

Curriculum Vitae

Diego Veliz-Otani, MSc, PhD

Graduate Research Assistant
Epidemiology and Human Genetics Program
University of Maryland School of Medicine
Email: diego.m.veliz.otani@gmail.com

I. Professional Summary

PhD in Molecular Epidemiology specializing in statistical and population genetics. Developed a coalescent-based framework (GeSi) to refine how distant and recent relatedness is modeled in genome-wide association studies of admixed populations, improving heritability estimates and phenotype predictions. Experienced in high-performance computing, