulated SNP-heritability whereas stratified LD score regression (S-LDSC)[63,64] and SumHer[34] yield biases between -64% and 28%. For completeness, we also contrast the GRE estimator with several REML-based methods in simulations at lower sample sizes (due to the computational burden of most REML methods) and find that, consistent with recent reports[67], all REML-based methods are biased when their model assumptions are violated, and multi-component REML methods that stratify SNPs by MAF and LD score (GREML-LDMS-I[67]) are more accurate than single-component REML methods. The performance of the GRE estimator is similar to that of GREML-LDMS-I, confirming that SNP-heritability can be accurately estimated without stratifying SNPs or specifying a heritability model.

Finally, we use marginal association statistics and in-sample LD from 290K “unrelated white British” individuals and 460K SNPs (MAF > 1%) to estimate SNP-heritability for 22 complex traits in the UK Biobank[11]. Consistent with simulations, estimates from S-LDSC and SumHer differ from the GRE estimates by a median of -9% and 11%, respectively, across the 18 traits with SNP-heritability estimates exceeding 0.05. For example, for height, estimates from S-LDSC (0.56) and SumHer (0.63) are approximately 7% lower and 5% higher, respectively, than our estimate of 0.60. Similarly, for hypertension, estimates from S-LDSC (0.14) and SumHer (0.18) are ±12.5% different from our estimate of 0.16. Taken together, our results demonstrate that SNP-heritability can be accurately estimated from biobank-scale data without prior knowledge of the genetic architecture the trait, motivating the development of scalable methods with fewer modeling assumptions.

2.2 Results

2.2.1 Overview of the approach

We investigate the utility of an estimator derived under a model that makes minimal assumptions on genetic architecture. We model the standardized phenotype of an individual as $y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$, where $\mathbf{x}$ is an M-vector of standardized genotypes, $\boldsymbol{\beta}$ is the corresponding vector of standardized effects, and $\epsilon \sim N(0, \sigma_e^2)$ is environmental noise (Ch. 2.4.1 Methods). The effect size of each SNP is assumed to have mean zero and a finite SNP-specific variance (σ_i^2 for SNP i) that is allowed to be 0; the covariance between all pairs of effects is assumed to be zero. We term this model the "generalized random effects" (GRE) model as, to the best of our knowledge, all existing methods impose additional assumptions on top of this model. For example, the single-component GREML model[52] assumes $\sigma_i^2 = h_g^2/M$ for $i = 1, \dots, M$, whereas the most recent LDAK model[35] assumes $\sigma_i^2 \propto w_i[f_i(1-f_i)]^{0.75}$ (where w_i is a SNP-specific LD weight and f_i is MAF) (Table 2.1). Under the GRE model, the SNP-heritability explained by the M SNPs is the sum of the SNP-specific variances: $h_g^2 \equiv \mathrm{Var}[\mathbf{x}^T\boldsymbol{\beta}]/\mathrm{Var}[y] \ = \sum_{i=1}^M \sigma_i^2$ (Ch. 2.4.1 Methods).

Given genotype measurements across N individuals at M SNPs and assuming $N > M$, the estimator $\hat{h}_g^2 = \frac{N\widehat{\boldsymbol{\beta}}^T\widehat{\mathbf{V}}^\dagger\widehat{\boldsymbol{\beta}}-q}{N-q}$, where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated marginal effects, $\widehat{\mathbf{V}}^\dagger$ is the pseudoinverse of the in-sample LD matrix, and q is the rank of the in-sample LD, is an unbiased estimator of SNP-heritability under the GRE model. That is, $\mathrm{E}\left[\hat{h}_g^2\right] = \sum_{i=1}^M \sigma_i^2 = h_g^2$ (Ch. 2.4.2 Methods). Unfortunately, even the largest biobanks currently have $N < M$ (i.e. UK Biobank has genotyped $M \approx 593\mathrm{K}$ SNPs in $N \approx 337\mathrm{K}$ unrelated British individuals), which limits the utility of the above estimator. We therefore extend our approach by partitioning the genome by chromosome:

$$\hat{h}_{\mathrm{GRE}}^2 = \sum_{k=1}^{22} \frac{N\widehat{\boldsymbol{\beta}}_k^T\widehat{\mathbf{V}}_k^\dagger\widehat{\boldsymbol{\beta}}_k - q_k}{N - q_k} \tag{2.1}$$

where for chromosome k with p_k SNPs, $\widehat{\boldsymbol{\beta}}_k$ is the p_k-vector of estimated effects, $\widehat{\mathbf{V}}_k^\dagger$ is the pseudoinverse of the in-sample LD matrix, and q_k is the rank of the in-sample LD. Although this estimator introduces bias, we show through extensive simulations that the magnitude of the bias is extremely small when N is sufficiently larger than p_k.

2.2.2 The GRE estimator is robust with respect to genetic-architecture parameters

To investigate the bias and variance of $\hat{h}^2_{\mathrm{GRE}}$, we perform simulations starting from real genotypes ($N = 337{,}205$, UK Biobank). First, we simulate 64 MAF/LD-dependent quantitative trait architectures from chromosome 22 ($M = 9654$ typed SNPs) by varying the SNP-heritability (h^2_g), proportion of causal variants (p_{causal}), distribution of causal variant MAF (CV MAF), and strength of coupling between effect size and MAF/LD; we use "LDAK-LD-dependent" to describe causal effects that are coupled with "LDAK weights" (Ch. 2.4.4 Methods). To compare estimates across different values of h^2_g, we assess bias as a percentage of the simulated value of h^2_g (relative bias). Errors of individual estimates are also expressed as percentages of h^2_g. Consistent with analytical derivations, the GRE estimator restricted to chromosome 22 is unbiased across the 64 architectures (bias p-value < 0.05/16 is considered significant in order to correct for 16 tests (architectures) at each value of h^2_g; Ch. 2.4.5 Methods) (Figure 2.1a, Figure 2.1c, Supplementary Table 1). The average relative bias across the 64 architectures is $0.00015\% \times h^2_g$ and the largest bias under any single architecture is approximately $\pm 0.2\% \times h^2_g$ (Supplementary Figure 1a, Supplementary Table 1 or ref.[37]). In simulations of unascertained case-control studies (Ch. 2.4.4 Methods), the GRE estimator is approximately unbiased across a range of disease prevalences (for $h^2_g = 0.10$, relative bias range is [-0.20%, 0.30%]) and has larger variance for lower prevalences (Supplementary Figure 2a and Supplementary Table 2). For ascertained case-control studies, estimates are downward-biased but invariant to architecture (when $h^2_g = 0.10$, prevalence = 0.10, and $N_{\mathrm{case}} = N_{\mathrm{control}}$, relative bias is approximately -4%) (Supplementary Table 3). Masking 0%, 50%, or 100% of causal SNPs from the observed summary statistics induces downward-bias when CV MAF = [0.01, 0.05] due to lower average LD between the observed SNPs and masked causal SNPs

(Supplementary Figure 3). The analytical estimator of the standard error (Ch. 2.4.3 Methods) is well-calibrated (Supplementary Figure 4a, Supplementary Table 4). As expected, partitioning chromosome 22 into disjoint, non-independent blocks induces upward bias that increases as block size decreases (Supplementary Figure 5, Supplementary Table 5).

Next, we perform genome-wide simulations ($N = 337\text{K}$ individuals, $M = 593\text{K}$ SNPs) to assess $\hat{h}^2_{\text{GRE}}$ with the 22-block approximation (Equation 2.1). Despite the approximation, $\hat{h}^2_{\text{GRE}}$ is highly accurate and robust across all 64 MAF- and LDAK-LD-dependent quantitative trait architectures (Figure 2.1b, 2.1c). Across the 64 architectures, the bias ranges from 0.07% to 2.1%× h^2_g (average = 0.97% × h^2_g) (Supplementary Figure 1b, Supplementary Table 6). Across all 6400 simulations (64 genetic architectures × 100 simulation replicates), the largest error of any single estimate is approximately $17\% \times h^2_g$ (Figure 2.1c). As N/M increases, the variance of $\hat{h}^2_{\text{GRE}}$ decreases while the relative bias appears to be approximately fixed, ranging between 0.91% ($N = 100\text{K}$) and 0.99% ($N = 200\text{K}$) (Figure 2.1d). These trends hold for a range of p_{causal} (Supplementary Figure 6, Supplementary Table 6), for unascertained case-control studies (Supplementary Figure 2b, Supplementary Table 7), and in a smaller set of simulations with $N = 7685$ individuals of South Asian ancestry and $M = 1642$ SNPs (Supplementary Table 8; Ch. 2.4.5 Methods). Most importantly, the accuracy of the GRE estimator is invariant to the underlying architecture (Figure 2.1b). The analytical estimator for the standard error is downward-biased (and invariant to genetic architecture) with respect to the empirical standard deviation of $\hat{h}^2_{\text{GRE}}$ estimates (Supplementary Figure 4b, Supplementary Table 9). For example, across 16 architectures where $h^2_g = 0.25$, the empirical standard deviation of 100 independent estimates ranges from 0.0049 to 0.0064, whereas our estimated standard errors are approximately 0.0036 across all architectures

(Supplementary Figure 4b, Supplementary Table 9).

We investigate the effects of unmodeled substructure and/or cryptic relatedness by filtering individuals at different kinship coefficient thresholds (Ch. 2.4.4 Methods) and find that using stricter relatedness thresholds increases the variance of the estimates (due to smaller sample size) while reducing bias, albeit not significantly (Supplementary Figure 7, Supplementary Table 10). To assess the impact of population stratification, we simulated an effect of the first genetic principal component (PC) on phenotype and computed OLS association statistics both with and without adjusting for the first PC (Ch. 2.4.4 Methods). As expected, OLS without PC adjustment yields inflated estimates while OLS with PC adjustment yields approximately unbiased estimates (Supplementary Figure 8, Supplementary Table 11). However, even when a relatively large proportion of phenotypic variance is explained by the first PC (e.g., $h_g^2 = 0.25$, $\sigma_s^2 = 0.05$), the maximum bias we observe using unadjusted association statistics is $5\% \times h_g^2$ (bias p-value = 2.7×10^{-9}). Together, these results indicate that the GRE estimator is robust to modest amounts of unmodeled substructure and/or stratification. In all subsequent analyses, we compute $\hat{h}_{\mathrm{GRE}}^2$ with the 22-block approximation as this provides sufficiently accurate estimates and a fair comparison to other methods.

2.2.3 Comparison of methods to estimate SNP-heritability

We compare $\hat{h}_{\mathrm{GRE}}^2$ with existing state-of-the-art methods that are easily scalable to the full UK Biobank data ($N = 337$K): LD score regression (LDSC), which assumes $\alpha = -1$ and no coupling of effects with LD[62]; stratified LD score regression (S-LDSC), which partitions h_g^2 by a set of annotations of interest[63,64]; and SumHer, a scalable extension of LDAK which explicitly models

MAF/LD-dependency through a specific form of the SNP-specific variances[34] (Table 2.1). To ensure a fair comparison, LD scores for all methods are computed using in-sample LD among the *M* SNPs, and in all simulations we aim to estimate the SNP-heritability explained by the same *M* SNPs (Ch. 2.4.5 Methods).

As expected, $\hat{h}^2_{\text{GRE}}$ is robust across all architectures while LDSC, S-LDSC, and SumHer are sensitive to model misspecification. For example, when $h_g^2 = 0.25$ (Figure 2.2), LDSC is approximately unbiased under the "single-component GREML model" (relative bias $= 0.04\%$, $p = 0.86$) but is sensitive to CV MAF and the degree of coupling between effect size and MAF/LD (e.g., when $p_{\text{causal}} = 1\%$, relative bias ranges from -44% to 50%) (Supplementary Table 12). Similarly, SumHer is accurate under the "LDAK model" (relative bias $= 5.3\%$) but highly sensitive to other architectures (when $p_{\text{causal}} = 1\%$, relative bias ranges from -19% to 22%) (Figure 2.2, Supplementary Table 13). S-LDSC (MAF), which partitions h_g^2 by 10 MAF bins (Supplementary Table 14; Ch. 2.4.5 Methods), is less biased than LDSC when effects are coupled with only MAF, but is significantly downward-biased when effects are also coupled with LDAK weights (for $h_g^2 = 0.25$, relative bias range is [1.9%, 7.0%] when $\gamma = 0$ and [-58%, -37%] when $\gamma = 1$) (Figure 2.2, Supplementary Table 15). S-LDSC with 10 MAF bins and an additional "level of LD" annotation, denoted S-LDSC (MAF+LLD) (Methods), produces similar results (for $h_g^2 = 0.25$, relative bias range is [1.8%, 6.5%] when $\gamma = 0$ and [-80%, -33%] when $\gamma = 1$) (Supplementary Table 16). In contrast, the relative bias of $\hat{h}^2_{\text{GRE}}$ ranges from 0.45% to 1.3% across the same 16 architectures where $h_g^2 = 0.25$ and $p_{\text{causal}} = 1\%$ (Figure 2.2, Supplementary Table 6). These trends hold for a range of h_g^2 and p_{causal}: across 112 LDAK-LD- and/or MAF-dependent architectures, the average and range of the relative bias of each method are 0.96% [-0.06%, 2.1%]

(GRE), -2.2% [-71%, 70%] (LDSC), -22% [-62%, 8.7%] (S-LDSC (MAF)), -29% [-89%, 9.0%] (S-LDSC (MAF+LLD)), and 2.8% [-27%, 28%] (SumHer) (Figure 2.1b, Figure 2.2, Supplementary Figures 9-12 and Supplementary Tables 6,12,13,15,16). Across 14 alternative LD-dependent architectures where SNP-specific variances are coupled with inverse LD scores instead of LDAK weights ("LD-score-dependent" architectures; Ch. 2.4.4 Methods, Supplementary Figure 13), $\hat{h}^2_{\text{GRE}}$ remains nearly unbiased (relative bias range [0.52%, 1.3%]) whereas S-LDSC (MAF), S-LDSC (MAF+LLD), and SumHer are generally downward-biased (Supplementary Figure 14, Supplementary Table 17).

For completeness, we compare to four widely used REML-based methods: GREML, which assumes $\alpha = -1$ and no coupling of effects with LD; GREML-LDMS-I, a multi-component extension of GREML that partitions SNPs by MAF and LD score; BOLT-REML, a computationally efficient variance components estimation method with assumptions similar to those of GREML; and LDAK, which assumes a specific form of the SNP-specific LD weights and recommends setting $\alpha = -0.25$ (Table 2.1). Because it is computationally intractable to apply the REML-based methods to thousands of genome-wide simulations with 337K individuals, we perform simulations using a reduced number of individuals ($N = 8430$) and SNPs ($M = 14821$) (Ch. 2.4.5 Methods). As expected, the single-component methods (GREML, BOLT-REML, and LDAK) are sensitive to MAF/LD-dependency whereas the GRE estimator is robust across all architectures. For example, when $h_g^2 = 0.25$ (Figure 2.3), GREML and BOLT-REML are accurate under the GREML model (GREML: relative bias $= -1.4\%$, $p = 6.0 \times 10^{-3}$, Supplementary Table 18; BOLT-REML: relative bias $= -0.16\%$, $p = 0.75$, Supplementary Table 19) and LDAK is approximately unbiased under the LDAK model (relative bias $= 0.16\%$, $p = 0.77$, Supplementary Table 20), but all three are sensitive to CV MAF, α, and γ. Across 12 architectures where $p_{\text{causal}} = 1\%$ (Figure 2.3), the relative biases are within [-15%, 7.9%] (GREML), [-14%, 9.1%] (BOLT-REML), and [-34%, 8.2%] (LDAK) (Supplementary Tables 18-20). In contrast, for the same 12 architectures, $\hat{h}^2_{\text{GRE}}$ yields relative biases in the range [-2.1%, 1.7%], which is comparable to the relative bias of GREML-LDMS-I (range [-2.9%, 1.5%]) using 8 GRMs (4 LD quartiles × 2 MAF bins) that align with CV MAF (Figure 2.3, Supplementary Tables 21, 22). These trends hold over a range of h_g^2 and p_{causal}: across 112 LDAK-LD- and/or MAF-dependent architectures (Supplementary Figures 15-19), the average and range of the relative bias are 0.09% [-4.9%, 6.4%] (GRE), -0.6% [-5.9%, 2.3%] (GREML-LDMS-I), -2.9% [-27%, 15%] (GREML), -1.8% [-25%, 18%] (BOLT-REML), and -8.2% [-44%, 13%] (LDAK) (Supplementary Tables 18-22). Similar trends are observed for LD-score-dependent architectures (Supplementary Figure 20, Supplementary Table 23). In an extreme example where CV MAF is tightly concentrated near 1%, GREML-LDMS-I with the same 8 GRMs as before is downward-biased whereas the GRE estimator remains robust (Supplementary Figure 21, Supplementary Tables 18-22). While the variance of our estimator is larger than the variances of the REML-based methods (Figure 2.3), our approach is designed for sample sizes several orders of magnitude larger than what we used in these simulations. In summary, our results confirm that it is possible to accurately estimate h_g^2 under the GRE model.

2.2.4 SNP-heritability of 22 complex traits in the UK Biobank

Finally, we compute $\hat{h}^2_{\text{GRE}}$ for 22 complex traits in the UK Biobank (290K unrelated British individuals, 460K SNPs; Ch. 2.4.6 Methods). For comparison, we also provide estimates from LDSC, S-LDSC (controlling for the baseline-LD model[63,64]), and SumHer. Of the 22 traits analyzed (6 quantitative, 16 binary), we focus on 18 traits for which $\hat{h}^2_{\text{GRE}} > 0.05$ (Table 2.2). For the 6 quantitative traits, $\hat{h}^2_{\text{GRE}}$ ranges from 0.12 (smoking status) to 0.60 (height). Across the 12 binary traits, $\hat{h}^2_{\text{GRE}}$ ranges from 0.064 (autoimmune disorders) to 0.16 (hypertension) (Table 2.2). These estimates are robust to filtering of individuals based on relatedness (Supplementary Table 24). We also computed $\hat{h}^2_{\text{GRE}}$ from two additional sets of SNPs (MAF > 0.1% and MAF > 0.01%) and found that the estimates increase slightly for lower MAF thresholds (Supplementary Table 25), which is expected due to the increased number of SNPs. To enable a direct comparison between $\hat{h}^2_{\text{GRE}}$ and the quantities estimated by LDSC, S-LDSC, and SumHer, we run the summary-statistics-based methods with LD scores and regression weights computed from in-sample LD and estimate h^2_g defined as a function of the same set of SNPs (Ch. 2.4.6 Methods). Across the 18 traits, S-LDSC (baseline-LD/in-sample) and SumHer (in-sample) differ from $\hat{h}^2_{\text{GRE}}$ by a median of -9% and 11%, respectively (expressed as a percentage of $\hat{h}^2_{\text{GRE}}$) (Figure 2.4, Table 2.2). As expected[62], LDSC (in-sample) yields inflated estimates.

To compare $\hat{h}^2_{\text{GRE}}$ to estimates reported in the literature, we also run the summary-statistics methods with their recommended parameter settings and with LD scores and regression weights computed from the 1000 Genomes Phase 3 reference panel[7] (489 Europeans) – we note that when running these methods as recommended, their estimands are not equivalent to our definition of h^2_g (see Ch. 2.4.6 Methods and refs.[34,36,62,63] for details). Across the 18 traits for which $\hat{h}^2_{\text{GRE}} > 0.05$, the median differences with respect to $\hat{h}^2_{\text{GRE}}$ are -11% for LDSC (1KG), -14% for S-LDSC (baseline-LD/1KG), and 38% for SumHer (1KG) (Supplementary Figure 22, Supplementary Table 26). For 9 of these traits, a previous study reported single-component BOLT-REML estimates (computed from a similar UK Biobank cohort[75]) that differ from our estimates by a median of 8%

(Supplementary Table 26).

2.2.5 Runtime and memory requirement

We report the runtime and memory requirements for computing $\hat{h}^2_{\mathrm{GRE}}$ with the 22-block approximation from 337K individuals and 593K SNPs. First, computing chromosome-wide LD has complexity $O(Np_k^2)$ for chromosome k with p_k SNPs. In practice, this step does not impose a computational bottleneck because the computations can be parallelized over SNPs. Second, the pseudoinverse of each LD matrix is computed via truncated SVD, which has complexity $O(p_k^3)$ for chromosome k. For 50K typed SNPs this takes about 3 hours and 60GB of memory. Lastly, given the pseudoinverse LD matrices and OLS association statistics, computing $\hat{h}^2_{\mathrm{GRE}}$ has complexity $O(p_1^2 + \cdots + p_{22}^2)$. For any of the traits analyzed in this work, this takes less than 1 hour and requires 24GB of memory; most of this time is spent loading the data into memory. For comparison, running LDSC, S-LDSC, or SumHer consists of precomputing LD scores and SNP-specific weights and performing linear regression to estimate the variance parameters. Precomputing LD scores and SNP-specific weights can be parallelized over blocks of SNPs. The second step (least squares regression) is $O(C^2M)$ for M SNPs in the regression and C variance parameters.

2.3 Discussion

In this work, we show that SNP-heritability can be accurately estimated under minimal assumptions on genetic architecture. Our proposed estimator allows the SNP-specific variances to capture arbitrary relationships between effect size and MAF/LD, and we demonstrate through simulations that its accuracy is invariant to genetic architecture. We show that all existing methods impose additional assumptions on the GRE model, and we confirm through simulations that these methods can be sensitive to model misspecification. One practical advantage of our approach over summary-statistics methods is that the estimand of our approach is always the same for a given genotype matrix, whereas the definitions and interpretations of the estimands of LDSC, S-LDSC, and SumHer depend on which SNPs are used in each step of inference (e.g., the SNPs used to compute LD scores need not be the same SNPs defining the estimand)[36,62,63]. Overall, our results show that while existing methods can yield biases, for the purpose of estimating total SNP-heritability, most methods are relatively robust.

We conclude with several caveats and future directions. First, the utility of $\hat{h}^2_{\text{GRE}}$ critically depends on the ratio between the number of SNPs (*M*) and the number of individuals (*N*) – as M/N increases, the eigenstructure of the in-sample LD matrix becomes increasingly distorted (larger eigenvalues are overestimated; smaller eigenvalues are underestimated)[76]. We mitigate this by assuming that chromosomes are approximately independent; as long as *N* exceeds the number of array SNPs per chromosome, $\hat{h}^2_{\text{GRE}}$ provides meaningful estimates of SNP-heritability. While the utility of our approach is limited by the availability of individual-level biobank-scale data, this concern will abate as more biobanks are established[13,14,17]. A major limitation remains with respect to imputed/sequencing data as *M* will continue to be orders of magnitude larger than *N* for the foreseeable future. We defer an investigation of regularized estimation of LD in high-dimensional settings ($M > N$) to future work.

Second, the theoretical guarantees of $\hat{h}^2_{\text{GRE}}$ rely on the assumption that OLS association statistics and LD are estimated from the same genotypes. While summary statistics have been made publicly available for hundreds of large-scale GWAS, in-sample LD is usually unavailable for these studies since most are meta-analyses[77]. In addition, summary statistics are often computed using linear mixed models to control for confounding, and previous works have noted that the LD computation must be adjusted to accommodate mixed model association statistics[77,78]. Thus, the sensitivity of $\hat{h}^2_{\text{GRE}}$ to reference panel LD (with or without regularized LD estimation) and/or mixed model association statistics remains unclear[74,79]. Furthermore, we simulate phenotypes from typed SNPs because imputed genotypes have highly irregular LD patterns[35,67]. Although it would be more realistic to simulate from sequencing data, our simulation design required individual-level genotype measurements in biobank-scale sample sizes.

Third, $\hat{h}^2_{\text{GRE}}$ does not correct for population structure/stratification. In real data, we mitigate this by considering only unrelated individuals (> 3rd degree relatives) and including age, sex, and the top 20 PCs as covariates when computing association statistics. While recent work has found evidence of assortative mating for some traits in the UK Biobank (e.g., height)[80], our estimates are robust to different relatedness thresholds, suggesting that adjusting for the top 20 PCs sufficiently controls for population stratification. Still, it remains unclear how to quantify the bias of our genome-wide estimator due to structure or assortative mating in real data. Future work is needed to extend the GRE approach to control for ascertainment bias[65,66,81,82].

Finally, while previous works applied similar estimators (defined under fixed effects models) to estimate local SNP-heritability within small regions[73,74], additional work is needed to extend our approach to perform partitioning of SNP-heritability by functional annotations. Existing methods for partitioning SNP-heritability make various assumptions on genetic architecture[34,61,63,64,83], motivating the development of new methods in this area.

2.4 Methods

2.4.1 The generalized random effects model

We model the phenotype for an individual n randomly sampled from the population as $y_n = \mathbf{x_n^T}\boldsymbol{\beta} + \epsilon_n$, where $\mathbf{x_n} = (x_{n1} \dots x_{nM})^T$ is a vector of standardized genotypes measured at M SNPs for individual n, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)^T$ is an M-vector of the corresponding standardized SNP effects, and $\epsilon_n \sim N(0, \sigma_e^2)$ is environmental noise. We assume $\mathrm{Var}[y_n] = 1$ and that the genotype at each SNP i is centered and scaled in the population such that $\mathrm{E}[x_{ni}] = 0$ and $\mathrm{Var}[x_{ni}] = 1$; i.e. $x_{ni} = (g_{ni} - 2f_i)/\sqrt{2f_i(1-f_i)}$, where $g_{ni} \in \{0,1,2\}$ is the number of copies of the effect allele at SNP i for individual n, and f_i is the population frequency of the effect allele at SNP i. We define the population LD between two SNPs i and j to be $v_{ij} \equiv \mathrm{E}[x_{ni}x_{nj}]$ for all $i \neq j$. The population LD matrix among the M SNPs is therefore $\mathbf{V} \equiv \mathrm{Cov}[\mathbf{x_n^T}]$. For simplicity, we use "SNP effects" in lieu of "standardized SNP effects" to refer to $\boldsymbol{\beta}$. We assume that $\mathbf{x_n}$ and $\boldsymbol{\beta}$ are independent given allele frequencies $(f_1, \dots, f_M)$ and $\mathbf{V}$.

Under the generalized random effects (GRE) model, the first two moments of β_i are $\mathrm{E}[\beta_i] = 0$ and $\mathrm{Var}[\beta_i] = \sigma_i^2$, where σ_i^2 can be any arbitrary nonnegative finite number. We assume the covariance between the effects of different SNPs is 0 (i.e. $\mathrm{Cov}[\beta_i, \beta_j] = \mathrm{E}[\beta_i\beta_j] = 0$ for all $i \neq j$). Because the SNP-specific variances can capture any degree of polygenicity and any relationship between genomic features (e.g., MAF and LD) and effect size, the GRE model encompasses most realistic genetic architectures (Table 2.1).

We define total SNP-heritability (h_g^2) to be the proportion of phenotypic variance attributable to the additive effects of a set of M SNPs whose genotypes are directly measured:

$$
\begin{aligned}
h_g^2 &\equiv \frac{\mathrm{Var}[\mathbf{x}_n^T\boldsymbol{\beta}]}{\mathrm{Var}[y_n]} \\
&= \mathrm{E}[\mathrm{Var}[\mathbf{x}_n^T\boldsymbol{\beta}|\boldsymbol{\beta}]] + \mathrm{Var}[\mathrm{E}[\mathbf{x}_n^T\boldsymbol{\beta}|\boldsymbol{\beta}]] \\
&= \mathrm{E}[\boldsymbol{\beta}^T\mathrm{Var}[\mathbf{x}_n^T]\boldsymbol{\beta}] + \mathrm{Var}[\mathrm{E}[\mathbf{x}_n^T]\boldsymbol{\beta}] \\
&= \mathrm{E}[\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}] + 0 \\
&= \mathrm{E}[\mathrm{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^T)] \\
&= \mathrm{tr}(\mathbf{V}\mathrm{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T])
\end{aligned}
$$

Since $\mathrm{E}\left[\beta_i\beta_j\right] = 0$ for all $i \neq j$, this simplifies to

$$h_g^2 = \sum_{i=1}^{M} \sigma_i^2 \tag{2.2}$$

Thus, h_g^2 is defined with respect to a given population and a given set of SNPs. By definition, $0 \leq h_g^2 \leq 1$. Similarly, we define regional SNP-heritability (h_k^2) to be the proportion of phenotypic variance due to the additive effects of the genotyped SNPs in region k. We assume that the set of SNPs that defines h_k^2 is a subset of the M SNPs that define h_g^2 (thus, $0 \leq h_k^2 \leq h_g^2$). If region k is the whole genome, $h_k^2 = h_g^2$.

2.4.2 Estimating SNP-heritability under the GRE model

We are interested in estimating h_g^2 under the GRE model (Equation 2.2). In a GWAS with N individuals genotyped at M SNPs, let $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$ be the $N \times M$ matrix of standardized genotypes (each column of $\mathbf{X}$ has been standardized to have mean 0 and variance 1), $\mathbf{y} = (y_1, \ldots, y_N)^T$ be the N-vector of standardized phenotypes, and $\widehat{\mathbf{V}} = (1/N)\mathbf{X}^T\mathbf{X}$ be the $M \times M$ in-sample LD matrix (an estimate of population LD, $\mathbf{V}$) with rank q, where $1 \leq q \leq M$. Let $\mathbf{X} =$

$(\mathbf{X}_1, \dots, \mathbf{X}_K)$ be the genotype matrices for K independent regions spanning all M SNPs (e.g., chromosomes). For region k containing p_k SNPs, $\mathbf{X}_k$ is the $N \times p_k$ standardized genotype matrix and $\widehat{\mathbf{V}}_k$ is the corresponding $p_k \times p_k$ in-sample LD matrix with rank q_k $(1 \leq q_k \leq p_k)$. We propose the following estimator for genome-wide SNP-heritability:

$$\hat{h}^2_{\text{GRE}} = \sum_{k=1}^{K} \frac{N \widehat{\boldsymbol{\beta}}_k^T \widehat{\mathbf{V}}_k^{\dagger} \widehat{\boldsymbol{\beta}}_k - q_k}{N - q_k}$$

where $\widehat{\boldsymbol{\beta}}_k = (1/N)\mathbf{X}_k^T \mathbf{y}$ is the p_k-vector of marginal SNP effects estimated by ordinary least squares (OLS) for region k and $\widehat{\mathbf{V}}_k^{\dagger}$ is the pseudoinverse of $\widehat{\mathbf{V}}_k$. Detailed derivations for $\hat{h}^2_{\text{GRE}}$ can be found in the Supplementary Note, which is freely available online[37].

2.4.3 Analytical variance of the GRE estimator

Following quadratic form theory[74,84], the variance of $\hat{h}^2_{\text{GRE}}$ in the single-block case is

$$\text{Var}\left[\hat{h}^2_{\text{GRE}}\right] = \left(\frac{N}{N - q}\right)^2 \left(2q\left(\frac{1 - h_g^2}{N}\right) + 4h_g^2\right)\left(\frac{1 - h_g^2}{N}\right) \tag{2.3}$$

When using the K-block approximation, which assumes that the blocks are independent, we approximate Equation 2.3 as the sum of the variances of the local SNP-heritabilities:

$$\text{Var}\left[\hat{h}^2_{\text{GRE}}\right] = \sum_{\text{k}=1}^{\text{K}} \left(\frac{N}{N - q_k}\right)^2 \left(2q_k\left(\frac{1 - h_k^2}{N}\right) + 4h_k^2\right)\left(\frac{1 - h_k^2}{N}\right) \tag{2.4}$$

Equation 2.3 is estimated by plugging in $\hat{h}^2_{\text{GRE}}$ and Equation 2.4 is estimated by plugging in $(\hat{h}^2_1, \dots, \hat{h}^2_K)$, the estimates of the regional SNP-heritabilities.

2.4.4 Simulation Framework

We simulated quantitative phenotypes from real genotype array data (UK Biobank[11]) under a range of genetic architectures. We obtained a set of $N = 337205$ unrelated British individuals by extracting individuals with self-reported British ancestry who are > 3rd degree relatives (pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$) and excluding individuals with putative sex chromosome aneuploidy[11]. In all simulations, we standardize the genotypes before drawing phenotypes. That is, for each SNP i and individual n, we compute $x_{ni} = (g_{ni} - 2f_i)/\sqrt{2f_i(1-f_i)}$, where $g_{ni} \in \{0,1,2\}$ is the number of minor alleles and f_i is the in-sample minor allele frequency (MAF).

2.4.4.1 Simulations of quantitative traits with no population stratification

Given $\mathbf{X}$ and a fixed value of h_g^2, phenotypes are drawn according to the following model. The proportion of causal variants, p_{causal}, is set to 1, 0.01, or 0.001. Let $c_i \in \{0,1\}$ be the causal status of SNP i. If $p_{\text{causal}} = 1$, $c_i = 1$ for $i = 1, \dots, M$. If $0 \leq p_{\text{causal}} < 1$, we draw $p_{\text{causal}} \times M$ SNPs from the set of SNPs with MAF in one of three ranges: (0, 0.5], (0.01, 0.05], or (0.05, 0.5]. We use "CV MAF" to refer to the MAF range from which the causal variants are drawn. Standardized effects and phenotypes are then drawn according to the model

$$\sigma_i^2 \propto c_i \cdot w_i^{\gamma} [2f_i(1-f_i)]^{1+\alpha} \tag{2.5}$$

$$(\beta_1, \dots, \beta_k)^T \sim N\big(0, \operatorname{diag}(\sigma_1^2, \dots, \sigma_M^2)\big) \tag{2.6}$$

$$(y_1, \dots, y_N)^T | \boldsymbol{\beta} \sim N\big(\mathbf{X}\boldsymbol{\beta}, (1 - h_g^2)\mathbf{I}_N\big) \tag{2.7}$$

where α controls the coupling of MAF and effect size, w_i is a SNP-specific LD weight, and $\gamma \in \{0,1\}$ specifies whether effects are coupled with the LD weights. We simulate two types of LD-dependent architectures by defining $w_1, \dots, w_M$ to be either (1) the default "LDAK weights" computed by the LDAK software[35,59], or (2) the inverse unpartitioned "LD score" of each SNP computed within a 2-Mb window ($w_i^{-1} = \sum_j v_{ij}^2$ where j indexes the set of SNPs within a 2-Mb window centered on SNP i)[62]. When $\gamma = 1$, both the LDAK weights and inverse LD score weights cause SNPs in regions of higher LD to have smaller effects than do SNPs in regions of lower LD. We set α to one of two values: $\alpha = -1$ (a relatively strong inverse relationship between MAF and effect size) or $\alpha = -0.25$ (a weaker inverse relationship between MAF and effect size). Each per-SNP variance is multiplied by a scaling factor so that $\sum_{i=1}^{M} \sigma_i^2 = h_g^2$. Note that $\sigma_i^2 = 0$ if $c_i = 0$.

Finally, given phenotypes $\mathbf{y} = (y_1, \dots, y_N)^T$ and genotypes $\mathbf{X} = (\mathbf{x}_1^T, \dots, \mathbf{x}_N^T)^T$, we compute marginal association statistics through ordinary least squares (OLS): $\widehat{\boldsymbol{\beta}} = (1/N)\mathbf{X}^T\mathbf{y}$.

2.4.4.2 Simulations of case-control phenotypes with no population stratification

To simulate case-control studies, we first draw each individual's continuous liability (l_n for individual *n*) according to Equation 2.7. For a given population prevalence ($0 \leq d_{pop} \leq 1$), we compute the corresponding liability threshold $L = \Phi^{-1}(1 - d_{pop})$, where Φ is the CDF of the standard normal distribution. Each l_n is then converted into a case-control status: $y_n = 1$ if $l_n \geq L$ or $y_n = 0$ if $l_n < L$. For unascertained case-control studies, we assume that the proportion of cases in the study is equal to the population prevalence ($d_{GWAS} = d_{pop}$). For ascertained case-control studies ($d_{GWAS} > d_{pop}$), we set $d_{GWAS} = 0.5$ and select a random set of controls to satisfy $N_{case} = N_{control}$.

We compute association statistics by regressing the binary case-control statuses on genotypes. The GRE estimator produces an estimate of SNP-heritability on the *observed* scale

($\hat{h}^2_{obs}$). Assuming we know the population prevalence, we convert $\hat{h}^2_{obs}$ to the *liability* scale with the transformation $\hat{h}^2_{liab} = \hat{h}^2_{obs} d^2_{pop} \left(1 - d_{pop}\right)^2 / ([f(L)]^2 d_{GWAS}(1 - d_{GWAS}))$, where f is the standard normal probability density function[85].

2.4.4.3 Simulations with population stratification

To simulate GWAS with population stratification, we draw phenotypes from a model where a covariate that is correlated to genotypes has a nonzero effect on phenotype. To this end, we simulate an effect of the first genetic principal component ($\mathbf{PC}_1$). Letting σ^2_s be the proportion of total phenotypic variance explained by $\mathbf{PC}_1$, phenotypes are drawn from the model

$$(y_1, \dots, y_N)^T | \boldsymbol{\beta} \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{PC}_1 \beta_s, \left(1 - h^2_g - \sigma^2_s\right)\mathbf{I}_N\right)$$

where $\mathrm{Var}[\mathbf{PC}_1 \beta_s] / \mathrm{Var}[\mathbf{y}] = \beta^2_s \mathrm{Var}[\mathbf{PC}_1] = \sigma^2_s$. We compute association statistics from one of two models: $\mathbf{y} = \mathbf{X}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$, which ignores population stratification and other sources of confounding, or $\mathbf{y} = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{PC}_1 \beta_s + \boldsymbol{\epsilon}$, which controls for the effect of $\mathbf{PC}_1$.

2.4.5 Comparison of methods in simulations

Unless otherwise specified, in all genome-wide simulations, we use real genotypes of N = 337,205 unrelated British individuals measured at M = 593,300 array SNPs to draw causal effects for all M SNPs and phenotypes for all N individuals. OLS summary statistics are computed for all M SNPs using the simulated phenotypes and real genotypes of all N individuals. We compare to three methods that operate on summary statistics and are computationally tractable for these simulations: LD score regression (LDSC)[62], stratified LD score regression (S-LDSC)[63,64], and SumHer[34].

For LDSC and S-LDSC, we compute the unpartitioned LD score of each SNP as a function of its LD to all other SNPs in a 2-Mb window centered on the SNP. For each annotation included in S-LDSC, the partitioned LD score of each SNP is a function of its LD to all SNPs within a 2-Mb window that are in the annotation. For both LDSC and S-LDSC, LD scores are computed with the LDSC software (https://github.com/bulik/ldsc/) from a random sample of 40K individuals to reduce the amount of memory required by the LDSC software. We run the regression with an unconstrained intercept, using all M SNPs as observations in the response variable. Each SNP is weighted to account for heteroscedasticity and correlations between association statistics[62]. For both methods, h_g^2 is estimated as a function of all M SNP-specific variances by using the flags --not-M-5-50 and --chisq-max 99999 (the latter option prevents the LDSC software from dropping high-effect SNPs).

We run S-LDSC in two ways to account for MAF/LD-dependent architectures. S-LDSC (MAF) refers to S-LDSC with 10 binary MAF bin annotations (each bin contains exactly 10% of the typed SNPs), which is intended to mirror the 10 MAF annotations in the "baseline-LD model"[63,64] (see Supplementary Table 14 for precise MAF bin ranges for the UK Biobank Axiom Array). S-LDSC (MAF+LLD) refers to S-LDSC with the same 10 MAF bins and an additional continuous "level of LD" (LLD) annotation computed by quantile-normalizing the unpartitioned LD scores within each MAF bin to a standard normal distribution[64]. While our definition of LLD is intended to mirror the LLD annotation in the baseline-LD model, we do not set the LLD of variants with MAF < 0.05 to 0 because our estimand of interest includes the effects of SNPs with MAF < 0.05.

To run SumHer, we use the LDAK software (https://dougspeed.com/ldak/) to compute the default "LDAK weights" using in-sample LD[34,35,59]. We then compute "LD tagging" (i.e. LD scores) using 1-Mb windows centered on each SNP and setting $\alpha = -0.25$ as recommended[34]. The LDAK software is memory-efficient, allowing us to use all 337K individuals to compute LDAK weights and LD tagging. Unless otherwise specified, all default parameter settings are used to run SumHer in simulations.

We also perform simulations with N = 8,430 unrelated individuals at M = 14,821 array SNPs. These individuals and SNPs are a subset of the data used in the genome-wide simulations, chosen by selecting approximately 2.5% of individuals and the first 2.5% of SNPs from the beginning of each chromosome in order to preserve the LD structure among the SNPs. We run single-component GREML[52,86] (GCTA software: https://cnsgenomics.com/software/gcta/) and single-component BOLT-REML[61] (https://data.broadinstitute.org/alkesgroup/BOLT-LMM/) with default parameters. We run GREML-LDMS-I[60,67] using 8 GRMs created from 2 MAF bins (MAF $\leq$ 0.05 and MAF > 0.05) and 4 LD score quartiles; LD scores were computed using the GCTA software with the default window size of 200-kb. We run LDAK using the default LDAK weights, setting $\alpha = -0.25$ as recommended[35,59].

A third set of simulations was performed using 7,685 individuals of South Asian ancestry in the UK Biobank. This group was composed of individuals of Indian (n = 5,716), Pakistani (n = 1,748), and Bangladeshi (n = 221) ancestry. Due to the small sample size, we used a reduced set of 803 SNPs from chromosome 21 and 839 SNPs from chromosome 22 (1,642 SNPs in total) which were chosen so that N/p_k for each chromosome k was similar to N/p_k in the "white British" cohort.

For a given genetic architecture, we generate 100 simulation replicates and obtain 100 estimates of h_g^2 from each method. We estimate the bias of an estimator $\hat{h}_g^2$ under a given architecture as $\text{bias}\left[\hat{h}_g^2\right] = \text{E}\left[\hat{h}_g^2\right] - h_g^2 \approx (1/100)\sum_{i=1}^{100} \hat{h}_g^2(i) - h_g^2$ where $\hat{h}_g^2(i)$ is the estimate from the i-th simulation. To test whether the bias is statistically significant (null hypothesis: $\text{bias}\left[\hat{h}_g^2\right] = 0$), we assess the z-score of the bias ($z_{\text{bias}} = \text{bias}\left[\hat{h}_g^2\right]/\text{SEM}[\hat{h}_g^2]$, where $\text{SEM}\left[\hat{h}_g^2\right]$ is the standard error of the mean of the 100 estimates) which follows a $N(0,1)$ distribution under the null hypothesis. The p-value of the bias is computed with a two-tailed test. To enable a comparison of estimators across different values of h_g^2, we assess the relative bias of an estimator under a single architecture ($\text{bias}\left[\hat{h}_g^2\right]/h_g^2$) as a percentage of h_g^2. In Figure 2.1a and 2.1c, we compute the error of a single estimate as $(\hat{h}_g^2(i) - h_g^2)/h_g^2$; errors are also reported as percentages of h_g^2.

2.4.6 Analysis of UK Biobank phenotypes

We estimate SNP-heritability for 22 complex traits (6 quantitative, 16 binary) in the UK Biobank[11]. We use PLINK[87,88] (https://www.cog-genomics.org/plink2) to exclude SNPs with MAF < 0.01 and genotype missingness > 0.01 as well as SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7}. We keep only the individuals with self-reported white British ancestry and no kinship (i.e. > 3rd degree relatives, defined in ref.[11] as pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$). After removing individuals who are outliers for genotype heterozygosity and/or missingness, we obtain a set of N = 290,641 individuals to use in the real data analyses. For all traits, marginal association statistics are computed through OLS in PLINK, using age, sex, and the top 20 genetic principal components (PCs) as covariates in the regression; these 20 PCs were precomputed by UK Biobank from a superset of 488,295 individuals. Additional covariates were used for waist-to-hip ratio (adjusted for BMI) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives). We compute $\hat{h}^2_{\text{GRE}}$ for each trait using in-sample LD estimated from all *N* individuals.

When using LDSC, S-LDSC, or SumHer to estimate SNP-heritability, it is necessary to define and distinguish between the following sets of SNPs: the set of SNPs containing all possible causal SNPs of interest (used to compute LD scores and LDAK weights), the set of SNPs used as observations in the regression, and the set of SNPs that defines the SNP-heritability estimand of interest. We run two versions of LDSC, S-LDSC (controlling for the most recent baseline-LD model[83]), and SumHer. First, to enable a direct comparison between $\hat{h}^2_{\text{GRE}}$ and the estimands of LDSC, S-LDSC, and SumHer, we run an "in-sample LD" version of each method where the *M* typed SNPs are used to compute LD scores and LDAK weights, perform the regression, and define the SNP-heritability estimand of interest. We refer to these as LDSC (in-sample), S-LDSC (baseline-LD/in-sample), and SumHer (in-sample). To run LDSC (in-sample) and S-LDSC (baseline-LD/in-sample), we use the LDSC software to compute LD scores and regression weights within 2-Mb windows centered on each SNP, using a random sample of 40K individuals to reduce the memory requirement. To run SumHer (in-sample), we use the LDAK software to compute LD tagging from the genotypes of all *N* individuals, using 1-Mb windows centered on each SNP and setting $\alpha = -0.25$ as recommended[35]. Unless otherwise specified, all other parameters were set to the default settings.

To enable comparisons between $\hat{h}^2_{\text{GRE}}$ and estimates reported in the literature, we also run each method with its recommended parameter settings and LD estimated from reference panel sequencing data. We refer to these methods as LDSC (1KG), S-LDSC (baseline-LD/1KG), and SumHer (1KG) to indicate that LD is estimated from 489 Europeans in the 1000 Genomes Phase 3 reference panel[7]. We run LDSC (1KG) and S-LDSC (baseline-LD/1KG) with LD scores and regression weights (1-cM windows) from 9,997,231 SNPs with minor allele count greater than 5 in the reference panel, and we define the SNP-heritability estimand to be a function of the array SNPs with MAF > 0.05. We run SumHer (1KG) using 8,569,062 SNPs with MAF > 0.01 in the reference panel to compute LDAK weights and LD tagging (1-cM windows) and to define the SNP-heritability estimand; we control for a multiplicative inflation of test statistics as recommended[34]. See refs.[34,62–64] for details about the definitions and interpretations of the estimands of LDSC, S-LDSC, and SumHer.

2.5 Figures

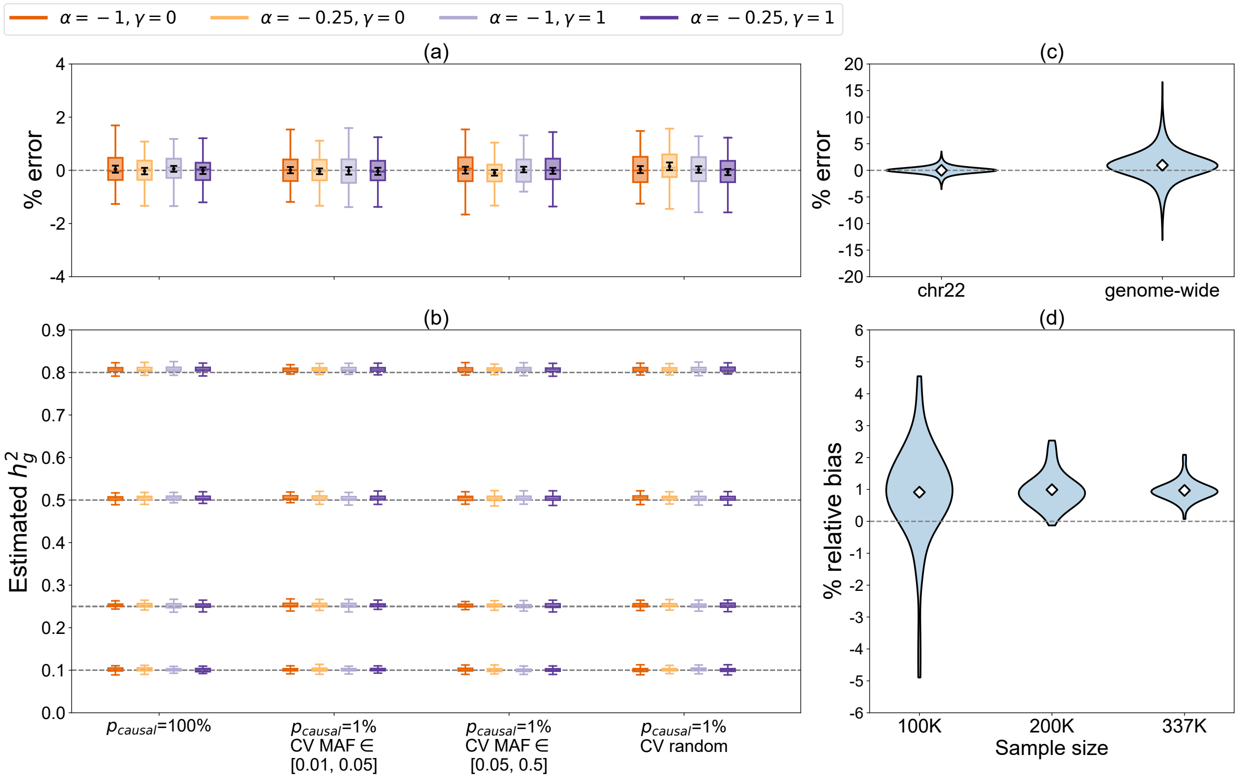

Figure 2.1 Performance of GRE estimator in simulations.

Performance of $\hat{h}^2_{\mathrm{GRE}}$ in simulations under 64 distinct MAF/LD-dependent architectures (*N*=337205). (a) Distribution of errors $\hat{h}^2_{\mathrm{GRE}}(i) - h^2_g$ as a percentage of h^2_g, where $\hat{h}^2_{\mathrm{GRE}}(i)$ is the estimate from the i-th simulation under a given genetic architecture, in simulations on chromosome 22 (*M*=9654 SNPs). $\hat{h}^2_{\mathrm{GRE}}$ was computed with 1 chromosome-wide LD block. Black points and error bars mark the mean and ± 2 standard errors of the mean (SEM). (b) Distribution of $\hat{h}^2_{\mathrm{GRE}}$ in genome-wide simulations (*M*=593300 SNPs) where $\hat{h}^2_{\mathrm{GRE}}$ was computed with 22 chromosome-wide LD blocks. In both (a) and (b), each boxplot represents estimates from 100 simulations. Boxplot whiskers extend to the minimum and maximum estimates located within $1.5 \times$ IQR from the first and third quartiles, respectively. (c) Errors for chromosome 22 and genome-wide simulations. Each violin plot represents the errors of 6400 estimates (64 genetic architectures × 100 simulation replicates). (d) Relative bias (as a percentage of h^2_g) as a function of sample size *N* in genome-wide simulations. Each violin plot represents 64 estimates of relative bias. In (c) and (d), the white diamonds mark the mean of each distribution.

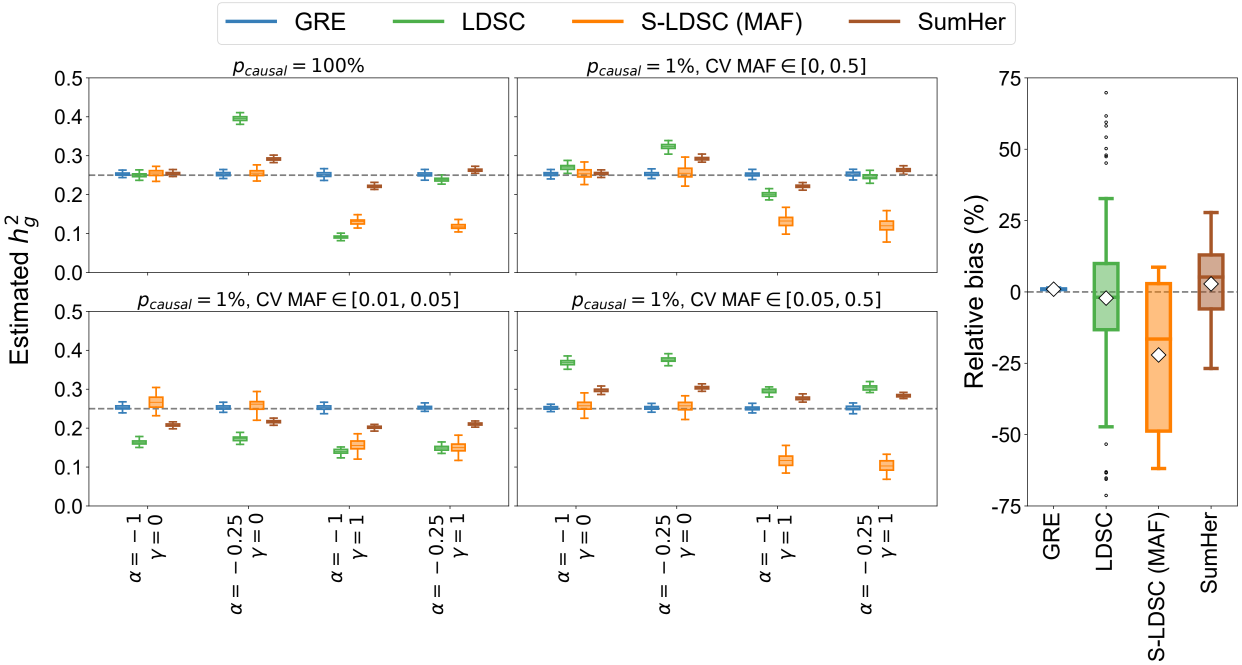

Figure 2.2 Comparison of GRE, LDSC, S-LDSC, and SumHer in genome-wide simulations. Left: Phenotypes were drawn under one of 16 MAF- and/or LDAK-LD-dependent architectures by varying p_{causal}, α, γ, and CV MAF (Methods). Each boxplot contains estimates of h_g^2 from 100 simulations. Right: Relative bias of each method (as a percentage of h_g^2) across 112 distinct MAF- and LDAK-LD-dependent architectures (Methods). Each boxplot contains 112 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture. The white diamonds mark the average of each distribution. Boxplot whiskers extend to the minimum and maximum estimates located within $1.5 \times$ IQR from the first and third quartiles, respectively.

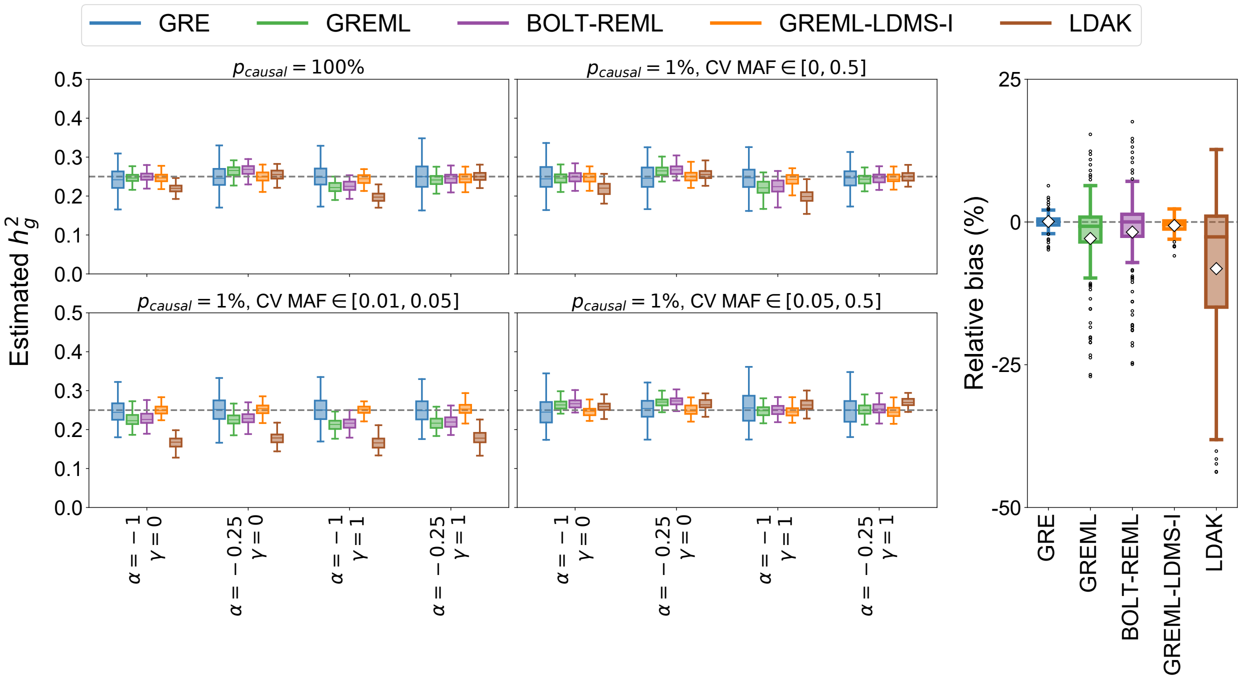

Figure 2.3 Comparison of GRE with REML-based methods (N=8430, M=14821 SNPs). Left: Phenotypes were drawn under one of 16 MAF- and/or LDAK-LD-dependent architectures by varying p_{causal}, α, γ, and CV MAF (Methods). Each boxplot contains estimates of h_g^2 from 100 simulations. Right: Relative bias of each method (as a percentage of the true h_g^2) across 112 distinct MAF- and LDAK-LD-dependent architectures (Methods). Each boxplot represents the distribution of 112 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture. The white diamonds mark the average of each distribution. Boxplot whiskers extend to the minimum and maximum estimates located within $1.5 \times$ IQR from the first and third quartiles, respectively.

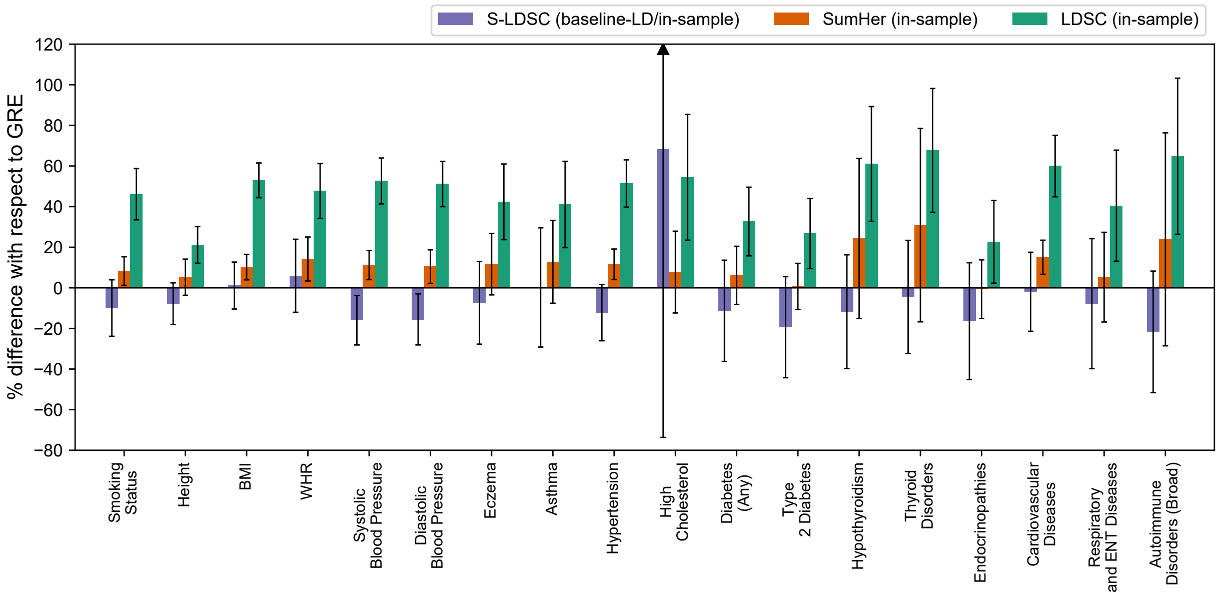

Figure 2.4 Comparison of GRE, LDSC, S-LDSC, and SumHer for 18 traits in the UK Biobank. Estimates from LDSC, S-LDSC, and SumHer expressed as % difference with respect to $\hat{h}^2_{\mathrm{GRE}}$ for 18 complex traits and diseases in the UK Biobank for which $\hat{h}^2_{\mathrm{GRE}} > 0.05$ (N=290K unrelated British individuals, M=460K typed SNPs, in-sample LD; Methods). Each bar represents the difference between the estimated h^2_g from one of the methods (LDSC, S-LDSC, or SumHer) and $\hat{h}^2_{\mathrm{GRE}}$ as a percentage of $\hat{h}^2_{\mathrm{GRE}}$. Black bars mark ± 2 standard errors.

2.6 Tables

Model	Assumptions on $\boldsymbol{\beta_i}$	Description
Generalized random effects	$\mathrm{E}[\beta_i] = 0$, $\mathrm{Var}[\beta_i] = \sigma_i^2$, $\sigma_i^2 \geq 0$	Each SNP i has a nonnegative SNP-specific variance σ_i^2. Total SNP-heritability is $h_g^2 \equiv \sum_{i=1}^{M} \sigma_i^2$.
GREML-SC [52,61,66]	$\beta_i \sim N(0, h_g^2/M)$	Each SNP explains an equal portion of h_g^2. In other words, $\sigma_i^2 = h_g^2/M$ for all $i = 1, \ldots, M$.
GREML-MC [60,61,85,89]	$\beta_i \sim N(0, \sum_{c \in C}[\mathrm{SNP}_i \in c] h_c^2/m_c)$	h_g^2 is partitioned by a set of disjoint SNP partitions C that span all M SNPs. Partition $c \in C$ contains m_c SNPs that have per-SNP variances h_c^2/m_c. Total SNP-heritability is $h_g^2 = \sum_{c \in C} h_c^2$.
LDAK[35,59]	$\beta_i \sim N(0, \sigma_i^2)$, $\sigma_i^2 \propto w_i[f_i(1-f_i)]^{1+\alpha}$	Each SNP-specific variance is proportional to a function of f_i (the MAF of SNP i) and to w_i (a SNP-specific weight that is a function of the inverse of the LD score of SNP i). α controls the relationship between σ_i^2 and f_i. The most recent recommendation by ref.[9] is to assume $\alpha = -0.25$.
LDSC[62]	$\mathrm{E}[\beta_i] = 0$, $\mathrm{Var}[\beta_i] = h_g^2/M$	Each SNP explains an equal portion of h_g^2 (similar to the GREML-SC model when h_g^2 is defined with respect to the same set of M SNPs).
S-LDSC[63,64,83]	$\mathrm{E}[\beta_i] = 0$, $\mathrm{Var}[\beta_i] = \sum_{a \in A} \tau_a a(i)$	Each SNP-specific variance is a linear function of a set of annotations A where each $a \in A$ represents a binary or continuous-valued annotation. $a(i)$ is the value of annotation a at SNP i. τ_a is the expected contribution of a one-unit increase in annotation a to each SNP-specific variance.
SumHer[34]	$\mathrm{E}[\beta_i] = 0$, $\mathrm{Var}[\beta_i] \propto w_i[f_i(1-f_i)]^{1+\alpha}$	An extension of the LDAK model to operate on summary-level data; can also efficiently partition h_g^2 by multiple annotations. The most recent recommendations by refs.[9,14] is to set $\alpha = -0.25$.

Table 2.1 The assumptions made by existing methods are subsumed under the GRE model. Existing methods to estimate SNP-heritability impose additional assumptions on top of the generalized random effects (GRE) model. Under the GRE model, the causal effects at any two SNPs are assumed to be independent ($\mathrm{E}[\beta_i\beta_j] = 0$ for all $i \neq j$) and genome-wide SNP-heritability is defined as $h_g^2 \equiv \sum_{i=1}^{M} \sigma_i^2$, where each σ_i^2 can be an arbitrary nonnegative real number as long as $0 \leq h_g^2 \leq 1$ (Methods). All existing methods make assumptions on the distribution of β_i and/or the form of σ_i^2 that can be subsumed under the GRE model. To simplify notation, we assume for each model that phenotypes are standardized in the population (i.e. $\mathrm{Var}[y_n] = 1$ for every individual n).

Trait	GRE	S.E.	LDSC	S.E.	S-LDSC	S.E.	SumHer	S.E.
Smoking Status	0.122	3.90E-03	0.178	7.70E-03	0.110	8.50E-03	0.132	4.30E-03
Height	0.602	4.70E-03	0.730	2.70E-02	0.555	3.10E-02	0.634	2.70E-02
BMI	0.285	4.20E-03	0.436	1.20E-02	0.289	1.70E-02	0.315	9.00E-03
WHR	0.173	4.00E-03	0.256	1.20E-02	0.184	1.60E-02	0.198	9.40E-03
Systolic Blood Pressure	0.159	4.20E-03	0.243	9.00E-03	0.134	9.70E-03	0.177	5.70E-03
Diastolic Blood Pressure	0.154	4.20E-03	0.233	8.60E-03	0.130	9.70E-03	0.170	6.40E-03
Eczema	0.116	4.20E-03	0.165	1.10E-02	0.107	1.20E-02	0.130	8.80E-03
Asthma	0.116	4.90E-03	0.163	1.20E-02	0.116	1.70E-02	0.131	1.20E-02
Hypertension	0.162	4.00E-03	0.244	9.40E-03	0.142	1.10E-02	0.180	6.10E-03
High Cholesterol	0.082	5.10E-03	0.127	1.30E-02	0.138	5.80E-02	0.088	8.30E-03
Diabetes (Any)	0.070	3.70E-03	0.093	5.90E-03	0.062	8.70E-03	0.074	5.00E-03
Type 2 Diabetes	0.071	3.80E-03	0.090	6.10E-03	0.057	8.80E-03	0.071	4.00E-03
Hypothyroidism	0.088	5.20E-03	0.142	1.30E-02	0.078	1.20E-02	0.110	1.70E-02
Thyroid Disorders	0.084	5.20E-03	0.141	1.30E-02	0.080	1.20E-02	0.110	2.00E-02
Endocrinopathies	0.069	5.10E-03	0.084	7.00E-03	0.058	9.90E-03	0.068	5.00E-03
Cardiovascular Diseases	0.143	5.30E-03	0.228	1.10E-02	0.140	1.40E-02	0.164	6.00E-03
Respiratory and ENT Diseases	0.086	5.20E-03	0.120	1.20E-02	0.079	1.40E-02	0.090	9.50E-03
Psoriasis	0.019	5.00E-03	0.071	3.10E-02	0.035	1.20E-02	0.059	4.20E-02
Dermatologic Disorders	0.023	5.00E-03	0.049	1.40E-02	0.034	9.90E-03	0.031	1.10E-02
Rheumatoid Arthritis	0.008	5.00E-03	0.041	2.10E-02	0.010	7.90E-03	0.021	1.20E-02
Autoimmune Disorders (Broad)	0.063	5.10E-03	0.105	1.20E-02	0.050	9.50E-03	0.079	1.70E-02
Autoimmune Disorders (Certain)	0.015	5.00E-03	0.052	2.60E-02	0.005	7.60E-03	0.047	3.40E-02

Table 2.2 Estimates from GRE, LDSC, S-LDSC, and SumHer for 22 complex traits (N=290K). Estimates of h_g^2 from the GRE approach, LDSC (in-sample), S-LDSC (baseline-LD/in-sample), and SumHer (in-sample) for 22 complex traits and diseases in the UK Biobank (*N*=290K unrelated British individuals, *M*=460K typed SNPs).

3 Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data

3.1 Introduction

Genetic and phenotypic variations among humans have been shaped by many factors, including migration histories, geodemographic events, and environmental background[38–41,44]. As a result, the underlying genetic architecture of a given complex trait – defined here in terms of 'polygenicity' (the number of variants with nonzero effects)[30,54,71,72,90] and the coupling of causal effect sizes with minor allele frequency (MAF)[52,91], linkage disequilibrium (LD)[35,59,64], and other genomic features[63] – varies among ancestral populations. While the vast majority of genome-wide association studies (GWAS) to date have been performed in individuals of European descent[1,6,92,93], growing numbers of studies performed in individuals of non-European ancestry[23,94–99] have created opportunities for well-powered transethnic genetic studies[42,49,100–103].

Risk regions identified through GWAS tend to replicate across populations[1,42,104,105], indicating that complex traits have shared genetic components among populations. Indeed, for certain post-GWAS analyses such as disease mapping[95,102,106] and statistical fine-mapping[100,107–110], under the assumption that two populations share one or more causal variants, population-

This chapter is published in Shi*, Burch*, et al. *Am J Hum Genet* (2020). Citations have been updated and renumbered. Supplemental material is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273527/

specific LD patterns can be leveraged to improve performance over approaches that model a single population. On the other hand, several studies have shown that heterogeneity in genetic architectures limits transferability of polygenic risk scores (PRS) across populations[43–48,111–113]; critically, if applied in a clinical setting, existing PRS may exacerbate health disparities among ethnic groups[114]. The population-specificity of existing PRS as well as estimates of transethnic genetic correlations less than one reported in the literature[49,50,115–117] indicate that (1) LD tagging and allele frequencies of shared causal variants vary across populations, (2) that a sizeable number of causal variants are population-specific, and/or (3) that causal effect sizes vary across populations due to, for example, different gene-environment interactions. For example, due to population-specific LD, a single genetic variant that is significantly associated with a trait in two populations may actually be tagging distinct population-specific causal variants (Figure 3.1). Conversely, two distinct associations in two populations may be driven by the same underlying causal variants (i.e. colocalization). Thus, identifying shared and population-specific components of genetic architecture could help improve transethnic analyses (e.g., transferability of PRS across populations[43,45,48,93,111]) and uncover novel disease etiologies.

In this work, we introduce PESCA (Population-spEcific/Shared Causal vAriants), an approach that requires only GWAS summary association statistics and ancestry-matched estimates of LD to infer genome-wide proportions of population-specific and shared causal variants for a single trait in two populations; the genome-wide estimates are then used as priors in an empirical Bayes framework to localize and test for enrichment of population-specific/shared causal variants in regions of interest. In this context, a “causal variant” is a variant measured in the given GWAS that either has a nonzero effect on the trait (e.g., a nonsynonymous variant that alters protein folding) or tags a nonzero effect at an unmeasured variant through LD. It is therefore important to note that the set of "causal variants" that PESCA aims to identify is defined with respect to the set of variants included in the GWAS and can contain variants with indirect nonzero effects that are statistical rather than biological in nature (this is analogous to the definition of SNP-heritability, which is also a function of a specific set of SNPs[37,52,62,74]). Through extensive simulations, we show that our method yields approximately unbiased estimates of the proportions of population-specific/shared causal variants if in-sample LD is used and slightly upward-biased estimates if LD is estimated from an external reference panel. We then show that using these estimates as priors to perform fine-mapping (Ch. 3.2.3 Material and Methods) produces well-calibrated per-SNP posterior probabilities and enrichment test statistics. We note that the definition of enrichment used here is related to, but conceptually distinct from, definitions of SNP-heritability enrichment[63,64]. Under our framework, an enrichment of causal SNPs greater than 1 indicates that, compared to the genome-wide background, there are more causal SNPs in that region than expected[118,119] (Ch. 3.2.6 Material and Methods). In contrast, an enrichment of SNP-heritability greater than 1 indicates that the average per-SNP effect size in the region is larger than the genome-wide average per-SNP effect size.

We apply our approach to publicly available GWAS summary statistics for 9 complex traits and diseases in individuals of East Asian (EAS) and European (EUR) ancestry (average $N_{EAS} = 94{,}621$, $N_{EUR} = 103{,}507$) (Table 3.1), restricting to common SNPs (MAF > 5%) and using 1000 Genomes[7] to estimate ancestry-matched LD. On average across the 9 traits, we estimate that approximately 80% (S.D. 15%) of common SNPs that are causal in EAS and 84% (S.D. 8%) of those in EUR are shared by the other population. Consistent with previous studies based on SNP-heritability[61,74], we find that high-posterior SNPs are distributed uniformly across the genome. We observe that population-specific GWAS risk regions have, on average across the 9 traits, a 2.8x enrichment of shared high-posterior SNPs relative to the genome-wide background, suggesting that many EAS-specific and EUR-specific GWAS risk regions harbor shared causal SNPs that are undetected in the other population due to differences in LD, allele frequencies, and/or GWAS sample size. The effects of SNPs with posterior probability > 0.8 of being causal (for any causal configuration) are highly correlated between EAS and EUR, concordant with replication slopes between EAS and EUR marginal effects close to 1 that have been reported for several complex diseases[42] and with strong transethnic genetic correlations previously reported for the same traits analyzed in this work (average $\hat{\rho}_g = 0.79 \pm 0.07$ s.e.m. across the 9 traits)[116]. Finally, we show that regions flanking genes that are specifically expressed in trait-relevant tissues[120] harbor a disproportionate number of shared high-posterior SNPs. Many of the same tissue-specific gene sets are also enriched with SNP-heritability, implying that SNP-heritability enrichments are driven by many low-effect SNPs rather than a small number of high-effect SNPs. Our results suggest that common causal SNPs have similar etiological roles in EAS and EUR and that transferability of PRS and other GWAS findings across populations can be improved by explicitly correcting for population-specific LD and allele frequencies.

We apply our approach to publicly available GWAS summary statistics for 9 complex traits and diseases in individuals of East Asian (EAS) and European (EUR) ancestry (average N_{EAS} = 94,621, N_{EUR} = 103,507) (Table 3.1), restricting to common SNPs (MAF > 5%) and using 1000 Genomes[59] to estimate ancestry-matched LD. On average across the 9 traits, we estimate that approximately 80% (S.D. 15%) of common SNPs that are causal in EAS and 84% (S.D. 8%) of those in EUR are shared by the other population. Consistent with previous studies based on SNP-heritability[55,60], we find that high-posterior SNPs are distributed uniformly across the genome. We observe that population-specific GWAS risk regions have, on average across the 9 traits, a 2.8x enrichment of shared high-posterior SNPs relative to the genome-wide background, suggesting that many EAS-specific and EUR-specific GWAS risk regions harbor shared causal SNPs that are undetected in the other population due to differences in LD, allele frequencies, and/or GWAS sample size. The effects of SNPs with posterior probability > 0.8 of being causal (for any causal configuration) are highly correlated between EAS and EUR, concordant with replication slopes between EAS and EUR marginal effects close to 1 that have been reported for several complex diseases[33] and with strong transethnic genetic correlations previously reported for the same traits analyzed in this work (average $\hat{\rho}_g = 0.79 \pm 0.07$ s.e.m. across the 9 traits)[51]. Finally, we show that regions flanking genes that are specifically expressed in trait-relevant tissues[61] harbor a disproportionate number of shared high-posterior SNPs – many of the same tissue-specific gene sets are also enriched with SNP-heritability, implying that SNP-heritability enrichments are driven by many low-effect SNPs rather than a small number of high-effect SNPs. Our results suggest that common causal SNPs have similar etiological roles in EAS and EUR and that transferability of PRS and other GWAS findings across populations can be improved by explicitly correcting for population-specific LD and allele frequencies.

3.2 Material and Methods

3.2.1 Distribution of GWAS summary statistics in two populations

For a given complex trait, we model the causal statuses of SNP i in two populations as a binary vector of size two, $\boldsymbol{C}_i = c_{i1}c_{i2}$, where each bit, $c_{i1} \in \{0,1\}$ and $c_{i2} \in \{0,1\}$, represents the causal status of SNP i in populations 1 and 2, respectively. $\boldsymbol{C}_i = 00$ indicates that SNP i is not causal in either population; $\boldsymbol{C}_i = 01$ and $\boldsymbol{C}_i = 10$ indicate that SNP i is causal only in the first and second population, respectively; and $\boldsymbol{C}_i = 11$ indicates that SNP i is causal in both populations. We assume $\boldsymbol{C}_i$ follows a multivariate Bernoulli (MVB) distribution[121,122]

$$\boldsymbol{C}_i \sim \mathrm{MVB}(f_{00}, f_{01}, f_{10}, f_{11})$$

in order to facilitate optimization and interpretation (Supplemental Material and Methods). Assuming the causal status vector of a SNP is independent from those of other SNPs ($\boldsymbol{C}_i \perp \boldsymbol{C}_j$ for $i \neq j$), the joint probability of the causal statuses of p SNPs is $\Pr\left(\boldsymbol{C}_1, \cdots, \boldsymbol{C}_p\right) = \prod_{i=1}^{p} \Pr(\boldsymbol{C}_i)$.

Given two genome-wide association studies with sample sizes n_1 and n_2 for the first and second populations, respectively, we derive the distribution of Z-scores, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ (both are $p \times 1$ vectors), conditional on the causal status vectors for each population, $\boldsymbol{c}_1 = \left(c_{11}, \cdots, c_{p1}\right)^T$ and $\boldsymbol{c}_2 = \left(c_{12} \cdots c_{p2}\right)^T$. Although it is reasonable to suspect that there are nonzero cross-population correlations of effect sizes at shared causal SNPs, to facilitate inference, we impose the (potentially strong) assumption that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are independent given $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Thus, for population j,

$$\boldsymbol{Z}_j |\ \boldsymbol{c}_j \sim MVN(\boldsymbol{0}, \boldsymbol{V}_j + \sigma_j^2 \boldsymbol{V}_j \mathrm{diag}(\boldsymbol{c}_j) \boldsymbol{V}_j)$$

where $\boldsymbol{V}_j$ is the $p \times p$ LD matrix for population j; $\mathrm{diag}(\boldsymbol{c}_j)$ is a diagonal matrix in which the *k*-th diagonal element is 1 if $c_{kj} = 1$ and 0 if $c_{kj} = 0$; and $\sigma_j^2 = \frac{n_j h_{gj}^2}{|\boldsymbol{c}_j|}$, where h_{gj}^2 and $|\boldsymbol{c}_j|$ are the SNP-heritability of the trait and the number of causal SNPs, respectively, in population j (Supplemental Material and Methods).

Finally, we derive the joint probability of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ by integrating over all possible causal status vectors in the two populations:

$$Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}) = \sum_{\boldsymbol{c}_1} \sum_{\boldsymbol{c}_2} \left[\prod\nolimits_{i=1}^{p} Pr(\boldsymbol{C}_i = c_{i1} c_{i2}) \prod_{j=1}^{2} N\left(\boldsymbol{Z}_j; \boldsymbol{0}, \boldsymbol{V}_j + \sigma_j^2 \boldsymbol{V}_j diag(\boldsymbol{c}_j) \boldsymbol{V}_j\right) \right] \quad (1)$$

where $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$ is the vector of parameters of the MVB distribution. In practice, we partition the genome into approximately independent regions[123] and model the distribution of Z-scores at all regions as the product of the distribution of Z-scores in each region (Ch. 3.2.5 Material and Methods; Supplemental Material and Methods).

3.2.2 Genome-wide proportions of population-specific/shared causal SNPs

We use Expectation-Maximization (EM) coupled with Markov Chain Monte Carlo (MCMC) to maximize the likelihood function in Equation (3.1) over the MVB parameters $\boldsymbol{f}$. We initialize $\boldsymbol{f}$ to $\boldsymbol{f} = (\boldsymbol{0}, \boldsymbol{-3.9}, \boldsymbol{-3.9}, \boldsymbol{3.9})$ which corresponds to 2% of SNPs being causal in population 1, 2% being causal in population 2, and 2% being shared causals. In the expectation step, we approximate the surrogate function $\boldsymbol{Q}\left(\boldsymbol{f} | \boldsymbol{f}^{(t)}\right)$ using an efficient Gibbs sampler; in the maximization step, we maximize $\boldsymbol{Q}\left(\boldsymbol{f} | \boldsymbol{f}^{(t)}\right)$ using analytical formulae (Supplemental Material and Methods). From the estimated $\boldsymbol{f}$, denoted $\boldsymbol{f}^*$, we recover the proportions of population-specific and shared causal SNPs. For computational efficiency, we apply the EM algorithm to each chromosome in parallel and aggregate the chromosomal estimates to obtain estimates of the genome-wide proportions of population-specific/shared causal SNPs.

3.2.3 Per-SNP posterior probabilities of being causal in one or both populations

We estimate the posterior probability of each SNP to be causal in a single population (population-specific) or both populations (shared), using the estimated genome-wide proportions of population-specific and shared causal variants (obtained from $\boldsymbol{f}^*$) as prior probabilities in an empirical Bayes framework. Specifically, for each SNP i, we evaluate the posterior probabilities $\Pr(\boldsymbol{C}_i = 01|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$, $\Pr(\boldsymbol{C}_i = 10|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$, and $\Pr(\boldsymbol{C}_i = 11|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$. Since evaluating these probabilities requires integrating over the posterior probabilities of all $2^{(2p)}$ possible causal status configurations, we use a Gibbs sampler to efficiently approximate the posterior probabilities (Supplemental Material and Methods).

3.2.4 Estimating numbers of population-specific/shared causal SNPs in a region

We infer the posterior expected numbers of population-specific/shared causal SNPs in a region (e.g., an LD block or a chromosome) conditional on the Z-scores ($\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$) by summing, across all SNPs in the region, the per-SNP posterior probabilities of being causal in a single or both populations. For example, in a region with p SNPs, the posterior expected number of shared causal SNPs is $\mathrm{E}[q_{11}|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*] = \sum_{i=1}^{p} \mathrm{E}\left[1_{\{\boldsymbol{C}_i=11\}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*\right] = \sum_{i=1}^{p} \Pr(\boldsymbol{C}_i = 11|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$. Since SNPs in a region are highly correlated, invalidating the use of jackknife to estimate standard errors, we refrain from reporting standard errors of the posterior expected regional numbers of population-specific/shared causal SNPs.

3.2.5 Defining LD blocks that are approximately independent in two populations

For computational efficiency, PESCA assumes that, in both populations, a SNP in a given block is independent from all SNPS in all other blocks. This assumption requires defining blocks of SNPs that are approximately LD-independent in both populations. To this end, we first compute the "transethnic LD matrix" ($\boldsymbol{V}_{trans}$) from the East Asian- and European-ancestry LD matrices ($\boldsymbol{V}_{EAS}$ and $\boldsymbol{V}_{EUR}$) by setting each element in the transethnic LD matrix to the larger of the East Asian-specific and European-specific pairwise LD; i.e. $\boldsymbol{V}_{trans,ij} = \boldsymbol{V}_{EAS,ij}$ if $|\boldsymbol{V}_{EAS,ij}| > |\boldsymbol{V}_{EUR,ij}|$ and $\boldsymbol{V}_{trans,ij} = \boldsymbol{V}_{EUR,ij}$ if $|\boldsymbol{V}_{EUR,ij}| > |\boldsymbol{V}_{EAS,ij}|$. The resulting matrix $\boldsymbol{V}_{trans}$ is block diagonal due to shared recombination hotspots in both populations; in practice, we apply this procedure to each chromosome separately to obtain 22 chromosome-wide transethnic LD matrices. We then apply LDetect[123] to define LD blocks within the transethnic LD matrix. Applying this procedure using the 1000 Genomes Phase 3 reference panel[7] to create the transethnic LD matrix produces 1,368 LD blocks (average length of 2-Mb) that are approximately independent in individuals of East Asian and European ancestry.

3.2.6 Enrichment of causal SNPs in functional annotations

We define the enrichment of population-specific/shared causal SNPs in a functional annotation as the ratio between the posterior and prior expected numbers of population-specific/shared causal SNPs. Specifically, we estimate the enrichment of population-specific/shared causal SNPs in a functional annotation *k* relative to the genome-wide background as

$$\hat{\alpha}_{k,\boldsymbol{b}} = \frac{E\left[q_{k,\boldsymbol{b}} \middle| \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*\right]}{E\left[q_{k,\boldsymbol{b}} \middle| \boldsymbol{f}^*\right]} = \frac{\sum_{i \in \psi(k)} \Pr(\boldsymbol{C}_i = \boldsymbol{b} | \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)}{p_k \Pr(\boldsymbol{C}_i = \boldsymbol{b})}$$

where $\boldsymbol{b} \in \{01, 10, 11\}$, $q_{k,\boldsymbol{b}}$ is the number of population-specific ($\boldsymbol{b} = 01$ or $\boldsymbol{b} = 10$) or shared ($\boldsymbol{b} = 11$) causal variants, $\psi(k)$ is the set of SNPs in functional annotation k, and p_k is the number of SNPs in functional annotation k. The numerator, $E[q_{k,\boldsymbol{b}} | \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]$, and denominator, $E[q_{k,\boldsymbol{b}} | \boldsymbol{f}^*]$, represent the posterior (conditioned on Z-scores) and prior expected numbers of causal SNPs in functional annotation k, respectively. We estimate the standard error of $\hat{\alpha}_{k,\boldsymbol{b}}$ using block jackknife over 1,368 non-overlapping approximately LD-independent blocks across the entire genome. The resulting enrichment test statistics, $\frac{\hat{\alpha}_{k,\boldsymbol{b}} - 1}{\mathrm{SE}(\hat{\alpha}_{k,\boldsymbol{b}})}$, approximately follow a t-distribution with degrees of freedom equal to the number of blocks minus one. Since we are interested in identifying categories of SNPs that harbor more population-specific/shared causal SNPs than expected (i.e. enrichment > 1), we report *P*-values from a one-tailed t-test where the null hypothesis is enrichment ≤ 1.

We note that our definition of enrichment of causal SNPs is related to, but conceptually different from, enrichment of SNP-heritability. A positive enrichment of causal SNPs in a functional category indicates that, compared to the genome-wide background, there are more causal SNPs in that category than expected; a positive enrichment of SNP-heritability in a category indicates that the average per-SNP effect size in the category is larger than the genome-wide average per-SNP effect size[63].

3.2.7 Simulation framework

We used real chromosome 22 genotypes of 10,000 individuals of East Asian ancestry from CONVERGE[124,125] and 50,000 individuals of white British ancestry from the UK Biobank[11,126] to simulate causal effects and phenotypes. First, we used PLINK[87] (v1.9) to remove redundant SNPs in the 1000 Genomes Phase 3 reference panel such that there are no pairs of SNPs with $r_{ij}^2 > 0.95$ $(i \neq j)$. We also removed strand-ambiguous SNPs and SNPs with MAF < 1% in either reference panel, resulting in a total of *M*=8,599 SNPs on chromosome 22 to use in simulations.

Given genotypes at *M* SNPs for n_1 and n_2 individuals in populations 1 and 2, respectively, we assume the standard linear models $\boldsymbol{y_1} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1}$ (population 1) and $\boldsymbol{y_2} = \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_1}$ (population 2). We assume the phenotypes are standardized within each population such that $\mathrm{E}[\boldsymbol{y_1}] = \mathbf{0}$, $\mathrm{Var}[\boldsymbol{y_1}] = \mathbf{I}$ and $\mathrm{E}[\boldsymbol{y_2}] = \mathbf{0}$, $\mathrm{Var}[\boldsymbol{y_2}] = \mathbf{I}$. Given $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, the index sets of causal SNPs in each population, the effects at the *i*-th causal SNP in each population, β_{1i} and β_{2i}, are drawn from

$$\boldsymbol{\beta}_{1c_1}|\boldsymbol{c}_1 \sim N\left(\mathbf{0}, \frac{h_{g1}^2}{|\boldsymbol{c}_1|}\mathbf{I}_{c_1}\right), \qquad \boldsymbol{\beta}_{2c_2}|\boldsymbol{c}_2 \sim N\left(\mathbf{0}, \frac{h_{g2}^2}{|\boldsymbol{c}_2|}\mathbf{I}_{c_2}\right)$$

where $|\boldsymbol{c_1}| = \sum_{i=1}^{M} c_{i1}$ and $|\boldsymbol{c_2}| = \sum_{i=1}^{M} c_{i2}$ are the total numbers of causal SNPs in each population, h_{g1}^2 and h_{g2}^2 are the total SNP-heritabilities in each population, and $\mathrm{E}\left[\beta_{1i}\beta_{1j}\right] = \mathrm{Cov}\left[\beta_{1i}, \beta_{1j}\right] = 0$ and $\mathrm{E}\left[\beta_{2i}\beta_{2j}\right] = \mathrm{Cov}\left[\beta_{2i}, \beta_{2j}\right] = 0$ for SNPs $i \neq j$. The effects at non-causal SNPs are set to 0. The environmental effects for the *n*-th individual in each population are drawn i.i.d. from $\epsilon_{1n} \sim N\left(0, 1 - h_{g1}^2\right)$ and $\epsilon_{2n} \sim N(0, 1 - h_{g2}^2)$.

Finally, given the real genotypes and simulated phenotypes for each population, we compute Z-scores for all SNPs in population *k* as $\boldsymbol{Z_k} = \frac{1}{\sqrt{n_k}}\boldsymbol{y}_k^T\boldsymbol{X}_k$.

3.2.8 Application to 9 complex traits and diseases

We downloaded publicly available East Asian- and European-ancestry GWAS summary statistics for body mass index (BMI), mean corpuscular hemoglobin (MCH), mean corpuscular volume

(MCV), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), triglycerides (TG), major depressive disorder (MDD), and rheumatoid arthritis (RA) from various sources (Table 3.1). The European-ancestry BMI GWAS is doubly corrected for genomic inflation factor[127], which induces downward-bias in the estimated SNP-heritability; we correct this bias by re-inflating the Z-scores for this GWAS by a factor of 1.24. For all traits, we restrict to SNPs with MAF > 5% in both populations to reduce noise in the LD matrices estimated from 1000 Genomes. We use PLINK (v.19) to remove redundant SNPs such that $\hat{r}_{ij}^2 < 0.95$ for all SNPs $i \neq j$ in both ancestry-matched 1000 Genomes reference panels. The resulting numbers of SNPs that were analyzed for each trait are listed in Table 3.1.

For each trait, we test for enrichment of population-specific/shared causal SNPs in 53 publicly available tissue-specific gene annotations[120], each of which represents a set of genes that are "specifically expressed" in a GTEx[128] tissue (referred to as "SEG annotations"). We set the threshold for statistical significance to P-value < 0.05/53 (Bonferroni correction for the number of tests performed per trait).

3.3 Results

3.3.1 Performance of PESCA in simulations

We assessed the performance of PESCA in simulations starting from real genotypes of individuals with East Asian (EAS) or European (EUR) ancestry (N_{EAS} = 10K, N_{EUR} = 50K, M = 8,599 SNPs) (Ch. 3.2.7 Material and Methods). First, we find that when in-sample LD from the GWAS is available, PESCA yields approximately unbiased estimates of the numbers of population-specific/shared causal SNPs (Figure 3.2, top panel). For example, in simulations where we randomly selected 50 EAS-specific, 50 EUR-specific, and 50 shared causal SNPs, we obtained estimates (and corresponding standard errors) of 37.8 (4.5) EAS-specific, 40.3 (4.9) EUR-specific, and 64.9 (6.3) shared causal SNPs, respectively. When external reference LD is used (in this case, from 1000 Genomes), PESCA yields a slight upward bias (Figure 3.2, bottom panel); on the same simulated data, we obtained estimates of 48.0 (5.9) EAS-specific, 53.7 (7.44) EUR-specific, and 78.8 (7.6) shared causal SNPs.

We observe a slight decrease in accuracy as the effective sample size, the product of SNP-heritability and sample size ($N \times h_g^2$), decreases (Figures S1-S5). This is expected as the likelihood of the GWAS summary statistics is a function of $N \times h_g^2$ (Ch. 3.2.1 Material and Methods) – as the expected per-SNP variance at causal SNPs ($N \times h_g^2$ divided by the number of causal SNPs) decreases, GWAS summary statistics provide less information on the causal status of each SNP. Since it is often the case that the sample size of one GWAS is larger than that of the other, we perform simulations in which SNP-heritability is fixed to 0.05 in both populations, the EAS sample size is fixed to $N_{EAS} = 10^4$, and the EUR sample size is varied such that the effective sample size of the EUR GWAS is 1-5x larger than that of the EAS GWAS. We find that the genome-wide estimators are relatively robust with in-sample LD; with external estimates of LD, when effective sample size differs by a factor of 2 or more, the estimator for the number of EUR-specific causal SNPs becomes less biased while the EAS-specific and shared causal estimators become increasingly inflated (Figure S6). In addition, while it seems likely that the effect sizes of shared causal SNPs would be positively correlated across populations, the PESCA model assumes zero cross-population correlation in order to facilitate inference (Ch. 3.2.1 Material and Methods). We therefore perform simulations under an alternative model in which EAS and EUR effect sizes at shared causal SNPs are positively correlated and find that our estimates of the genome-wide numbers of shared and population-specific causal SNPs become increasingly inflated and deflated, respectively, as the correlation increases from 0 to 1 (Figure S7).

Next, we use the estimated genome-wide proportions of population-specific/shared causal SNPs to evaluate per-SNP posterior probabilities of being causal in a single population (EAS only or EUR only) or in both populations (Ch. 3.2.3 Material and Methods). For each of the three causal configurations of interest (EAS only, EUR only, and shared), we observe an increase in the average correlation between the per-SNP posterior probabilities and the true causal status vector for that configuration as $N \times h_g^2$ increases and as the total number of causal SNPs decreases (i.e. as per-SNP causal effect sizes increase) (Figures S8-S9). As expected, as the simulated proportion of shared causal SNPs increases, the average correlation between the posterior probabilities and true causal status vectors increases for the shared causal configuration and decreases for the population-specific causal configurations (Figures S8-S9). Since we did not have access to individual-level genotypes sampled from an ancestral group with shorter LD blocks (e.g., African-ancestry individuals), we use the EAS and EUR LD scores of each SNP as proxies for the strength of LD in the region housing the SNP to investigate the impact of population-specific LD patterns on the per-SNP posterior probabilities. Among the true causal SNPs (shared or population-specific), the posterior probabilities are relatively invariant to the magnitude of the EAS and EUR LD scores (Figure S10). In other words, under the PESCA framework, power to detect a given true causal SNP does not depend on its LD score in either population. Restricting to a set of "high-posterior SNPs" (defined here as SNPs with posterior probability greater than some threshold t), we investigate whether PESCA systematically misclassifies SNPs based on the magnitude of their LD scores. Again, we observe that the average EAS and EUR LD scores do not vary significantly between the true and false positive classifications (Table S1). We then assessed whether our proposed statistics for testing for enrichment of population-specific/shared causal SNPs in functional annotations (Ch. 3.2.6 Material and Methods) are well-calibrated under the null hypothesis of no enrichment. Overall, when both population-specific and shared causal SNPs are drawn at random, the enrichment test statistics are conservative at different levels of polygenicity and GWAS power ($N \times h_g^2$), irrespective of whether in-sample LD or external reference LD is used (Figures S11-S16).

Finally, we evaluated the computational efficiency of each stage of inference. In the first stage of inference – estimating genome-wide proportions of population-specific/shared causal SNPs – the maximization step of the EM algorithm uses Gibbs sampling to efficiently sample from the posterior of the causal status vectors (Supplemental Material and Methods). We set both the number of burn-in iterations and the number of samples to 5,000 for the MCMC within the maximization step and found that the overall EM typically converged within 200 iterations (Figures S17-S19). Run-time per EM-iteration increases with the number of causal SNPs (Figure S20); for example, in simulations with a total of 8,589 SNPs, when the maximum number of EM iterations was set to 200, PESCA took an average of 90 minutes to obtain estimates in simulations with 20 randomly selected causal variants and 360 minutes in simulations with 100 randomly selected causal SNPs. This is expected because the likelihood function being maximized is proportional to the Bayes factor of only the causal SNPs (Supplemental Material and Methods). In the second stage of inference – evaluating posterior probabilities for each SNP – we set both the number of burn-in iterations and the number of samples to 5,000 for the MCMC and, to ensure stable estimates of the posterior probability, we report the average posterior probability from 20 iterations of the Gibbs sampling procedure. The average run-time was 5 minutes in simulations with 20 causal variants and 28 minutes in simulations with 100 causal variants (Figure S20). We note that both stages of inference can be parallelized to decrease run time.

3.3.2 Genome-wide proportions of shared causal SNPs for 9 complex traits

We obtained publicly available GWAS summary statistics for 9 (non-independent) complex traits and diseases in individuals of EAS and EUR ancestry (average N_{EAS} = 94,621, N_{EUR} = 103,507) (Table 3.1) and applied PESCA to estimate the genome-wide proportions of population-specific/shared common causal SNPs (Ch. 3.2.8 Material and Methods). To ensure convergence, we applied 750 EM iterations for each trait (Figures S21-S23). Across the 9 traits, the estimated proportions of common causal SNPs in each population (the sum of the numbers of population-specific and shared causal SNPs) are consistent with previously reported estimates of polygenicity in single populations[54,71,74,129,130]. For example, we estimate that approximately 10% of common SNPs have nonzero effects on BMI in both EAS and EUR and that 2-3% have nonzero effects on the lipids traits (Table 3.1). The low estimates for major depressive disorder and rheumatoid arthritis may be explained in part by their small GWAS sample sizes. While there is heterogeneity in the estimated proportions of shared causal SNPs across the 9 traits, we find that most common causal SNPs are shared between the populations, consistent with findings from previous studies[42]. For example, for BMI, we estimate that approximately 96% of common causal SNPs in each population are also causal in the other; for total cholesterol (TC), we estimate that 73% of common causal SNPs in EAS and 77% of those in EUR are shared by both populations (Table 3.1).

3.3.3 High-posterior SNPs are distributed nearly uniformly across the genome

We define 1,368 regions that are approximately LD-independent in both populations and estimate the posterior expected numbers of population-specific/shared causal SNPs in each region (Ch. 3.2.5 Material and Methods). For all 9 traits, high-posterior SNPs for both the population-specific and shared causal configurations are spread nearly uniformly across the genome (Figure 3.3, Figures S24-S31). For example, mean corpuscular hemoglobin (MCH) harbored, on average, 0.68 (S.D. 0.42) EAS-specific, 0.53 (S.D. 0.40) EUR-specific, and 2.19 (S.D. 1.46) shared high-posterior SNPs per region (Figure 3.3, Figure S29). Aggregating posterior probabilities by chromosome, we find that the posterior expected numbers of EAS-specific, EUR-specific, and shared causal SNPs per chromosome are highly correlated with chromosome length (Figures S32-S34), recapitulating previous findings based on regional SNP-heritability[61,74].

3.3.4 Distributions of high-posterior SNPs across GWAS risk regions

We aggregate per-SNP posterior probabilities within GWAS risk regions that are EAS-specific, EUR-specific, or shared by both populations and find that most GWAS risk regions harbor two or more shared high-posterior SNPs (Figure 3.4, Figures S35-S39), concordant with previous findings on allelic heterogeneity of complex traits[131]. On average across the 9 traits, we observe a 2.8x enrichment of shared high-posterior SNPs in population-specific GWAS risk regions relative to the genome-wide background. For example, for mean corpuscular hemoglobin (MCH), the EAS-specific and EUR-specific GWAS risk regions harbor an average of 3.0 (S.D. 1.7) and 3.3 (S.D. 1.5) shared high-posterior SNPs per region, respectively, whereas the average number of shared high-posterior SNPs per region across all regions is 2.0 (S.D. 1.3) (Figure 3.4). While BMI, the blood traits (MCH and MCV), and rheumatoid arthritis have similar numbers of EAS-specific and EUR-specific high-posterior SNPs in their population-specific GWAS risk regions, the lipids traits (HDL, LDL, total cholesterol and triglycerides) have significantly more EAS-specific high-posterior SNPs in all GWAS risk regions (Figure 3.4, Figures S35-S39).

For each causal configuration (EAS-specific, EUR-specific, or shared), we examine the effect sizes of high-posterior SNPs (posterior probability > 0.8) in EAS and EUR (Figure 3.5). Across the 9 traits, the majority of EAS-specific high-posterior SNPs are nominally significant ($p_{GWAS} < 5 \times 10^{-6}$) either in the EAS GWAS only or in both GWASs. While five EUR-specific high-posterior SNPs are nominally significant in only the EAS GWAS, the majority are nominally significant either in the EUR GWAS only or in both GWASs. We observe strong correlations between the effect sizes in EAS and EUR for all three sets of high-posterior SNPs (Pearson r^2 of 0.79 [EAS-specific], 0.73 [EUR-specific], and 0.80 [shared]) that are driven by SNPs that are nominally significant in both GWASs (Figure 3.5). Taken together, these results suggest that most population-specific GWAS risk regions harbor shared causal variants that are undetected in the other population due to heterogeneity in LD structures, allele frequencies, and/or GWAS sample sizes.

3.3.5 Enrichment of high-posterior SNPs in genes that are specifically expressed in trait-relevant tissues

Motivated by recent work that found enrichment of SNP-heritability in regions near genes that are "specifically expressed" in trait-relevant tissues and cell types (referred to as "SEG annotations"), we tested for enrichments of population-specific and shared causal SNPs in the same 53 tissue-specific SEG annotations[120]. For a given causal configuration, the enrichment of causal SNPs in an annotation is defined as the ratio between the posterior and prior expected numbers of causal SNPs in the annotation (Ch. 3.2.6 Methods). For 8 of the 9 traits, we find significant enrichment of shared high-posterior SNPs in at least one SEG annotation (P-value < 0.05/53 to correct for 53 tests per trait) (Figures S40-S44). All SEG annotations with significant enrichments of population-specific high-posterior SNPs are also enriched with shared high-posterior SNPs for the same trait, providing additional evidence that many signatures of population-specific genetic architecture are induced by population-specific LD and allele frequencies rather than distinct genetic etiologies. We do not find enrichment of any high-posterior SNPs in any SEG annotation for major depressive disorder (MDD) (Figure S44), which could be due to low GWAS sample sizes (Table 3.1). Finally, for each SEG annotation, we obtain a meta-analyzed transethnic SNP-heritability enrichment by computing the inverse-variance weighted average of the EAS and EUR SNP-heritability enrichments (which are obtained separately using stratified LD score regression[63,64]). We observe a strong correlation between the meta-analyzed SNP-heritability enrichments and the enrichments of shared high-posterior SNPs (Figure 3.6), suggesting that SNP-heritability enrichments are largely driven by many low-effect SNPs rather than a small number of high-effect SNPs.

3.4 Discussion

We have presented PESCA, a method for estimating the genome-wide proportions of SNPs with nonzero effects in a single population (population-specific) or in two populations (shared) from GWAS summary statistics and estimates of LD. We applied PESCA to EAS and EUR GWAS summary statistics for 9 complex traits and find that, while the lipids traits have significantly more

EAS-specific common causal SNPs compared to the remaining traits, the majority of common causal SNPs are shared by both populations. Regions that harbor statistically significant GWAS associations for one population are enriched with SNPs with high-posterior probability of being causal in both populations; moreover, high-posterior SNPs (posterior probability > 0.8 for any causal configuration) have highly correlated effect sizes in EAS and EUR, recapitulating results of previous studies[42]. For all traits except MDD, we identify tissue-specific SEG annotations[120] enriched with shared high-posterior SNPs and observe that all SEG annotations enriched with population-specific high-posterior SNPs are a subset of those enriched with shared high-posterior SNPs. Taken together, our results indicate that most population-specific GWAS risk regions contain shared common causal SNPs that are undetected in the second population due to differences in LD or allele frequencies. This suggests that localizing shared components of genetic architecture and explicitly correcting for population-specific LD and allele frequencies may help improve transferability of results from well-powered European-ancestry studies to other understudied populations. Based on the simulation results in Figure S1 (in which 100% of causal SNPs are shared) and our estimates of SNP-heritability for the traits in Table 3.1, we recommend applying PESCA to summary statistics for which the *effective per-SNP sample size*, $N \times h_g^2$ divided by the number of causal SNPs, is at least 3 for both GWASs. For a typical quantitative trait (e.g., Table 3.1), this corresponds to a total effective sample size of approximately $N \times h_g^2 > 10{,}000$.

We conclude by discussing the caveats and limitations of our analyses. First, the estimated proportions of causal SNPs must be interpreted with caution as they can be influenced by gene-environment interactions. For example, if a SNP has a nonzero effect on a trait only in the presence of environmental factors that are specific to EAS-ancestry individuals, PESCA will interpret that

SNP as an EAS-specific causal SNP even though it would have a nonzero effect in Europeans in the presence of the same environmental factors.

Second, we chose to analyze a set of traits that were present in both the UK Biobank and Biobank Japan and for which GWAS summary statistics were publicly available. Since most publicly available summary statistics of large-scale GWAS are meta-analyses of smaller studies, in-sample LD is often unavailable. While PESCA with in-sample LD is relatively robust to differential GWAS power, with external LD, performance decreases when the GWAS effective sample sizes differ by more than a factor of 2x. We note, however, that for the real traits analyzed in this work, effective sample size differs by a maximum factor of 2x (mean corpuscular hemoglobin; Table 3.1). Additionally, PESCA currently cannot be applied to admixed populations if in-sample LD is unavailable. An extension of PESCA to properly account for external/noisy estimates of LD would thus increase its utility; we defer a thorough investigation of this to future work. In parallel, in light of ongoing efforts at several institutions to establish biobanks[13,14,17,132,133], we believe that well-powered GWASs (with in-sample LD) will become increasingly available for diverse and admixed populations. Another challenge is that many publicly available summary statistics were computed from fixed-effect meta-analyses or linear mixed models. Since the PESCA model is defined with respect to GWAS marginal effects estimated by ordinary least squares (OLS) regression, it is unclear whether PESCA is sensitive to non-OLS association statistics, which have different statistical properties; we defer a thorough investigation of this to future work.

Third, we restricted our analyses to SNPs with MAF > 5% in both populations to reduce noise in the LD matrices estimated from external reference panels. Consequently, the estimates we report in this work do not capture effects of low frequency or rare variants that are not well-tagged by common SNPs. Furthermore, since most common variants are shared across continental populations and rarer variants tend to localize among closely related populations[7,11], our study design undersamples population-specific causal variants. We note, however, that lower MAF thresholds can be used if in-sample LD is available. We also note that for the purpose of improving transferability of polygenic risk scores (PRS) across populations, prediction accuracy depends largely on the accuracy of the PRS weights at common SNPs (the average per-SNP contribution to total SNP-heritability is larger for common SNPs than for low frequency or rare variants[91]).

Finally, PESCA can be sensitive to model misspecification. For computational efficiency, PESCA relies on having regions that are approximately LD-independent in both populations; if there is LD leakage between regions, the estimated proportions of causal SNPs will be biased. We therefore recommend defining LD blocks for each pair of populations one analyzes. Similarly, to facilitate inference, PESCA does not explicitly model cross-population correlations of effect sizes at shared causal variants; we conjecture that modeling these correlations can further improve performance.

3.5 Supplemental Data

Supplemental Data include 44 figures, 1 table, and Supplemental Material and Methods. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273527/.

3.6 Figures

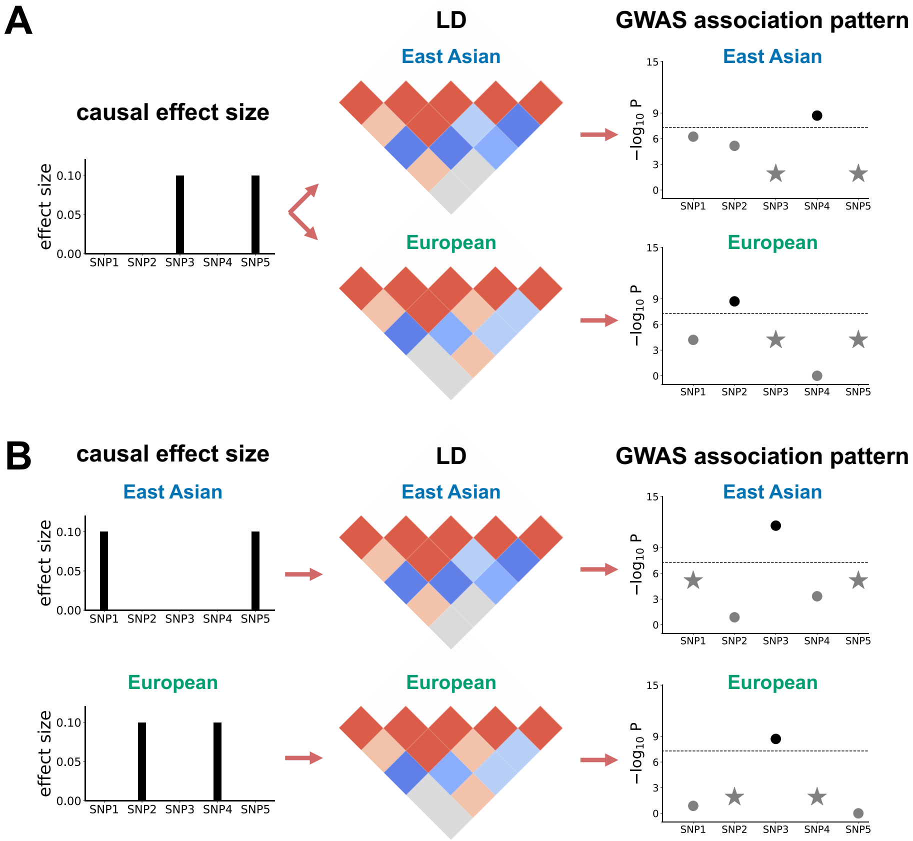

Figure 3.1 Toy examples illustrating impact of population-specific LD on GWAS associations. A) SNPs 3 and 5 are causal in both East Asians and Europeans and have the same population-specific causal effect size of 0.1. However, due to different LD patterns in East Asians and Europeans, SNPs 2 and 4 are observed to be GWAS-significant, respectively. B) Different SNPs are causal in East Asians (SNPs 1 and 5) and Europeans (SNPs 2 and 4). However, due to population-specific LD, SNP 3 is observed to be GWAS-significant in both populations. The stars in the rightmost plots represent the SNPs with true nonzero effects; the GWAS-significant SNP is highlighted in a darker color.

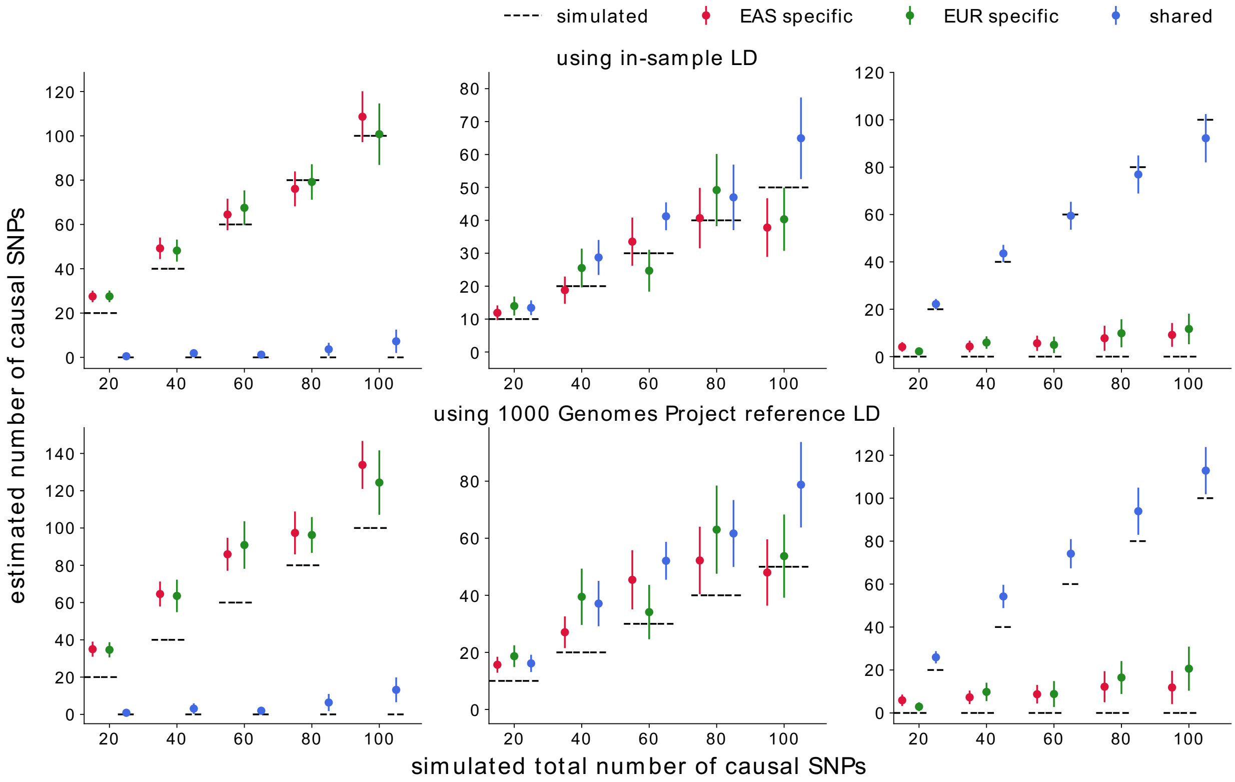

Figure 3.2 Estimates of the numbers of population-specific/shared causal SNPs in simulations. The estimates are approximately unbiased when in-sample LD is used (top panel) and upward-biased estimates when external reference LD is used (bottom panel). For both populations, we simulate such that the product of SNP-heritability and GWAS sample size is 500. Mean and standard errors were obtained from 25 independent simulations. Error bars represent ±1.96 of the standard error.

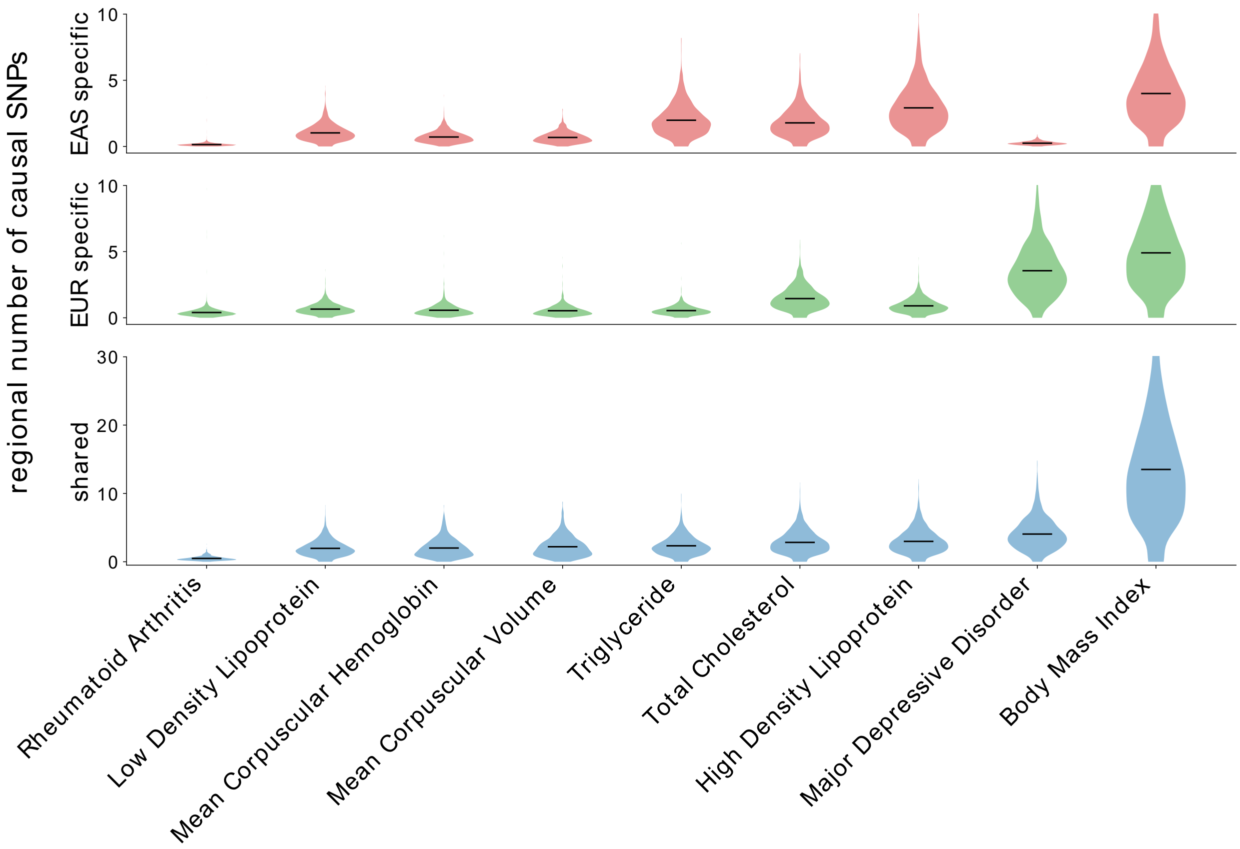

Figure 3.3 Number of causal SNPs shared between EAS and EUR across 1,368 LD blocks. Estimates of the numbers of population-specific/shared causal SNPs across LD blocks that are approximately independent in both EAS and EUR. Each violin plot represents the distribution of the posterior expected number of population-specific or shared causal SNPs per region; details on how the regions were defined can be found in the Methods. For a single region, the posterior expected number of SNPs in a given causal configuration is estimated by summing, across all SNPs in the region, the per-SNP posterior probabilities of having that causal configuration (Material and Methods). The dark lines mark the means of the distributions. The traits are sorted on the x-axis by the average number of shared high-posterior SNPs per region.

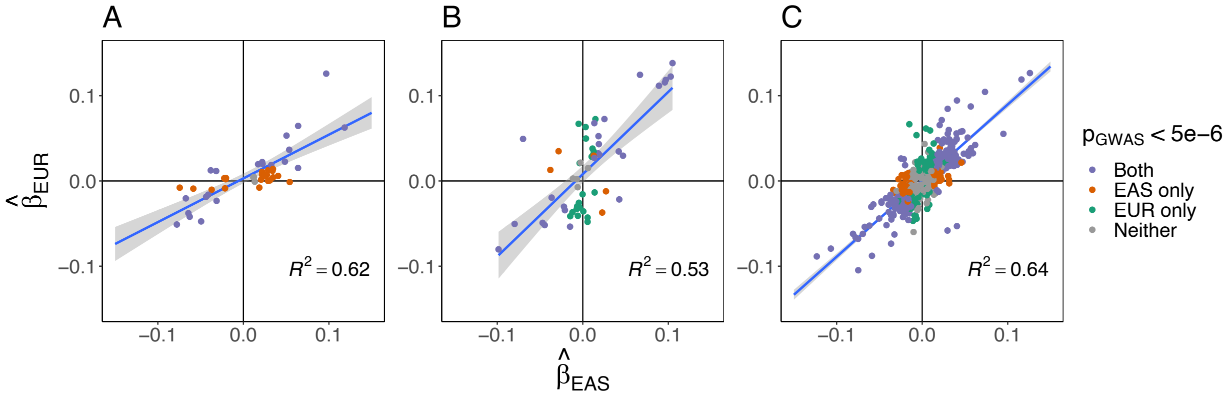

Figure 3.4 Marginal regression coefficients of high-posterior SNPs for 9 complex traits.

Each plot corresponds to one of the three causal configurations of interest: EAS-specific (A), EUR-specific (B), and shared (C). Each point represents a SNP with posterior probability > 0.8 for a single trait. The x-axis and y-axis mark the marginal regression coefficients in the EAS-ancestry GWAS and EUR-ancestry GWAS, respectively. The colors indicate whether the SNP is nominally significant ($p_{GWAS} < 5 \times 10^{-6}$) in both GWASs (purple), the EAS GWAS only (orange), the EUR GWAS only (green), or in neither GWAS (gray). The gray band marks the 95% confidence interval of the regression line.

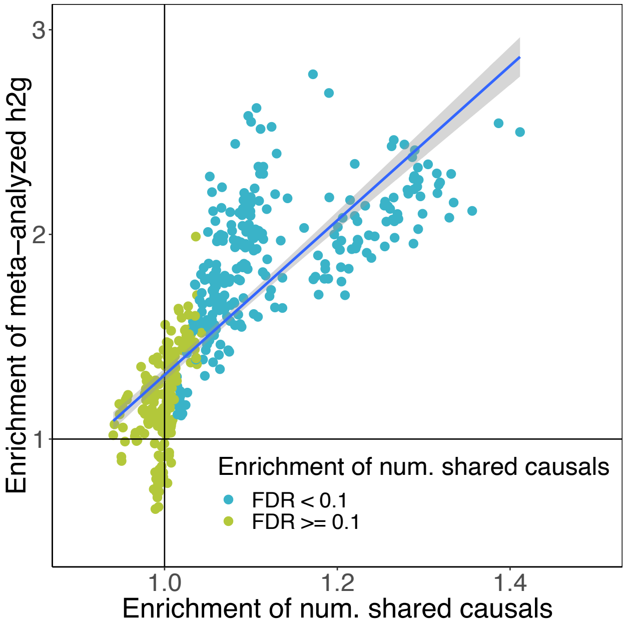

Figure 3.5 Enrichment of shared high-posterior SNPs is highly correlated with h2g enrichments. Each point represents a trait-tissue pair; each tissue-specific functional category represents a set of genes that are "specifically expressed" in one of 53 GTEx tissues (53 SEG annotations). The x-axis is the enrichment of shared high-posterior SNPs in the SEG annotation obtained from PESCA. The y-axis is the meta-analyzed transethnic SNP-heritability explained by the SEG annotation, defined as the inverse-variance weighted average of the EAS and EUR SNP-heritability enrichments (obtained separately using stratified LD score regression). The points are colored by whether the trait has a statistically significant enrichment of shared high-posterior SNPs in the corresponding SEG annotation (FDR < 0.1). Enrichment estimates and standard errors for each trait-tissue pair can be found in Figures S40-S44.

3.7 Tables

Trait name (abbrev.)	Pop.	Ref.	$\hat{h}_g^2$ (S.E.) %	Sample size (N)	Total # SNPs (MAF > 5%)
Body Mass Index (BMI)	EAS	22	19.8 (0.64)	224,698	258,130
	EUR	72	20.6 (0.91)	158,284	
Mean Corpuscular Hemoglobin (MCH)	EAS	21	18.6 (2.2)	108,054	480,684
	EUR	82	22.7 (3.2)	172,332	
Mean Corpuscular Volume (MCV)	EAS	21	21.0 (2.13)	108,256	480,678
	EUR	82	23.6 (3.1)	172,433	
High Density Lipoprotein (HDL)	EAS	21	20.7 (3.03)	70,657	268,198
	EUR	83	16.4 (2.2)	89,614	
Low Density Lipoprotein (LDL)	EAS	21	9.5 (1.3)	72,866	268,201
	EUR	83	13.6 (1.93)	85,491	
Total Cholesterol (TC)	EAS	21	8.1 (0.84)	128,305	268,197
	EUR	83	22.5 (2.1)	89,865	
Triglyceride (TG)	EAS	21	13.5 (3.3)	105,597	268,198
	EUR	83	13.6 (2.2)	86,502	
Major Depressive Disorder (MDD)	EAS	67	35.6 (3.4)	10,640	389,593
	EUR	84	19.0 (1.8)	18,759	
Rheumatoid Arthritis (RA)	EAS	36	28.9 (18.3)	22,515	526,206
	EUR	36	9.5 (1.9)	58,284	

Table 3.1 List of 9 complex traits analyzed in EAS and EUR.

We estimated genome-wide SNP-heritability using LD score regression with the intercept constrained to 1 (i.e. assuming no population stratification).

Trait	EAS-specific causals (S.E.)	EUR-specific causals (S.E.)	Shared causals (S.E.)	$\hat{\rho}_g$ (S.E.)[51]
BMI	982 (2) 0.4%	1,033 (2) 0.4%	25,641 (16) 10%	0.80 (0.02)
MCH	1,165 (6) 0.2%	728 (3) 0.2%	3,082 (4) 0.6%	0.88 (0.05)
MCV	1004 (4) 0.2%	737 (5) 0.2%	3,256 (8) 0.7%	0.89 (0.05)
HDL	3,167 (12) 1%	652 (2) 0.2%	4,789 (9) 2%	0.89 (0.06)
LDL	969 (5) 0.4%	742 (2) 0.3%	3,129 (6) 1%	0.66 (0.11)
TC	1,892 (3) 0.7%	1,493 (5) 0.6%	5,058 (12) 2%	0.91 (0.07)
TG	2,245 (3) 0.8%	511 (4) 0.2%	3,432 (7) 1%	0.93 (0.07)
MDD	88 (4) 0.02%	3,280 (6) 0.84%	7,830 (6) 2%	0.34 (0.07)
RA	3 (0.3) 6e-04%	124 (2) 0.02%	1,080 (6) 0.2%	0.87 (0.10)

Table 3.2 Estimated numbers of population-specific/shared common causal SNPs for 9 traits. Trans-ethnic genetic correlation estimates ($\hat{\rho}_g$) computed from a similar set of summary statistics were obtained from a previous study[116]. Standard errors of the estimated numbers of population-specific/shared causal SNPs were computed using the last 50 iterations of the EM-MCMC algorithm.

4 Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes

4.1 Introduction

Since the vast majority of risk variants identified through genome-wide association studies (GWAS) are located in noncoding regions, the genes and pathways driving complex traits are largely unknown[1,26,134]. For most complex traits, fundamental characteristics of genetic architecture—for example, the number of variants/genes with nonzero effects (polygenicity), the number of genes regulated by local versus distal variants, and the relative contributions of rare versus common variants to gene expression and phenotype—remain actively debated[33,68,81,135–142].

That complex-trait SNP-heritability is enriched in regulatory regions is well established[26,63,143,144]. However, since SNP-heritability is overwhelmingly driven by common variants of low effect—individual rare variants with large per-allele effects contribute very little to population-level phenotypic variance[52,145]—whether the largest heritability enrichments localize the most clinically relevant regions and/or genes for a trait is unclear. For example, a recent study estimates that the majority of complex-trait SNP-heritability mediated via the *cis*-genetic component of expression is explained by genes that individually have low *cis*-heritability of

This chapter is published in Burch*, Hou*, et al. *bioRxiv* (2021) as a non-peer-reviewed preprint. Citations have been updated and renumbered. Supplemental material is available at https://doi.org/10.1101/2021.08.17.456722 expression[53]. In addition, despite the inherent complexity of the biological processes driving complex traits, there is growing evidence that extreme complex-trait polygenicity may be explained in large part by negative/stabilizing selection, which purges high-effect alleles from the population, producing the remarkably even distribution of SNP-heritability among common variants genome-wide (the so-called "flattening" hypothesis)[4,54]. If the most critical genes for a trait are not necessarily localized by enrichments of total heritability[55,146], the open question of how to identify target genes using heritability enrichments or overlaps between GWAS and expression quantitative trait loci[147,148] becomes even murkier. Gene-based association tests that aggregate signal from multiple rare variants—for example, burden tests and sequence-based association tests (SKAT)—can increase power under different genetic-architecture scenarios[149–158]. However, such methods are generally designed to test for only rare-variant association or the combined effects of common and rare variants, and thus are not ideal for parsing the relative contributions of rare/common variants to the heritability of a single gene.

Here, we propose an approach to estimate the relative heritability contributions of common, low-frequency, and rare variants to a quantity we call *gene-level heritability* (h^2_{gene}), defined as the proportion of phenotypic variance explained by the additive effects of a given set of variants assigned to a gene of interest. While the method itself is general and can be applied to any small annotation of interest (Ch. 4.3 Discussion), our goal in this work is to use MAF-partitioned gene-level heritability estimates to identify disease-relevant genes, which may have different relative contributions to heritability across MAF classes. The key challenge in estimating gene-level heritability lies in the *uncertainty* about which variants are causal and what their causal effect sizes are; such uncertainty in fine-mapping increases as the strength of LD in the region increases and as GWAS sample size decreases[159]. Consider a toy example in which a variant in the gene of interest is in perfect LD (LD=1) with a second variant adjacent to the gene, the observed data are GWAS marginal association statistics and LD for the region (Figure 4.1a). Without additional information, it is impossible to definitively elucidate the underlying causal configuration. Even if the LD between the variants is 0.9 instead of 1, if this GWAS has 90% power to identify the associated region, to correctly reject the null hypothesis for the non-causal variant would require a sample size $\geq$ 4x larger than that of the original GWAS[159]. Since each causal configuration can yield a different gene-level heritability (with or without MAF-partitioning), randomly selecting one possible configuration (e.g., using variable selection methods such as the Lasso[160]) can yield inaccurate/misleading estimates. As an alternative approach, methods for partitioning genome-wide SNP-heritability across MAF bins can be employed. However, such methods are also ill-suited to our goals as they make distributional assumptions on the causal effects which (i) limit power to detect enrichment in small categories of variants (< 1% of the genome) and/or (ii) may not apply equally to rare and common variants[34,60,61,63,83,144,161]. Estimators for the SNP-heritability of a single region ("regional SNP-heritability") yield inflated estimates if any variants in the region of interest are in LD with the adjacent regions[55,73,74,162]. To address the fine-mapping uncertainty, we seek to propagate the uncertainty about which variants are causal to infer the posterior distribution over the entire gene of interest. Given GWAS summary statistics and estimates of LD, we sample from the posterior distribution of the causal effect sizes within a probabilistic fine-mapping framework[163] and use the posterior samples to approximate the posterior distribution of gene-level heritability, thus capturing uncertainty in the causal effects (Figure 4.1b). From the full posterior distribution of gene-level heritability, one can compute various summary statistics of interest for each gene. We report the posterior mean, which we denote $\hat{h}^2_{\mathrm{gene}}$, and ρ-level credible intervals, or ρ-CI, defined as the central interval containing the true gene-level heritability with probability ρ (Ch. 4.4.3 Material and Methods).

We confirm in simulations that accounting for uncertainty in the estimated causal effects significantly reduces the bias of $\hat{h}^2_{\mathrm{gene}}$. Although the corresponding ρ-CIs are not perfectly calibrated—for example, at $\rho = 0.9$, about 70% of credible intervals overlap h^2_{gene}—among the true causal genes, any mis-calibrated CIs overwhelmingly tend to underestimate rather than overestimate h^2_{gene}. Both $\hat{h}^2_{\mathrm{gene}}$ and ρ-CIs are robust to parameters such as causal effect sizes, gene length, allele frequencies of causal variants, and the strength of local LD. Assuming that total gene-level heritability can be expressed as $h^2_{\mathrm{gene,t}} = h^2_{\mathrm{gene,r}} + h^2_{\mathrm{gene,lf}} + h^2_{\mathrm{gene,c}}$, where each term refers to the component of $h^2_{\mathrm{gene,t}}$ explained by rare ($0.5\% \leq \mathrm{MAF} < 1\%$), low-frequency ($1\% \leq \mathrm{MAF} < 5\%$), and common ($\mathrm{MAF} \geq 5\%$) variants, respectively, we apply the same approach to estimate the posterior distributions of $h^2_{\mathrm{gene,r}}$, $h^2_{\mathrm{gene,lf}}$, and $h^2_{\mathrm{gene,c}}$ and observe similar trends and levels of accuracy (we note that there are many definitions of "rare" in the literature, and that we use 0.5% ≤ MAF < 1% because we analyze imputed genotypes).

Applying our approach to estimate gene-level heritability for 17,436 genes and 25 quantitative traits in the UK Biobank[11] (N=290K self-reported "white British", MAF > 0.5%), we find that $h^2_{\mathrm{gene,t}}$ is indeed dominated by $h^2_{\mathrm{gene,c}}$. Among genes with $h^2_{\mathrm{gene,t}}$ 90%-CI > 0 ("nonzero-heritability genes") for a given trait, 92% (s.d. 1%) have nonzero common-variant heritability, and 76% (s.d. 1%) have nonzero heritability exclusively from common variants (i.e. $h^2_{\mathrm{gene,t}} \approx h^2_{\mathrm{gene,c}}$). In contrast, only 2.5% (s.d. 0.6%) of nonzero-heritability genes, averaged across traits, have nonzero rare-variant heritability, and 0.8% (s.d. 0.4%) have nonzero heritability exclusively from rare variants ($h^2_{\text{gene,t}} \approx h^2_{\text{gene,r}}$). As a sanity check, we confirm that Mendelian-disorder genes from OMIM[164], genes intolerant to loss of function (LoF) variants[165], and a set of FDA-approved drug targets for 30 immune-related traits[166] have elevated estimates of all four heritability quantities (total, common, low-frequency, and rare). Among the 0.8% with $h^2_{\text{gene,t}} \approx h^2_{\text{gene,r}}$ (370 gene-trait pairs in total), we identify many examples of disease genes with known roles in phenotypically similar Mendelian disorders and other congenital growth and developmental disorders. 37% of the 370 gene-trait pairs were not identified by existing methods for gene-level association testing, likely because existing methods have low power to detect genes containing only rare variants of moderate or low effect. We observe an overrepresentation of LoF-intolerant genes, but not Mendelian-disorder genes, among the $h^2_{\text{gene,t}} \approx h^2_{\text{gene,r}}$ genes. Using gene-level heritability estimates to further explore genetic architecture reveals notable differences between total/rare-variant gene-level heritability; for example, while total/common-variant gene-level heritability increases with gene length, we observe a clear inverse relationship between the rare-variant component and gene length.

Taken together, our results show that the low-frequency/rare-variant component of total gene-level heritability is useful for identifying narrow sets of high-impact genes that are not necessarily located in regions enriched with common-variant heritability. Our results are also consistent with the hypothesis that a sizable amount of complex-trait variation is driven by dysregulation of genes that—if completely disrupted—cause phenotypically similar monogenic disorders and/or systemic congenital and developmental disorders[56]. Since some high-impact genes are disrupted/dysregulated by a combination of common and rare variants, we conclude that

$h^2_{\text{gene,r}}$ should be considered alongside common-variant heritability enrichments if one is interested in identifying high-impact disease genes under different degrees of selection. While we restrict our analyses to genes (±10-kb window), our method is general and can thus be applied to any small annotation of interest (e.g., enhancers, a set of genes involved in a pathway, a set of putative causal variants).

4.2 Results

4.2.1 Overview of the Methods

We propose a general approach for estimating the heritability explained by a given set of variants and assess its utility in estimating gene-level heritability. Given an assignment of m variants to a gene g of interest, total gene-level heritability is defined as $h^2_{\text{gene,t}} \equiv \text{Var}\left[\mathbf{x}_{\text{g}}^T \boldsymbol{\beta}_{\text{g}} | \boldsymbol{\beta}\right] = \boldsymbol{\beta}_{\text{g}}^T \mathbf{R}_{\text{g}} \boldsymbol{\beta}_{\text{g}}$, where $\boldsymbol{\beta}_{\text{g}}$ is the $m \times 1$ vector of unknown causal effect sizes and $\mathbf{R}_{\text{g}}$ is the $m \times m$ LD for SNPs in the gene (Ch. 4.4.1 Material and Methods). Our goal in this work is to estimate a *distribution* over $h^2_{\text{gene,t}}$ that captures uncertainty in the causal effects that arises from LD (Figure 4.1a). To this end, we adopt a probabilistic fine-mapping framework[162,163] which assumes a sparse prior on the causal effect sizes in the LD block containing gene g and infers the posterior distribution of the causal effect sizes, $p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}})$, where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated marginal effects from GWAS and $\widehat{\mathbf{R}}$ is an estimate of LD. We sample from the posterior of $\boldsymbol{\beta}$ to approximate to the posterior of $h^2_{\text{gene,t}}$ (Figure 4.1b, Ch. 4.4.2 Material and Methods). For each gene, we report the estimated posterior mean, denoted $\widehat{h}^2_{\text{gene,t}}$, and ρ-level credible intervals (ρ-CI), defined as the central interval that contains the true gene-level heritability with probability $\rho \in [0,1]$. Whereas previous works applied similar approaches to generate credible sets of causal variants[163] or to estimate regional SNP-heritability of LD blocks[162], our goal in this work is to estimate the heritability explained by any arbitrary (not necessarily contiguous) set of variants much smaller than an LD block. This allows us to partition by minor allele frequency (MAF) bins under the assumption that $h^2_{\text{gene,t}} = h^2_{\text{gene,r}} + h^2_{\text{gene,lf}} + h^2_{\text{gene,c}}$, where the subscripts represent the rare ($0.5\% \leq$ MAF $< 1\%$), low-frequency ($1\% \leq$ MAF $< 5\%$), and common (MAF $\geq$ 5%) variants assigned to the gene. (We note that, while there are many definitions of "rare" in the literature, we threshold at MAF ≥ 0.5% because we want to reduce potential noise from imputation; see Ch. 4.3 Discussion for details.)

4.2.2 Accuracy of gene-level heritability estimates in simulations

We perform simulations starting from real imputed genotypes of N=290,273 "unrelated white British" individuals in the UK Biobank (chromosome 1, MAF $> 0.5\%$, M=200,235 variants, 1,083 genes; Ch. 4.4.5 Material and Methods). In all simulations, the estimand of interest (gene-level heritability, h^2_{gene}) is the proportion of phenotypic variance explained by the variants in the gene body, as well as the MAF-partitioned counterpart. We note that our choice of variant assignment is arbitrary; there are many ways to assign variants to a gene, but our goal in this section is to provide a proof of concept. In brief, our simulation framework consists of three steps. First, for a given total heritability (variance explained by all *M* variants) and cumulative gene-level heritability (variance explained by all genes), we randomly select 3%, 8%, or 16% of the genes to be causal, where "causal" in this context refers to genes with $h^2_{\text{gene,t}} > 0$. Second, for each causal gene, we draw causal variants in the gene body and within 10-kb upstream/downstream of the gene start/end positions; the purpose of the latter is to create situations where the estimated effects of variants in the region of interest are inflated in part because they tag causal effects located adjacent to the region. Third, we sample noncoding "background" causal variants from the whole chromosome with frequency $p_{\text{causal}} = \{0.001, 0.01\}$. Under this model, the majority of simulated gene-level heritabilities are on the order of 10^{-6} to 10^{-3} (Supplementary Figure 1), similar to what we observe in real data in subsequent sections.

Overall, the estimated posterior means of total gene-level heritability, $\hat{h}^2_{\text{gene,t}}$, are highly concordant with the true gene-level heritabilities (Figure 4.2, Supplementary Figure 2). For each gene, we compute two metrics of accuracy from $s = 30$ simulation replicates: $\text{bias}\left[\hat{h}^2_{\text{gene,t}}\right] \approx 1/30 \sum_s (\hat{h}^2_{\text{gene,t(s)}} - h^2_{\text{gene,t}})$, and $\text{MSE}\left[\hat{h}^2_{\text{gene,t}}\right] = \left(\text{bias}\left[\hat{h}^2_{\text{gene,t}}\right]\right)^2 + \text{Var}\left[\hat{h}^2_{\text{gene,t}}\right]$ (mean squared error) (Ch. 4.4.6 Material and Methods). As expected, MSE increases as the background polygenicity (p_{causal}) and proportion of causal genes increase, i.e. as causal effect sizes of noncoding variants and gene-level heritabilities decrease (Supplementary Figure 3). Among the causal genes ($h^2_{\text{gene,t}} > 0$), $\hat{h}^2_{\text{gene,t}}$ tends to underestimate $h^2_{\text{gene,t}}$, with the median bias across genes ranging from approximately $-4\% \times h^2_{\text{gene,t}}$ for lower polygenicities to $-30\% \times h^2_{\text{gene,t}}$ for higher polygenicities (Figure 4.2, Supplementary Figure 4). There is a small positive correlation between bias and gene length (average Pearson R = 0.05 (s.d. 0.02) across simulation setups), i.e. the estimates tend to be more downward-biased for shorter genes; average LD score and average MAF of variants in the gene have no discernible impact on accuracy (Supplementary Figures 5-8). To visualize the impact of causal-effect uncertainty on gene-level heritability estimation, we compare

$\hat{h}^2_{\text{gene,t}}$ to a naive estimator that ignores LD between the gene and its adjacent regions, thus ignoring causal-effect uncertainty (Ch. 4.4.7 Material and Methods). As expected, the naive estimator is significantly inflated; in particular, many noncausal genes have dramatically upward-biased estimates (Figure 4.2, Supplementary Figures 2 and 9) due to LD between variants in the gene and nearby causal variants. We benchmark the estimators for the contributions of rare, low-frequency, and common variants to total gene-level heritability and find that they perform similarly to $\hat{h}^2_{\text{gene,t}}$ (Figure 4.3, Supplementary Figures 3, 4, 6-8, 10-12).

4.2.3 Calibration of credible intervals

Calibration of ρ-level credible intervals (ρ-CIs) was assessed using "empirical coverage," defined here as the proportion of simulation replicates in which ρ-CI contains the true gene-level heritability (Ch. 4.4.3 Material and Methods). Perfect calibration of ρ-CI would manifest as empirical coverage equal to ρ for all $\rho \in [0,1]$. In reality, we observe a downward bias in empirical coverage across all simulations that increases in magnitude as the proportion of causal genes increases (i.e. as per-variant causal effect sizes decrease). For example, at $\rho = 0.9$, empirical coverage ranges from approximately 0.75 when 3% of genes are causal to 0.65 when 16% are causal (Supplementary Figure 13). While downward bias in empirical coverage can be the result of ρ-CIs underestimating or overestimating $h^2_{\text{gene,t}}$, the credible intervals at $\rho = \{0.90, 0.95\}$ tend to underestimate the true gene-level heritability (Supplementary Table 1), consistent with the downward-bias we observe in $\hat{h}^2_{\text{gene,t}}$ (Figure 4.2). For example, at $\rho = 0.95$, the proportion of true causal genes that are underestimated vs. overestimated is approximately 14% vs. 6% (when

3% of genes are causal) and 30% vs. 3.5% (when 16% of genes are causal) (Supplementary Table 1). The ρ-CIs for $h^2_{\text{gene,r}}$ are more conservative; for the same parameters, among the genes with true $h^2_{\text{gene,r}} > 0$, the proportions of underestimated vs overestimated genes are 38% vs. 1.5% (when 3% of genes are causal) and 45% vs. <1% (when 16% of genes are causal) (Supplementary Table 2, Supplementary Figure 14).

4.2.4 Robustness to noise in estimates of LD

Finally, we assess whether $\hat{h}^2_{\text{gene,t}}$ is robust to the number of individuals used to estimate LD, i.e. the sample size of the "LD panel" (Ch. 4.4.8 Material and Methods). Compared to in-sample LD computed from the full set of individuals in the GWAS (N = 290,273), using a random subset of N={500, 1000, 2500, 5000} individuals from the original GWAS does not significantly impact the MSE of $\hat{h}^2_{\text{gene,t}}$ or $\hat{h}^2_{\text{gene,r}}$ (Supplementary Figure 15). Using 90%-CIs to identify potential causal genes (i.e. 90%-CI lower bound > 0), we observe a slight increase in the false positive rate for both $\hat{h}^2_{\text{gene,t}}$ and $\hat{h}^2_{\text{gene,r}}$ as N decreases (Supplementary Figure 16); this is accompanied by a slight increase in power for $\hat{h}^2_{\text{gene,t}}$ but not for $\hat{h}^2_{\text{gene,r}}$ (Supplementary Figure 17). Since the N=5,000 LD panel and the full in-sample LD yield similar false positive rates for both estimators, we recommend using an in-sample LD panel of no less than 5,000 individuals (see Ch. 4.3 Discussion for additional comments on LD panels).

4.2.5 Rare-variant component of gene-level heritability links complex traits to phenotypically related monogenic disorders

We estimate, and partition by MAF, the gene-level heritabilities of 17,437 genes for 25 quantitative traits in the UK Biobank (N=290,273 "unrelated white British" individuals, M=5,650,812 with MAF > 0.5%, imputed data; Ch. 4.4.9 Material and Methods). Unless otherwise stated, the quantity of interest, $h^2_{\mathrm{gene,t}}$, is a function of the variants located in the gene body *and* the variants located within 10-kb upstream/downstream from the gene start/end positions. A gene is classified as having "nonzero heritability" if it meets two criteria: (i) the 90%-CI for $h^2_{\mathrm{gene,t}}$ does not overlap zero and (ii) the 90%-CI for at least one MAF component ($h^2_{\mathrm{gene,r}}$, $h^2_{\mathrm{gene,lf}}$, or $h^2_{\mathrm{gene,c}}$) does not overlap zero. Using this definition, the number of nonzero-heritability genes ranges from 1,212 (7%) for corneal hysteresis to 2,469 (14%) for height (Table 4.1). Most of the estimated posterior means for these genes lie between 10^{-6} and 10^{-4} (Figure 4.4).

As expected, $\hat{h}^2_{\mathrm{gene,c}}$ behaves similarly to $\hat{h}^2_{\mathrm{gene,t}}$. The average Pearson R^2 of $\hat{h}^2_{\mathrm{gene,c}}$ and $\hat{h}^2_{\mathrm{gene,t}}$ across the 25 traits is 94% (s.d. 1%) (Figure 4.4, Supplementary Figure 18). 92% (s.d. 1%) of nonzero-heritability genes have significant common-variant heritability; 76% (s.d. 1%) have significant causal effects exclusively from common variants (Table 4.1). On the other hand, $\hat{h}^2_{\mathrm{gene,r}}$ is significantly less correlated with $\hat{h}^2_{\mathrm{gene,t}}$ (average R2 = 30% (s.d. 21%) across traits) (Figure 4.4, Supplementary Figure 18). Approximately 2.5% (s.d. 0.6%) of genes have significant rare-variant heritability, and only 0.8% (s.d. 0.4%)—370 gene-trait pairs in total—have significant heritability exclusively from rare variants (Table 4.1, Supplementary Table 3). Of these 370 gene-trait pairs with only rare-variant heritability (ranging from 4 genes for heel T-score and corneal hysteresis to

32 genes for height (Table 4.1, Supplementary Table 3)), 232 gene-trait pairs are also identified by MAGMA[167] (FDR < 0.05, Ch. 4.4.9 Material and Methods). These 232 gene-trait pairs have a median $\hat{h}^2_{\text{gene,t}} \approx \hat{h}^2_{\text{gene,r}}$ on the order of 10^{-4} whereas the median for the remaining gene-trait pairs not found by MAGMA is $\sim 10^{-6}$. This suggests that MAGMA likely has limited power to detect signal from rare causal variants of moderate effect, which is expected as MAGMA tests for association between the total causal-variant signal at a gene and phenotype; it is not designed for partitioning the signal into components from different allele-frequency classes.

The 138 additional gene-trait pairs identified with our approach (Supplementary Table 4) include several genes implicated in phenotypically related Mendelian disorders. For example, *AKT2* is identified for serum gamma-glutamyl transferase (90%-CI of $h^2_{\text{gene,t}} = [3 \times 10^{-5}, 1 \times 10^{-4}]$, MAGMA z-score: 1.1), which is used to test for the presence of liver disease; *AKT2* is implicated in monogenic forms of type 2 diabetes[168] and hypoinsulinemic hypoglycemia with hemihypertrophy[169]. The *AKT2* annotation used for this analysis contains a total of 104 variants; 24 are rare variants, of which 1 is identified as causal. For serum alkaline phosphatase (used to diagnose diseases related to the liver or skeletal system), we identify *MDM4* (90%-CI of $h^2_{\text{gene,t}} = [4 \times 10^{-7}, 5 \times 10^{-6}]$, MAGMA z-score: 1.3; annotation contains 273 variants; 144 are rare variants, of which ~5 are identified as causal), which encodes a negative regulator of p53-mediated transcription[170] that was recently implicated in an autosomal dominant bone marrow failure syndrome[171]. *COL4A4*, identified for serum apolipoprotein A1 (a test for atherosclerotic cardiovascular disease; 90%-CI of $h^2_{\text{gene,t}} = [4 \times 10^{-5}, 2 \times 10^{-4}]$; MAGMA z-score: 1.1; annotation contains 390 variants; 33 are rare variants, of which ~1 is identified as causal), is implicated in monogenic forms of kidney disease ranging in severity from hematuria to end-stage renal disease[172–175].

We also identify several genes implicated in congenital developmental and metabolic disorders. For example, *RTTN*, identified for mean corpuscular hemoglobin (90%-CI of $h^2_{\text{gene,t}} = [9 \times 10^{-6}, 2 \times 10^{-4}]$; MAGMA z-score: 2.2; annotation contains 369 variants; 83 are rare, of which ~2 are identified as causal), is implicated in microcephaly, short stature, and polymicrogyria with seizures[176–179]. *SLC25A24*, identified for serum cystatin C (90%-CI of $h^2_{\text{gene,t}} = [3 \times 10^{-5}, 2 \times 10^{-4}]$; MAGMA z-score: 1.8; annotation contains 243 variants; 21 are rare, of which ~1 is causal), is implicated in Fontaine progeroid syndrome[180,181]. *TBCK*, identified for red blood cell count (90%-CI of $h^2_{\text{gene,t}} = [3 \times 10^{-5}, 2 \times 10^{-4}]$; MAGMA z-score: 2.0; annotation contains 617 variants; 59 are rare, of which ~1 is causal), is implicated in infantile hypotonia with psychomotor retardation and characteristic facies[182–184].

Taken together, these findings indicate that the rare-variant contribution to total gene-level heritability is indeed useful for identifying disease-relevant genes, especially those with moderate or relatively low total heritability, which existing methods can be underpowered to detect. Our results are consistent with the hypothesis that complex-trait variation may be explained in part by dysregulation of genes that—if completely disrupted—cause phenotypically similar or related Mendelian disorders[56]. We emphasize that, since heritability reflects genetic and phenotypic variation at the population level, if a common variant and rare variant explain the same heritability (i.e. have the same standardized causal effect size), the allelic effect—the expected change in phenotype per additional copy of the effect allele—is significantly larger for the rare variant.

4.2.6 Loss-of-function intolerant genes are overrepresented among genes with only rare-variant heritability

We estimate, and partition by MAF, the gene-level heritabilities of three gene sets: (i) known Mendelian-disorder genes from OMIM[164] (n=3,446), (ii) loss-of-function (LoF)-intolerant genes (probability of LoF-intolerance (pLI) > 0.9)[165] (n=3,230), and (iii) a set of FDA-approved drug targets for 30 immune-related traits[166] (n=216) (Ch. 4.4.9 Material and Methods). Compared to a set of "null" genes (sampled from the set of genes not contained in any of the three gene sets), all three gene sets have significantly higher median estimates of total and MAF-partitioned gene-level heritability (Figure 4.5).

We investigate whether certain classes of nonzero-heritability genes are overrepresented in the Mendelian-disorder and LoF-intolerant gene sets. The Mendelian-disorder gene set comprises ~20% of all genes and is enriched for genes with nonzero heritability for at least one trait (Fisher's exact test, 95%-CI of OR: [1.2, 1.4]); the number of genes in both categories ranges from 261 for corneal hysteresis to 557 for height. The LoF-intolerant genes comprise ~19% of all genes and are also enriched for nonzero-heritability genes (Fisher's exact test, 95%-CI of OR: [1.5, 1.7]); the overlap between the two categories ranges from 314 genes for corneal hysteresis to 650 for height. In contrast, genes with exclusively rare-variant heritability are significantly enriched in the LoF-intolerant gene set (95%-CI of OR: [1.1, 2.1]) but not in the Mendelian-disorder gene set (95% CI of OR: [0.9, 1.7]). On average across traits, ~19% (s.d. 11%) of the previously identified $h^2_{\mathrm{gene,t}} = h^2_{\mathrm{gene,r}}$ genes and ~21% (s.d. 1%) of genes with only common-variant heritability are also in the Mendelian-disorder gene set. In contrast, ~32% (s.d. 16%) of genes with $h^2_{\mathrm{gene,t}} =$

$h^2_{\text{gene,r}}$ are also in the LoF-intolerant gene set, compared with ~23% (s.d. 1%) of genes with $h^2_{\text{gene,t}} = h^2_{\text{gene,c}}$.

4.2.7 MAF-partitioned gene-level heritability reveals unique insights into complex-trait genetic architecture

We investigated whether gene-level heritability estimates are correlated with gene length, average LD score of variants in the gene (a proxy for the strength of LD in the region), and average MAF of variants in the gene. $h^2_{\text{gene,c}}$ (and, to a large extent, $h^2_{\text{gene,lf}}$) is distributed very similarly to $h^2_{\text{gene,t}}$ with respect to these variables (Figure 4.6, Supplementary Figure 19). However, the distribution of $h^2_{\text{gene,r}}$ shows marked differences, particularly with respect to gene length. Specifically, we observe higher average $h^2_{\text{gene,r}}$ among shorter genes even though the number of causal variants per gene (across all allele frequencies) increases with gene length (Figure 4.6, Supplementary Figure 20). The expected per-causal variant effect size per gene is invariant to gene length for common and low-frequency variants, but for rare variants, the average across gene-trait pairs is nearly 10^{-4} in the shortest quintile of genes versus 10^{-6} in the longest (Figure 4.6). While this result initially seems paradoxical, it is not inconsistent with the literature; previous studies have reported strong inverse correlations between gene length and expression which could be due to, for example, natural selection favoring fewer/shorter introns in highly expressed genes due to the high energy/costs associated with transcription and splicing[185,186].

Using the empirical distributions of cumulative $h^2_{\text{gene,t}}$, $h^2_{\text{gene,c}}$, $h^2_{\text{gene,lf}}$, and $h^2_{\text{gene,r}}$, we loosely quantify differences in polygenicity at the level of genes (with the caveat that, since there is a high degree of gene overlap in some regions, cumulative $h^2_{\mathrm{gene,t}}$ may be more informative for some traits over others) (Figure 4.7). For example, if cumulative $h^2_{\mathrm{gene,t}}$ is divided equally among nonzero-heritability genes, the empirical CDF for $h^2_{\mathrm{gene,t}}$ would be the line y = x, where the x-axis is the rank ordering of genes from highest to lowest $h^2_{\mathrm{gene,t}}$; two traits with the same empirical CDF for $h^2_{\mathrm{gene,t}}$ can have different empirical CDFs for each MAF-partitioned component. Once again, we find that the cumulative distributions of $h^2_{\mathrm{gene,c}}$ are extremely similar to those of $h^2_{\mathrm{gene,t}}$ (Figure 4.7, Supplementary Figure 21). Although the curves generally have similar shapes across traits (i.e. similar spread of heritability across genes), some traits have a notable amount of heritability concentrated in just the top gene, and many of these gene-trait pairs have been functionally validated in the literature. For example, for serum urate concentration, *SLC2A9* — a known urate transporter[187–189] — is the single largest contributor to total, common-, and LF-variant gene-level heritability ($\hat{h}^2_{\mathrm{gene,t}} = 6.2\%$, $\hat{h}^2_{\mathrm{gene,c}} = 5.9\%$, $\hat{h}^2_{\mathrm{gene,lf}} = 0.3\%$, $\hat{h}^2_{\mathrm{gene,r}} = 0$), accounting for 46%, 51%, and 29% of the cumulative heritability for each estimand, respectively (Figure 4.7); certain loss-of-function mutations in *SLC2A9* are known to cause a rare form of renal hypouricemia[190–192], a disorder characterized in part by low serum urate levels. For serum alkaline phosphatase, we find that *ALPL* — which encodes the enzyme alkaline phosphatase — is the single largest contributor to total and LF-variant gene-level heritability ($\hat{h}^2_{\mathrm{gene,t}} = 4.1\%$, $\hat{h}^2_{\mathrm{gene,c}} = 1.8\%$, $\hat{h}^2_{\mathrm{gene,lf}} = 2.1\%$, $\hat{h}^2_{\mathrm{gene,r}} = 0\%$), explaining 15% and 39% of the respective cumulative heritability estimands (Figure 4.7); certain loss-of-function mutations in *ALPL* are known to cause hypophosphatasia, a monogenic disorder characterized in part by low alkaline phosphatase[193,194].

4.3 Discussion

We propose a general approach for estimating the heritability explained by any set of variants much smaller than an LD block and assess its utility in estimating/partitioning gene-level heritability. In simulations, we confirm that incorporating uncertainty about which variants are causal and what their effect sizes are dramatically improves specificity over naive approaches that ignore uncertainty in the causal effects. For 25 complex traits and >17K genes, we estimate gene-level heritability—the heritability explained by variants in the gene body plus a 10-kb window upstream/downstream from the gene start/end positions—and partition by allele-frequency class to explore differences in genetic architecture across traits. As expected, most gene-level heritability is dominated by common variants, but we identify several genes with nonzero heritability exclusively from rare or low-frequency variants. Notably, we identify many genes with nonzero gene-level heritability explained exclusively by rare variants that existing methods are underpowered to detect. Many of these genes have known roles in Mendelian disorders that are phenotypically similar or related to the complex trait; we also identify genes implicated in systemic congenital developmental and metabolic disorders. Our results demonstrate that the rare-variant contribution to total gene-level heritability is a useful quantity that can be considered alongside common-variant heritability enrichments to obtain a more comprehensive understanding of genetic architecture.

We conclude by discussing the limitations of our approach. First, multiple lines of evidence suggest that rare and “ultra-rare” variants, which are not well-tagged by variants on genotyping arrays, may explain much of the “missing heritability” not captured by genotyped or imputed variants[101,142,195,196]. Since imputed genotypes are noisier for rarer variants and variants in lower

LD regions, we analyze variants with MAF > 0.5%. Additional work is needed to assess the error incurred by using genotyped/imputed data in lieu of whole genome sequencing (WGS) as well as the signal that is missed by excluding variants with MAF < 0.5%. While our estimator can be applied to whole exome sequencing (WES) data, LD between coding and noncoding regions would significantly inflate gene-level heritability estimates; LD between exonic and intronic variants could also cloud interpretation, depending on the application. With multiple biobanks starting to sequence large numbers of individuals[13,197–199], we believe the availability of large-scale WGS data will gradually become less of an issue.

We correct for population structure using genome-wide principal components (PCs) computed from the same imputed genotypes that are used to perform each GWAS. This is a standard approach to correcting for population stratification, which typically reflects geographic separation, in estimates of genome-wide SNP-heritability and genome-wide functional enrichments, both of which are driven by common SNPs. However, rare variants generally have more complex spatial distributions and thus exhibit stratification patterns distinct from those of common SNPs[196,200]. It is unclear whether methods that are effective for controlling stratification of common SNPs are applicable to rare variants[201]. We leave the question of whether uncorrected structure among rare variants significantly influences our estimates of gene-level heritability for future work.

Our approach requires OLS association statistics and LD computed from a subset of individuals in the GWAS. While estimates of gene-level heritability and the MAF-partitioned components are robust to sample sizes as low as 5,000, the individuals used to estimate LD must be a subset of the individuals in the GWAS. Although summary association statistics are publicly available for hundreds of large-scale GWAS, most of these studies are meta-analyses and therefore do not have in-sample LD available. Moreover, many publicly available summary statistics were computed from linear mixed models rather than OLS, which is used throughout our simulations and derivations. Additional work is needed to extend our approach to allow external reference panel LD (e.g., 1000 Genomes) and/or mixed model association statistics. Biobanks can help to ameliorate potential issues stemming from noisy LD by releasing summary LD information in addition to summary association statistics[202].

Finally, gene-level heritabilities of different genes can have nonzero covariance due to physical overlap between genes and/or correlated causal effect sizes. Thus, the heritability estimates reported in this work have additional sources of noise/uncertainty which were not directly modeled or accounted for. Since modeling correlation of causal effect sizes would make inference considerably more challenging, we leave this for future work. Importantly, genes with credible intervals > 0 should not be interpreted as "causal" for the complex trait without additional functional validation, as nonzero gene-level heritability indicates association but not causality.

4.4 Material and Methods

4.4.1 Model and definitions of estimands

We model the phenotype of a given individual using a standard linear model, $y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon$, where $\mathbf{x}^{\mathrm{T}} = (x_1 \ldots x_M)^{\mathrm{T}}$ is the vector of standardized genotypes at M variants, i.e. $\mathbb{E}[x_i] = 0$ and $var[x_i] = 1$ for $i = 1, \ldots, M$. $\boldsymbol{\beta}$ is the $\mathrm{M} \times 1$ vector of standardized causal effect sizes, and $\epsilon \sim N(0, \sigma_e^2)$ is environmental noise. We assume that the phenotype is standardized in the population, i.e. $\mathbb{E}[y] = 0$, $var[y] = 1$. Linkage disequilibrium (LD) between variants i and j is defined as $r_{ij} \equiv cov[x_i, x_j] = \mathbb{E}[x_i x_j]$ and the full LD matrix for all M variants is $\mathbf{R} \equiv cov[\mathbf{x}^{\mathrm{T}}]$.

Letting $p_{\text{causal}} \in [0,1]$ such that $M \times p_{\text{causal}}$ is the total number of causal variants, we assume the causal effect of the *i*-th variant is $\beta_i \sim N\left(0, \frac{h_G^2}{M \times p_{\text{causal}}}\right)$ with probability p_{causal} or $\beta_i = 0$ with probability $1 - p_{\text{causal}}$. Under this model, total SNP-heritability h_G^2 is defined as the proportion of phenotypic variance explained by the M variants,

$$
\begin{aligned}
h_{\mathrm{G}}^2 &\equiv \frac{var[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}]}{var[y]} \\
&= \mathbb{E}_{\boldsymbol{\beta}}\left[var[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}|\boldsymbol{\beta}]\right] + var_{\boldsymbol{\beta}}\left[\mathbb{E}[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}|\boldsymbol{\beta}]\right] \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}^{\mathrm{T}} var[\mathbf{x}^{\mathrm{T}}]\boldsymbol{\beta}] + var_{\boldsymbol{\beta}}[\mathbb{E}[\mathbf{x}^{\mathrm{T}}]\boldsymbol{\beta}] \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{R}\boldsymbol{\beta}] + var_{\boldsymbol{\beta}}[0] \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{R}\boldsymbol{\beta}]
\end{aligned}
$$

where the second line follows from the Law of Total Variance.

Let g index a gene of interest. Given an assignment of m_g variants to gene g, let $\mathbf{x}_g^{\mathrm{T}}$ be the $m_g \times 1$ vector of genotypes at this set of variants and let $\mathbf{x}_{g'}^{\mathrm{T}}$ be the genotypes of the remaining $M - m_g$ variants. We can rewrite the total SNP-heritability of the trait in terms of gene g as

$$
\begin{aligned}
h_{\mathrm{G}}^2 &= \mathrm{Var}\left[\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g + \mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}\right] \\
&= \mathrm{Var}\left[\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g\right] + \mathrm{Var}\left[\mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}\right] + 2\mathrm{Cov}\left[\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g, \mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{R}_{g'}\boldsymbol{\beta}_{g'}\right] + 2\left[\mathrm{E}\left[\left(\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g\right)\left(\mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}\right)\right] - \mathrm{E}\left[\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g\right]\mathrm{E}\left[\mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}\right]\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{R}_{g'}\boldsymbol{\beta}_{g'}\right] + 2\mathrm{E}_{\boldsymbol{\beta}}\left[\mathrm{E}\left[\left(\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g\right)(\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{x}_{g'})\middle|\boldsymbol{\beta}\right]\right] \\
&\qquad\qquad - 2\mathrm{E}_{\boldsymbol{\beta}}\left[\mathrm{E}\left(\mathbf{x}_g^{\mathrm{T}}\boldsymbol{\beta}_g|\boldsymbol{\beta}\right)\right]\mathrm{E}_{\boldsymbol{\beta}}\left[\mathrm{E}\left(\mathbf{x}_{g'}^{\mathrm{T}}\boldsymbol{\beta}_{g'}|\boldsymbol{\beta}\right)\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{R}_{g'}\boldsymbol{\beta}_{g'}\right] + 2\mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathrm{E}\left[\mathbf{x}_{g'}\mathbf{x}_g^{\mathrm{T}}\right]\right] - 0 \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{R}_{g'}\boldsymbol{\beta}_{g'}\right] + 2\mathrm{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g\boldsymbol{\beta}_{g'}^{\mathrm{T}}\right]\mathrm{E}_{\mathrm{x}}\left[\mathbf{x}_{g'}\mathbf{x}_g^{\mathrm{T}}\right]
\end{aligned}
$$

where the fourth line follows from the Law of Total Expectation. If we additionally assume that $cov[\beta_i, \beta_j] = 0$ for all $i \neq j$, then $\mathbb{E}\left[\boldsymbol{\beta}_{(g)}\boldsymbol{\beta}_{(g')}^{\mathrm{T}}\right] = cov\left[\boldsymbol{\beta}_{(g)}, \boldsymbol{\beta}_{(g')}\right] = 0$, which simplifies the above equation to

$$h_{\mathrm{G}}^2 = \mathbb{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g\right] + \mathbb{E}_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}_{g'}^{\mathrm{T}}\mathbf{R}_{g'}\boldsymbol{\beta}_{g'}\right]$$

We refer to the first term, the component of heritability attributable to the causal effects in gene g, as *total gene-level heritability*, i.e.

$$h_{\mathrm{gene,t}}^2 = \boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g$$

Using the same assumptions as above, we can partition the variants in gene g by minor allele frequency such that

$$h^2_{\text{gene,t}} = h^2_{\text{gene,r}} + h^2_{\text{gene,lf}} + h^2_{\text{gene,c}}$$

where $h^2_{\text{gene,r}}$, $h^2_{\text{gene,lf}}$, and $h^2_{\text{gene,c}}$ are the components of $h^2_{\text{gene,t}}$ attributable to the causal effects of rare (MAF < 0.01), low-frequency (0.01 ≤ MAF < 0.05), and common (MAF ≥ 0.05) variants, respectively. The estimands of interest in this work are the four terms in $h^2_{\text{gene,t}} = h^2_{\text{gene,r}} + h^2_{\text{gene,lf}} + h^2_{\text{gene,c}}$.

4.4.2 Estimating the posterior distribution of gene-level heritability

Since we have neither the "true" causal effect sizes, $\boldsymbol{\beta}$, nor the population LD, $\mathbf{R}$, we must estimate both from data. We consider one approximately independent LD block at a time. Given a GWAS of N individuals, let $\mathbf{X} = [\mathbf{x}_1^{\mathrm{T}}, \dots, \mathbf{x}_N^{\mathrm{T}}]^{\mathrm{T}}$ be the $N \times M$ matrix of standardized genotypes measured at M variants, let $\mathbf{y} = (y_1, \dots, y_N)^{\mathrm{T}}$ be an $N \times 1$ vector of phenotypes, and let $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$ be environmental noise.

It is often the case that individual-level genotype data are inaccessible for privacy or logistical reasons. However, GWAS summary statistics—estimates of the causal effects and their standard errors—are publicly available for thousands of traits. Ordinary least squares (OLS) estimates of the causal effects are often provided, defined as

$$\widehat{\boldsymbol{\beta}}_{\text{GWAS}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \frac{1}{N}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}$$

It follows that

$$p\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}\middle|\boldsymbol{\beta}, \widehat{\mathbf{R}}, \sigma_e^2\right) \sim MVN\left(\widehat{\mathbf{R}}\boldsymbol{\beta}, \frac{\sigma_e^2}{N}\widehat{\mathbf{R}}\right)$$

In this scenario, the observed data D are not the individual-level genotypes and phenotypes $(\mathbf{X}, \mathbf{y})$, but rather $\mathrm{D} = \left(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}, \widehat{\mathbf{R}}\right)$, where $\widehat{\mathbf{R}}$ is an estimate of LD computed from either the genotypes of a set of individuals in the GWAS ("in-sample" LD) or from an external reference panel (e.g., 1000 Genomes[7]). By combining the prior on $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}|\boldsymbol{\lambda})$ ($\boldsymbol{\lambda}$ represents the hyperparameters in the prior over $\boldsymbol{\beta}$), and the likelihood of the observed data, $p\left(\widehat{\boldsymbol{\beta}}_{\text{GWAS}}\middle|\boldsymbol{\beta}, \widehat{\mathbf{R}}, \sigma_e^2\right)$, one can compute the posterior distribution of the causal effects, $p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{\text{GWAS}}, \widehat{\mathbf{R}}, \boldsymbol{\lambda}, \sigma_e^2)$. The hyperparameters $\boldsymbol{\lambda}$ and σ_e^2 can be estimated with an empirical Bayes procedure as in SuSiE[163] framework. We note that for computational efficiency, we can partition the whole genome into approximately independent LD blocks and estimate the posterior distribution of $\boldsymbol{\beta}$ separately for each LD block. Because each LD block is approximately independent of the rest of the genome, the genetic effects of SNPs outside of the LD block of interest are absorbed into the environmental noise. Correspondingly, the LD block-specific hyperparameters $(\boldsymbol{\lambda}, \sigma_e^2)$ are estimated independently for each LD block.

In general, the posterior of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}|\mathrm{D})$, is computationally intractable. Approximate inference, e.g., Markov Chain Monte Carlo (MCMC) or variance inference, can be used to approximate the exact posterior $p(\boldsymbol{\beta}|\mathrm{D})$ as $\tilde{p}(\boldsymbol{\beta}|\mathrm{D})$. In this work, we use SuSiE[163], a variational inference-based implementation of linear regression with a sparse prior. (In principle, it is straightforward to use other implementations of linear regression with a sparse prior). We draw K samples from the posterior of the causal effects, $\widetilde{\boldsymbol{\beta}}^{(1)}, \ldots, \widetilde{\boldsymbol{\beta}}^{(K)} \sim \tilde{p}(\boldsymbol{\beta}|\mathrm{D})$. This approximate distribution can in turn be used to approximate the full posterior distribution of h_{gene}^2, i.e.

$\left(\widetilde{\boldsymbol{\beta}}_g^{(1)}\right)^{\mathrm{T}}\widehat{\mathbf{R}}_g\left(\widetilde{\boldsymbol{\beta}}_g^{(1)}\right), \dots, \left(\widetilde{\boldsymbol{\beta}}_g^{(K)}\right)^{\mathrm{T}}\widehat{\mathbf{R}}_g\left(\widetilde{\boldsymbol{\beta}}_g^{(K)}\right)$. Finally, given the approximate posterior of h^2_{gene}, one can compute the posterior mean,

$$\begin{aligned}\hat{h}^2_{\text{gene}} &= \widehat{\mathrm{E}}\left[\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{R}_g\boldsymbol{\beta}_g \middle| \mathrm{D}\right] \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\left(\widetilde{\boldsymbol{\beta}}_g^{(k)}\right)^{\mathrm{T}}\widehat{\mathbf{R}}_g\left(\widetilde{\boldsymbol{\beta}}_g^{(k)}\right)\end{aligned}$$

and measures of uncertainty such as credible intervals (described below). Similar procedures can be applied to partition gene-level heritability (e.g., by MAF-based annotations).

4.4.3 Quantifying uncertainty in gene-level heritability estimates

$\widetilde{\boldsymbol{\beta}}^{(1)}, \dots, \widetilde{\boldsymbol{\beta}}^{(K)}$ provide an approximation to the full posterior distribution of $\boldsymbol{\beta}$, thus capturing *uncertainty* about the causal effect sizes arising from two main sources: LD and finite GWAS sample size (Figure 4.1). Therefore, by using the full posterior of $\boldsymbol{\beta}$ to approximate the full posterior of h^2_{gene}, we wish to capture uncertainty in the causal effects that propagates into our estimate of h^2_{gene}. (The noise in $\widehat{\mathbf{R}}$ is also an important factor but, for simplicity, we first investigate uncertainty in $\hat{h}^2_{\text{gene}}$ in simulations where $\widehat{\mathbf{R}} = \mathbf{R}$.)

We summarize the uncertainty in h^2_{gene} by computing ρ-level credible intervals (ρ-CIs). For a given $\rho \in [0,1]$, ρ-CI is defined as the central interval within which h^2_{gene} lies with probability ρ, i.e. the upper and lower bounds of ρ-CI are set to the empirical $\frac{1-\rho}{2}$ and $1-\left(\frac{1-\rho}{2}\right)$ quantiles of the posterior samples $\left(\widetilde{\boldsymbol{\beta}}_g^{(k)}\right)^{\mathrm{T}}\widehat{\mathbf{R}}_g\left(\widetilde{\boldsymbol{\beta}}_g^{(k)}\right), k = 1, \dots, K$.

4.4.4 Implementation details

We partition the genome into approximately independent LD blocks[123] and, for each gene of interest, we perform inference on the LD block containing the gene. For each LD block, we extract the marginal association statistics and estimate LD for all the variants in the LD block. We estimate the posterior distribution of effect sizes using the function "susie_suff_stat" with default parameters, as implemented in SuSiE v0.8. We use the function "susie_get_posterior_samples" to obtain 500 posterior samples.

4.4.5 Simulation framework

We obtain the real imputed genotypes of N=290,273 "unrelated white British" individuals in the UK Biobank by extracting individuals with self-reported British ancestry who are > third-degree relatives (pairs of individuals with kinship coefficient < $½^{(9/2)}$, as defined in ref.[11]). Filtering on MAF > 0.5% leaves 200,235 variants on chromosome 1. A list of 1,083 genes on chromosome 1 and their coordinates were downloaded from https://github.com/bogdanlab/gene_sets. For each variant, genotypes are standardized such that the mean is 0 and variance is 1 across individuals. Phenotypes were simulated under a variety of genetic architectures according to the following steps. First, we randomly select 3%, 8%, or 16% (out of the 1,083 genes) to be causal ($h^2_{\text{gene}} > 0$). Second, we draw causal variants in the causal gene bodies and within 10-kb upstream/downstream of the gene start/end positions; the causal variants in the window around the gene are intended to represent regulatory causal variants in transcription start sites (TSSs). The causal configuration is set to be either (1) 5 causal variants in gene body and 3 causal variants in TSS or (2) 10 causal variants in gene body and 6 causal variants in TSS. Third, we draw noncoding "background" causal variants across the whole chromosome with frequency $p_{\text{causal}} = \{0.001, 0.01\}$. Finally, conditional on the causal statuses of the variants, we draw independent causal effect sizes from a Gaussian distribution where the variance of each causal variant is standardized such that the gene bodies collectively have a heritability of 3%, TSSs collectively have 1%, and non-coding background variants together explain 1%. We note that the causal statuses and effect sizes for each variant are only drawn once; the environmental noise term is drawn 30 times independently to generate 30 simulation replicates.

4.4.6 Evaluating and comparing gene-level heritability estimates in simulations

Recall that for a given gene g, the causal effect sizes and LD of the variants assigned to the gene are denoted $\boldsymbol{\beta}_g$ and $\mathbf{R}_g$, and ground-truth gene-level heritability is defined as $h^2_{\text{gene}} = \boldsymbol{\beta}_g^{\mathrm{T}} \mathbf{R}_g \boldsymbol{\beta}_g$. The posterior mean estimated for a single simulation replicate s is denoted $\hat{h}^2_{\text{gene,(s)}}$. We estimate the bias of the estimator as $\text{bias}\left[\hat{h}^2_{\text{gene}}\right] \approx \frac{1}{30} \sum_s (\hat{h}^2_{\text{gene,(s)}} - h^2_{\text{gene}})$; the variance of the estimator as $\text{Var}\left[\hat{h}^2_{\text{gene}}\right] \approx \frac{1}{30} \sum_s \left(\hat{h}^2_{\text{gene,(s)}} - h^2_{\text{gene}}\right)^2$; and the mean squared error as $\text{MSE}\left[\hat{h}^2_{\text{gene}}\right] = \left(\text{bias}\left[\hat{h}^2_{\text{gene}}\right]\right)^2 + \text{Var}\left[\hat{h}^2_{\text{gene}}\right]$.

For each simulation replicate s, we also output ρ-level credible intervals, defined as $\text{CI}_{(s)} = \left(\hat{h}^2_{\text{gene},\frac{1-\rho}{2},(\text{s})}, \hat{h}^2_{\text{gene},1-\frac{1-\rho}{2},(\text{s})}\right)$, where the $\frac{1-\rho}{2}$ and $1 - \left(\frac{1-\rho}{2}\right)$ quantiles are estimated from the posterior samples. To assess the accuracy of credible intervals, we calculate *empirical coverage* across simulation replicates, defined as the proportion of simulation replicates in which the ρ-level credible interval covers the ground-truth gene-level heritability: $\frac{1}{30} \sum_s \mathbb{I}\left[\hat{h}^2_{\text{gene,(s)}} \in \text{CI}_{(s)}\right]$.

4.4.7 Comparison to “naïve” gene-level heritability estimator

We compare our approach to an alternative “naïve” estimator of gene-level heritability that does not model LD between the gene and its adjacent regions and thus ignores causal-effect uncertainty. This estimator is similar to existing methods that are meant to be applied to approximately independent LD blocks[73,74]. For each gene g, we extract the marginal association statistics, $\widehat{\boldsymbol{\beta}}_g$, and the estimated LD, $\widehat{\mathbf{R}}_g$, for the variants assigned to the gene, and we compute the alternative estimator as $\frac{N\widehat{\boldsymbol{\beta}}_g^\top \widehat{\mathbf{R}}_g^\dagger \widehat{\boldsymbol{\beta}}_g}{N-q}$, where $\widehat{\mathbf{R}}_g^\dagger$ and q are the pseudo-inverse and rank of $\widehat{\mathbf{R}}_g$, respectively[74].

4.4.8 Assessing robustness to LD panel sample size

To assess the robustness of our approach to the sample size of the LD panel used to estimate LD, we randomly draw a subset of N={500, 1000, 2500, 5000} individuals from the full 290,273 individuals. After extracting variants with MAF > 0.5%, genotypes are standardized to have mean 0 and variance 1, similar to the full-sample analysis. Since we are interested in assessing robustness to noisy estimates of LD, all analyses are performed using the same set of marginal association statistics used in the full-sample analysis, excluding the variants that were filtered from the LD panel based on MAF. The LD and marginal association statistics are fed into the *h2gene* software, similar to the full-sample analysis.

4.4.9 Analysis of 25 UK Biobank phenotypes

We analyzed 25 quantitative phenotypes in the UK Biobank. Phenotypes and imputed genotypes were filtered according to the same procedures used in the simulation analyses, leaving N=290,273 individuals and M=5,650,812 variants with MAF > 0.5%. Quantitative phenotypes were quantile-normalized to a Gaussian distribution with mean 0 and variance 1. We then performed a GWAS for each trait using the "assoc" option in PLINK[87,88] with age, sex, and the top 10 genetic principal components included as covariates. We computed in-sample LD for each approximately independent LD block. We downloaded gene names and coordinates from https://github.com/bogdanlab/gene_sets and, for each gene, we define the estimand of interest to be a function of the variants in the gene body *and* those located within 10-kb upstream/downstream of the gene start/end positions. Finally, given the in-sample LD and marginal association statistics, we infer the posterior distribution of the causal effect sizes one LD block at a time, and we estimate and partition gene-level heritability for all genes in each LD block. MAGMA[167] v1.09 was used for gene-level association with a 10kb window around each gene. The same list of genes and the same set of imputed variants were used for the MAGMA analysis.

4.5 Figures

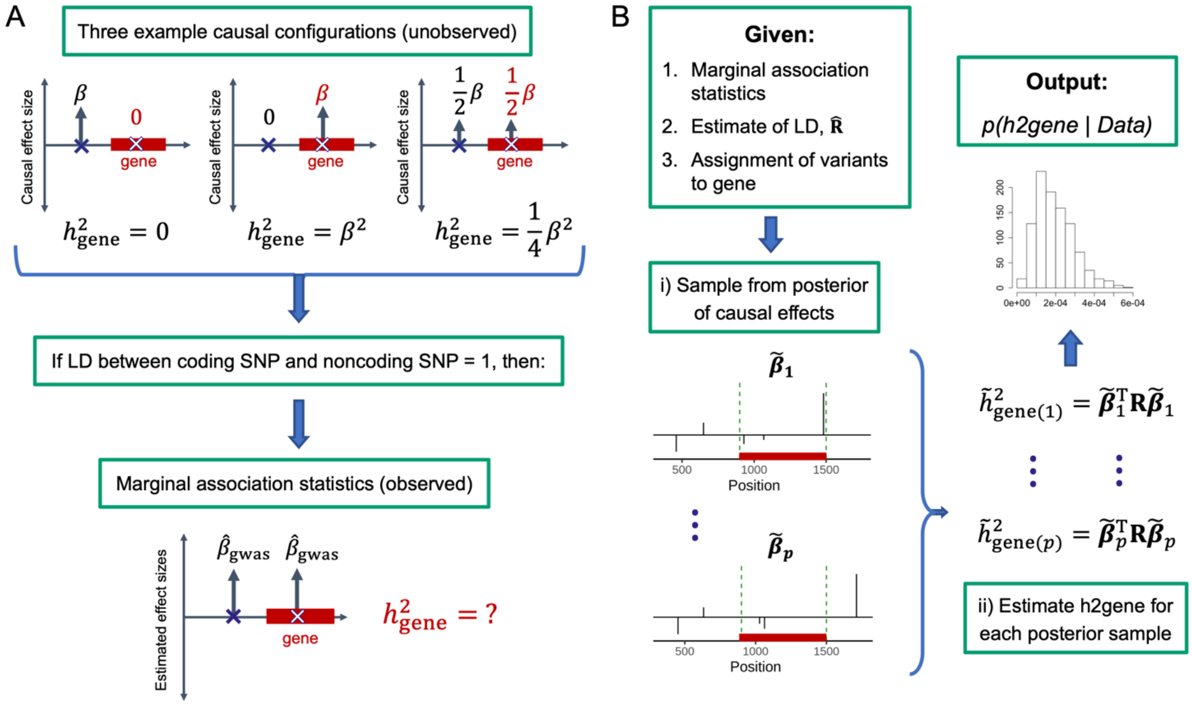
A
Three example causal configurations (unobserved)
Causal effect size
β
0
gene
$h^2_{\text{gene}} = 0$
Causal effect size
0
β
gene
$h^2_{\text{gene}} = \beta^2$
Causal effect size
$\frac{1}{2}\beta$
$\frac{1}{2}\beta$
gene
$h^2_{\text{gene}} = \frac{1}{4}\beta^2$
If LD between coding SNP and noncoding SNP = 1, then:
Marginal association statistics (observed)
Estimated effect sizes
$\hat{\beta}_{\text{gwas}}$
$\hat{\beta}_{\text{gwas}}$
gene
$h^2_{\text{gene}} = ?$
B
Given:
1. Marginal association statistics
2. Estimate of LD, $\hat{\mathbf{R}}$
3. Assignment of variants to gene
i) Sample from posterior of causal effects
$\tilde{\boldsymbol{\beta}}_1$
500
1000
1500
Position
$\tilde{\boldsymbol{\beta}}_p$
500
1000
1500
Position
Output:
p(h2gene | Data)
0
50
100
150
200
0e+00
2e-04
4e-04
6e-04
$\tilde{h}^2_{\text{gene}(1)} = \tilde{\boldsymbol{\beta}}_1^{\text{T}}\mathbf{R}\tilde{\boldsymbol{\beta}}_1$
$\tilde{h}^2_{\text{gene}(p)} = \tilde{\boldsymbol{\beta}}_p^{\text{T}}\mathbf{R}\tilde{\boldsymbol{\beta}}_p$
ii) Estimate h2gene for each posterior sample

Figure 4.1 Toy example illustrating (A) the impact of causal-effect uncertainty on gene-level heritability and (B) our approach to capturing uncertainty in gene-level heritability estimation.
(A) Toy example with two variants, one of which is assigned to the gene of interest. The top row depicts 3 example causal configurations corresponding to 3 different gene-level heritabilities (0, β^2, and $\beta^2/4$). Since the variants in are in perfect LD, all 3 causal configurations yield the same expected marginal association statistics. (B) Given marginal association statistics, an estimate of LD, and an assignment of variants to the gene of interest, our approach involves i) sampling from the posterior of the causal effect sizes (assuming a sparse prior) to capture our uncertainty about which variants are causal, and then ii) estimating gene-level heritability for each posterior sample to approximate the posterior distribution of gene-level heritability.

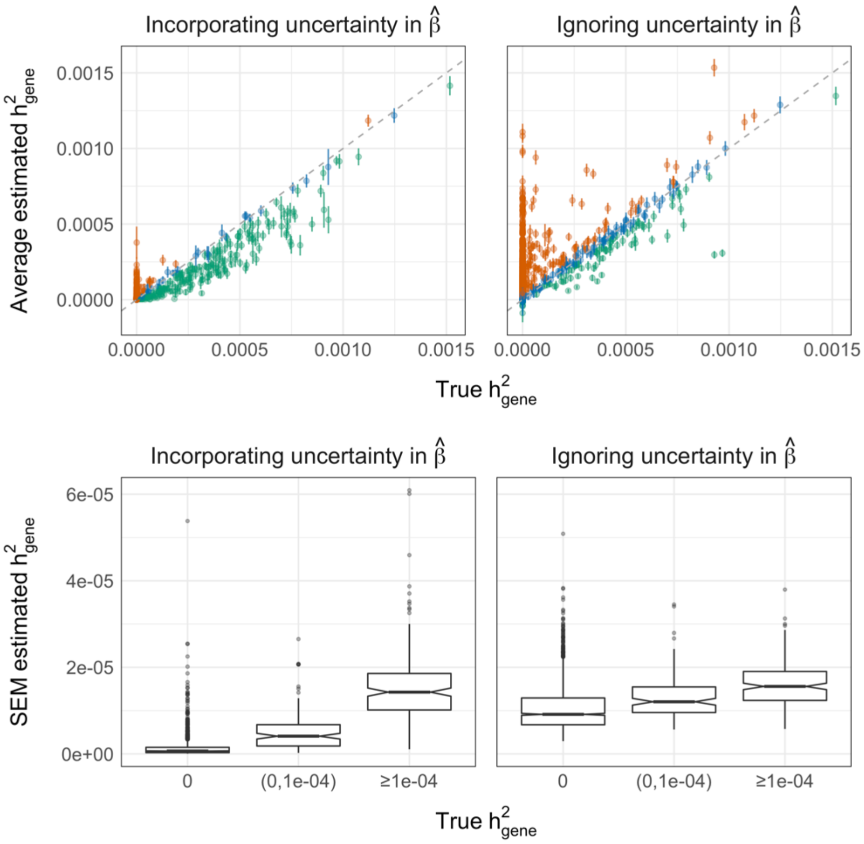

Figure 4.2 Ignoring uncertainty in gene-level h^2 estimation significantly increases false positives. Impact of causal-effect uncertainty on gene-level heritability estimation in simulations. Chromosome 1, MAF > 0.5%, p_{causal}=0.01, N=290K individuals, and 1,038 genes, of which 16% have nonzero gene-level heritability. Top row: each point is the average $\hat{h}^2_{\text{gene}}$ for a given gene across 30 simulation replicates; error bars mark 1.96 × standard error of the mean (SEM). Orange and green points are genes for which the estimator is significantly upward-biased and downward-biased, respectively. Bottom row: distributions of SEM with respect to gene-level heritability.

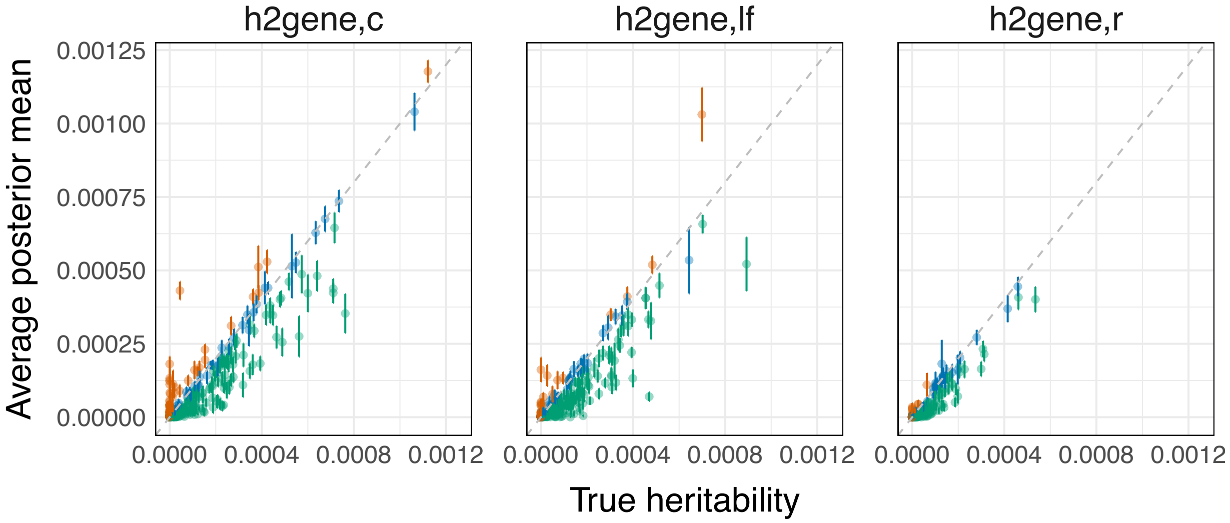

Figure 4.3 Incorporating uncertainty via MCMC enables accurate partitioning of gene-level h^2. Estimates of the heritability contributions of common, low-frequency, and rare variants in simulations. Chromosome 1, MAF > 0.5%, p_{causal}=0.01, N=290K individuals, and 1,083 genes, of which 16% have nonzero heritability. Each point is the average posterior mean for a given gene from 30 simulation replicates; error bars mark 1.96 x SEM. Orange and green points are genes for which the estimator is significantly upward-biased and downward-biased, respectively, where significance is determined by the error bars.

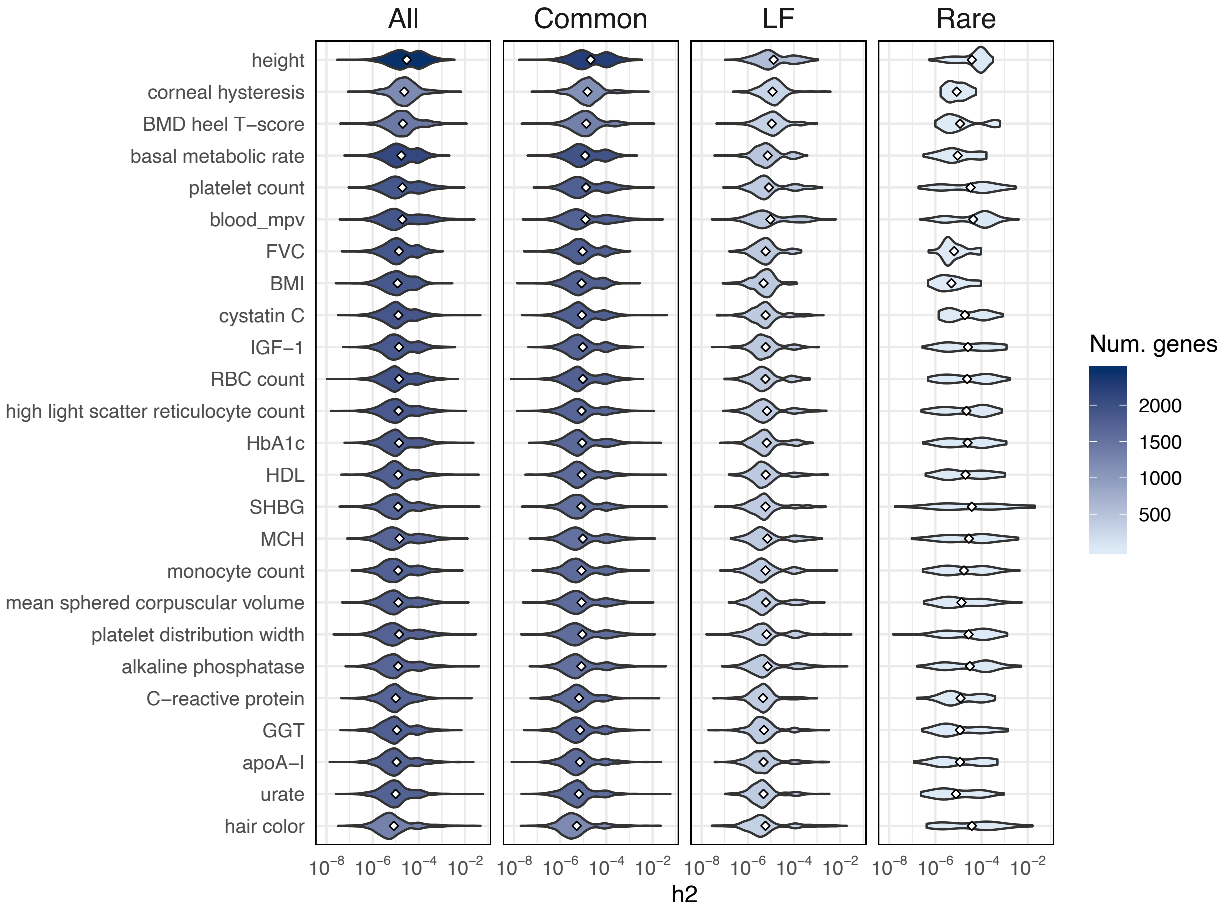

Figure 4.4 Distributions of total and MAF-partitioned gene-level heritability estimates for 25 traits. Each violin plot is the distribution of posterior mean estimates for genes with 90%-CI > 0 for one trait. The shading scales with the number of genes in the violin plot.

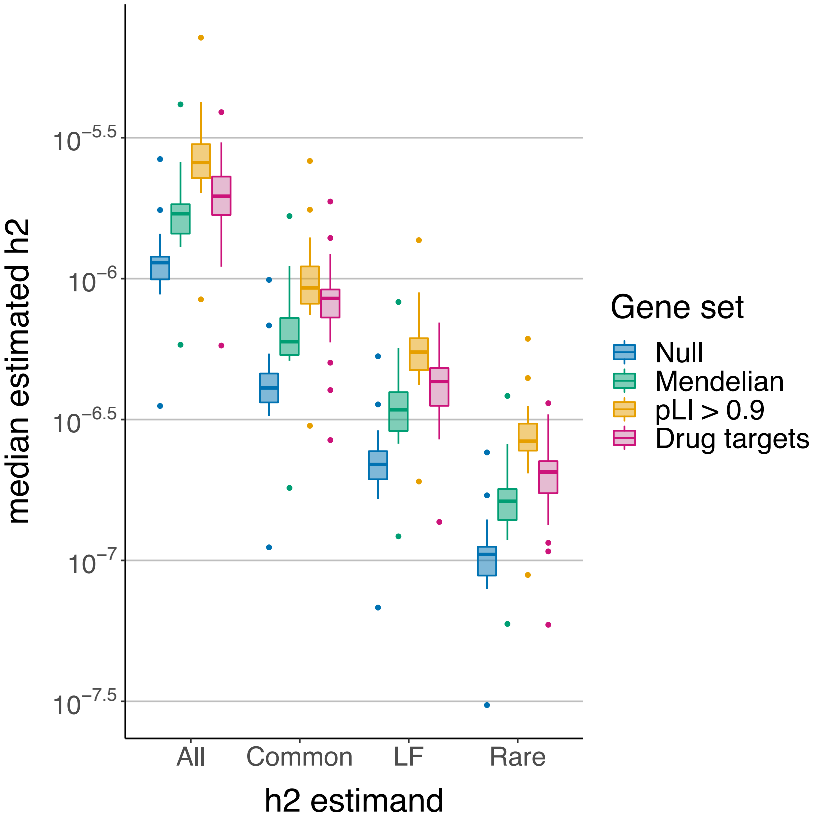

Figure 4.5 Total and MAF-partitioned gene-level heritability estimates for Mendelian-disorder genes (n=3,446), LoF-intolerant genes (n=3,230), and immune-related drug targets (n=216). Each point is the median posterior mean across genes for a given trait; each boxplot contains 25 quantitative traits in the UK Biobank.

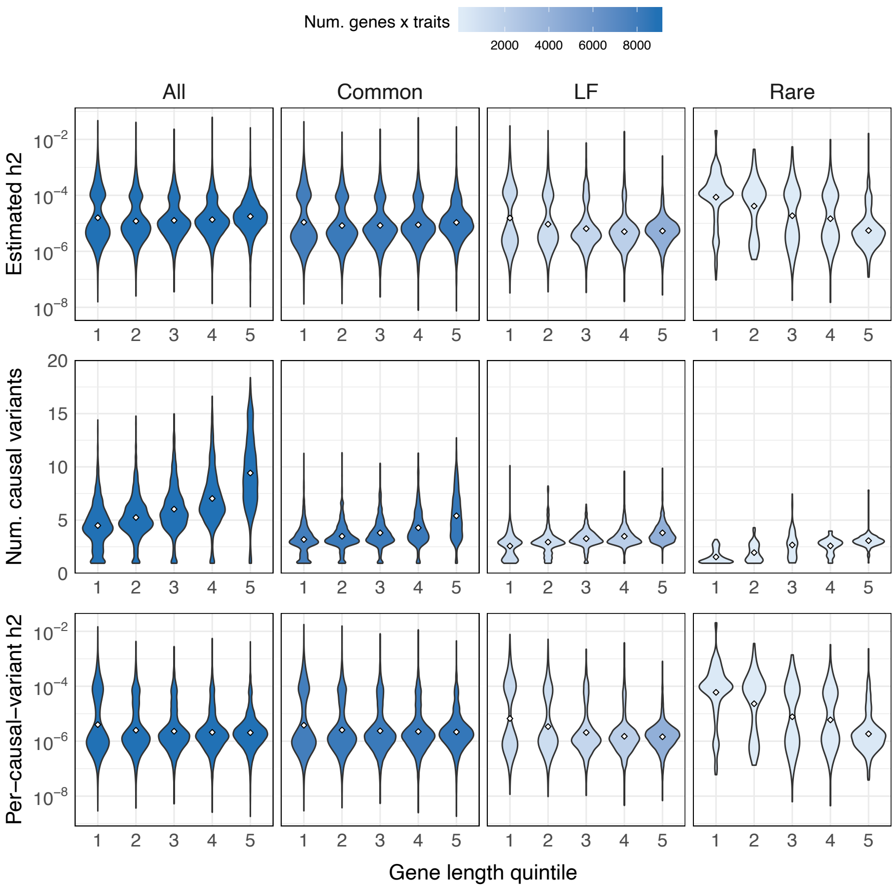

Figure 4.6 Inverse relationship between rare-variant gene-level h^2 estimates and gene length. Estimates of h2 (top), number of causal variants per gene (middle), and expected effect size per causal variant per gene (bottom) with respect to gene length (x-axis) for 25 traits. Each violin plot is the distribution of posterior mean estimates for nonzero-heritability genes with 90%-CIs > 0 for each h2 quantity. Color gradient indicates the number of estimates in each violin plot (number of gene-trait pairs).

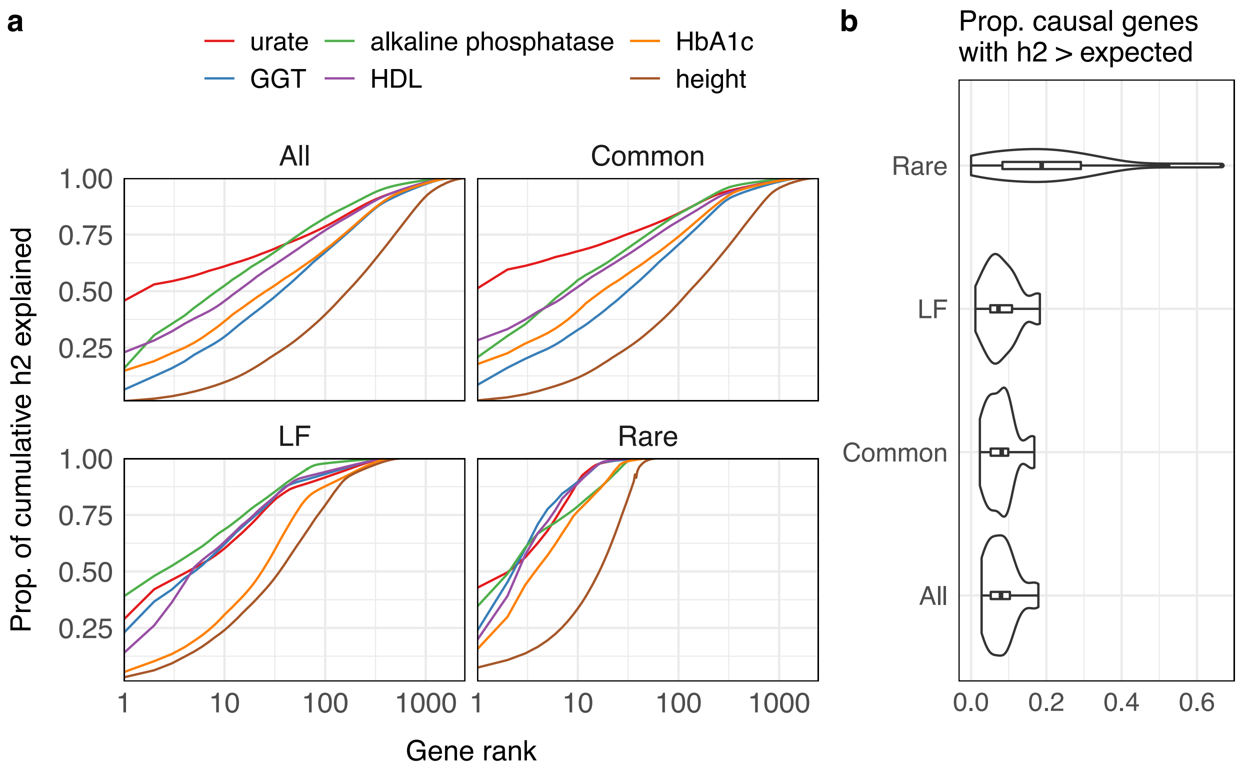

Figure 4.7 Total and MAF-partitioned gene-level h^2 capture differences polygenicity across traits. (a) Empirical distributions of cumulative heritability for six example traits (clockwise from top left: total, common, low-frequency, and rare). Each curve can be read as, “the top X genes explain Y% of the cumulative gene-level heritability for a given trait.” Cumulative gene-level h2 is estimated by summing the estimated posterior means for nonzero-h2 genes (90%-CI > 0). (Supplementary Figure 21 shows all 25 traits.) (b) Proportion of nonzero-h2 genes per trait with disproportionately large heritability estimates, defined as genes with 90%-CI > (cumulative heritability / number of causal genes)). Each violin plot represents 25 traits.

4.6 Tables

Trait	$h^2_{gene,t} > 0$	$\geq \frac{1}{2}\sum h^2_{gene,t}$	(%)	$h^2_{gene,t} = h^2_{gene,c}$	$h^2_{gene,t} = h^2_{gene,lf}$	$h^2_{gene,t} = h^2_{gene,r}$
Corneal Hysteresis	1212	42	3.5%	912	82	4
Hair Color	1328	6	0.5%	972	92	14
BMD Heel T-score	1430	48	3.4%	1098	90	4
Alkaline Phosphatase	1695	9	0.5%	1257	120	20
SHBG	1699	5	0.3%	1277	118	19
MCH	1701	41	2.4%	1253	137	18
C-reactive Protein	1702	5	0.3%	1293	98	7
apoA-I	1730	14	0.8%	1290	119	14
Platelet Distribution Width	1736	19	1.1%	1316	117	20
MSCV	1738	38	2.2%	1339	118	11
Urate	1744	2	0.1%	1319	119	14
Monocyte Count	1750	41	2.3%	1332	112	10
HDL	1766	14	0.8%	1321	126	11
GGT	1784	37	2.1%	1361	108	13
HbA1c	1813	26	1.4%	1345	145	17
High Light Scatter Reticulocyte Count	1858	56	3.0%	1399	129	25
IGF1	1859	62	3.3%	1402	128	12
Body Mass Index (BMI)	1879	184	9.8%	1430	116	8
Cystatin C	1900	22	1.2%	1452	121	9
Platelet Count	1910	64	3.4%	1471	119	25
Forced Vital Capacity	1910	157	8.2%	1465	123	6
Mean Platelet Volume	1912	32	1.7%	1408	140	25
RBC Count	1915	89	4.6%	1461	138	21
Basal Metabolic Rate	2099	181	8.6%	1608	128	11
Height	2469	168	6.8%	1860	182	32

Table 4.1 Number of nonzero-*h*^{2} genes identified (90%-CI), the spread of *h*^{2} signal across genes, and the relative contributions of different MAF classes for 25 quantitative traits in the UK Biobank. Nonzero-*h*^{2} genes have (i) $h^2_{\mathrm{gene,t}}$ 90%-CI > 0 and (ii) 90%-CI > 0 for at least one MAF bin. Columns 3-4: number (and %) of nonzero-*h*^{2} genes that explain at least 50% of cumulative $h^2_{\mathrm{gene,t}}$. Columns 5-7: numbers of genes with *h*^{2} exclusively from common, low-frequency, or rare variants

.

5 References

1. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
3. Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* **10**, 1579 (2019).
4. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
5. Uricchio, L. H. Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Hum. Genet.* **139**, 5–21 (2020).
6. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
7. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
9. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
10. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
11. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
12. Metspalu, A. The Estonian genome project. *Drug Dev. Res.* **62**, 97–101 (2004).
13. Leitsalu, L. *et al.* Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
14. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
15. Hirata, M. *et al.* Overview of BioBank Japan follow-up data in 32 diseases. *J. Epidemiol.* **27**, S22–S28 (2017).
16. Wei, C.-Y. *et al.* Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ Genom. Med.* **6**, 10 (2021).
17. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
18. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
19. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

20. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

21. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

22. Gharahkhani, P. *et al.* Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* **12**, 1258 (2021).

23. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

24. Matoba, N. *et al.* GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nat. Hum. Behav.* **4**, 308–316 (2020).

25. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

26. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

27. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14* (ACM Press, 2014). doi:10.1145/2649387.2660800.

28. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

29. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

30. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).

31. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

32. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49**, 1304–1310 (2017).

33. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

34. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).

35. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).

36. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).

37. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).

38. Campbell, M. C. & Tishkoff, S. A. The evolution of human genetic and phenotypic variation in Africa. *Curr. Biol.* **20**, R166-73 (2010).

39. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science* **259**, 639–646 (1993).

40. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208-15 (2010).

41. Laland, K. N., Odling-Smee, J. & Myles, S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat. Rev. Genet.* **11**, 137–148 (2010).

42. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* **9**, e1003566 (2013).

43. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

44. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

45. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

46. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9**, 96 (2017).

47. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).

48. Chen, C.-Y., Han, J., Hunter, D. J., Kraft, P. & Price, A. L. Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. *Genet. Epidemiol.* **39**, 427–438 (2015).

49. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

50. Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genetic Epidemiology* vol. 43 180–188 (2019).

51. Shi, H. *et al.* Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).

52. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

53. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

54. O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).

55. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).

56. Freund, M. K. *et al.* Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).

57. Burch, K. S. *et al.* Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *bioRxiv* 2021.08.17.456722 (2021) doi:10.1101/2021.08.17.456722.

58. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

59. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

60. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

61. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

62. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

63. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

64. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

65. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).

66. Wu, Y. & Sankararaman, S. A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics* **34**, i187–i194 (2018).

67. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).

68. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107 Suppl 1**, 1752–1756 (2010).

69. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).

70. Uricchio, L. H., Kitano, H. C., Gusev, A. & Zaitlen, N. A. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol. Lett.* **3**, 69–79 (2019).

71. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

72. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).

73. Gamazon, E. R., Cox, N. J. & Davis, L. K. Structural architecture of SNP effects on complex traits. *Am. J. Hum. Genet.* **95**, 477–489 (2014).

74. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).

75. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

76. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88**, 365–411 (2004).

77. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).

78. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206-13 (2015).

79. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).

80. Yengo, L. *et al.* Imprint of assortative mating on the human genome. *Nat. Hum. Behav.* **2**, 948–954 (2018).

81. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5272-81 (2014).

82. Weissbrod, O., Flint, J. & Rosset, S. Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.* **103**, 89–99 (2018).

83. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).

84. Elman, R. S. & Karpenko, N. *The algebraic and geometric theory of quadratic forms*. (American Mathematical Society, 2008). doi:10.1090/coll/056/03.

85. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

86. Pasaniuc, B. *et al.* Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).

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

88. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

89. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).

90. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).

91. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, (2019).

92. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

93. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).

94. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).

95. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).

96. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

97. Ng, M. C. Y. *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet.* **10**, e1004517 (2014).

98. Franceschini, N. *et al.* Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am. J. Hum. Genet.* **93**, 545–554 (2013).

99. Schick, U. M. *et al.* Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.* **98**, 229–242 (2016).

100. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

101. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 30–35 (2016).

102. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).

103. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).

104. Kraft, P., Zeggini, E. & Ioannidis, J. P. A. Replication in genome-wide association studies. *Stat. Sci.* **24**, 561–573 (2009).

105. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).

106. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

107. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet.* **9**, e1003379 (2013).

108. Asimit, J. L. *et al.* Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nat. Commun.* **10**, 3216 (2019).

109. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).

110. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).

111. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

112. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 1080 (2019).

113. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

114. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

115. Ikeda, M. *et al.* Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. *Schizophr. Bull.* **45**, 824–834 (2019).

116. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).

117. Guo, J. *et al.* Quantifying genetic heterogeneity between continental populations for human height and body mass index. *Sci. Rep.* **11**, 5240 (2021).

118. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

119. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

120. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

121. Shi, H., Pasaniuc, B. & Lange, K. L. A multivariate Bernoulli model to predict DNaseI hypersensitivity status from haplotype data. *Bioinformatics* **31**, 3514–3521 (2015).

122. Dai, B., Ding, S. & Wahba, G. Multivariate Bernoulli distribution. *Bernoulli (Andover.)* **19**, 1465–1483 (2013).

123. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

124. CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).

125. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

126. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

127. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

128. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

129. Johnson, R., Shi, H., Pasaniuc, B. & Sankararaman, S. A unifying framework for joint trait analysis under a non-infinitesimal model. *Bioinformatics* **34**, i195–i201 (2018).

130. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* **16**, e1008612 (2020).

131. Hormozdiari, F. *et al.* Widespread Allelic heterogeneity in complex traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).

132. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).

133. Chen, C.-H. *et al.* Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum. Mol. Genet.* ddw346 (2016) doi:10.1093/hmg/ddw346.

134. Cano-Gamez, E. & Trynka, G. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* **11**, 424 (2020).

135. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).

136. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022-1034.e6 (2019).

137. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, (2017).

138. Yao, C. *et al.* Dynamic role of trans regulation of gene expression in relation to complex traits. *Am. J. Hum. Genet.* **100**, 985–986 (2017).

139. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).

140. Caballero, A., Tenesa, A. & Keightley, P. D. The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. *Genetics* **201**, 1601–1613 (2015).

141. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).

142. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *Yearbook of Paediatric Endocrinology* (2019) doi:10.1530/ey.16.14.15.

143. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **95**, 126 (2014).

144. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

145. Hunt, K. A. *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498**, 232–235 (2013).

146. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).

147. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

148. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

149. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92**, 841–853 (2013).

150. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

151. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).

152. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455-64 (2014).

153. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).

154. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11**, e1005165 (2015).

155. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).

156. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

157. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).

158. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).

159. Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.* **34**, 463–468 (2010).

160. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **58**, 267–288 (1996).

161. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of genomes. *Nat. Commun.* **11**, 4020 (2020).

162. Benner, C., Havulinna, A. S., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and regional heritability. *bioRxiv* (2018) doi:10.1101/318618.

163. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

164. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-98 (2015).

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

166. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).

167. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

168. George, S. *et al.* A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science* **304**, 1325–1328 (2004).

169. Hussain, K. *et al.* An activating mutation of AKT2 and human hypoglycemia. *Science* **334**, 474–474 (2011).

170. Biderman, L., Manley, J. L. & Prives, C. Mdm2 and MdmX as regulators of gene expression. *Genes Cancer* **3**, 264–273 (2012).

171. Toufektchan, E. *et al.* Germline mutation of MDM4, a major p53 regulator, in a familial syndrome of defective telomere maintenance. *Sci. Adv.* **6**, eaay3511 (2020).

172. Mencarelli, M. A. *et al.* Evidence of digenic inheritance in Alport syndrome. *J. Med. Genet.* **52**, 163–174 (2015).

173. Mochizuki, T. *et al.* Identification of mutations in the alpha 3(IV) and alpha 4(IV) collagen genes in autosomal recessive Alport syndrome. *Nat. Genet.* **8**, 77–81 (1994).

174. Lemmink, H. H. *et al.* Benign familial hematuria due to mutation of the type IV collagen alpha4 gene. *J. Clin. Invest.* **98**, 1114–1118 (1996).

175. Badenas, C. *et al.* Mutations in theCOL4A4 and COL4A3 genes cause familial benign hematuria. *J. Am. Soc. Nephrol.* **13**, 1248–1254 (2002).

176. Kheradmand Kia, S. *et al.* RTTN mutations link primary cilia function to organization of the human cerebral cortex. *Am. J. Hum. Genet.* **91**, 533–540 (2012).

177. Shamseldin, H. *et al.* RTTN mutations cause primary microcephaly and primordial dwarfism in humans. *Am. J. Hum. Genet.* **97**, 862–868 (2015).

178. Rump, P. *et al.* Whole-exome sequencing is a powerful approach for establishing the etiological diagnosis in patients with intellectual disability and microcephaly. *BMC Med. Genomics* **9**, 7 (2016).

179. Shaheen, R. *et al.* Genomic and phenotypic delineation of congenital microcephaly. *Genet. Med.* **21**, 545–552 (2019).

180. Writzl, K. *et al.* De Novo mutations in SLC25A24 cause a disorder characterized by early aging, bone dysplasia, characteristic face, and early demise. *Am. J. Hum. Genet.* **101**, 844–855 (2017).

181. Ehmke, N. *et al.* De Novo Mutations in SLC25A24 Cause a Craniosynostosis Syndrome with Hypertrichosis, Progeroid Appearance, and Mitochondrial Dysfunction. *Am. J. Hum. Genet.* **101**, 833–843 (2017).

182. Alazami, A. M. *et al.* Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep.* **10**, 148–161 (2015).

183. Chong, J. X. *et al.* Recessive inactivating mutations in TBCK, encoding a Rab GTPase-activating protein, cause severe infantile syndromic encephalopathy. *Am. J. Hum. Genet.* **98**, 772–781 (2016).

184. Bhoj, E. J. *et al.* Mutations in TBCK, encoding TBC1-domain-containing kinase, lead to a recognizable syndrome of intellectual disability and hypotonia. *Am. J. Hum. Genet.* **98**, 782–788 (2016).

185. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418 (2002).

186. Grishkevich, V. & Yanai, I. Gene length and expression level shape genomic novelties. *Genome Res.* **24**, 1497–1503 (2014).

187. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437–442 (2008).

188. Anzai, N. *et al.* Plasma urate level is directly regulated by a voltage-driven urate efflux transporter URATv1 (SLC2A9) in humans. *J. Biol. Chem.* **283**, 26834–26838 (2008).

189. Caulfield, M. J. *et al.* SLC2A9 is a high-capacity urate transporter in humans. *PLoS Med.* **5**, e197 (2008).

190. Dinour, D. *et al.* Two novel homozygous SLC2A9 mutations cause renal hypouricemia type 2. *Nephrol. Dial. Transplant* **27**, 1035–1041 (2012).

191. Dinour, D. *et al.* Homozygous SLC2A9 mutations cause severe renal hypouricemia. *J. Am. Soc. Nephrol.* **21**, 64–72 (2010).

192. Matsuo, H. *et al.* Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *Am. J. Hum. Genet.* **83**, 795 (2008).

193. Weiss, M. J. *et al.* A missense mutation in the human liver/bone/kidney alkaline phosphatase gene causing a lethal form of hypophosphatasia. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7666–7669 (1988).

194. Sergi, C., Mornet, E., Troeger, J. & Voigtlaender, T. Perinatal hypophosphatasia: Radiology, pathology and molecular biology studies in a family harboring a splicing mutation (648+1A) and a novel missense mutation (N400S) in the tissue-nonspecific alkaline phosphatase (TNSALP) gene. *Am. J. Med. Genet.* **103**, 235–240 (2001).

195. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* **51**, 1349–1355 (2019).

196. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).

197. Younes, N. *et al.* A whole-genome sequencing association study of low bone mineral density identifies new susceptibility loci in the phase I Qatar Biobank cohort. *J. Pers. Med.* **11**, 34 (2021).

198. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

199. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).

200. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).

201. Bhatia, G. *et al.* Subtle stratification confounds estimates of heritability from rare variants. *bioRxiv* (2016) doi:10.1101/048181.

202. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).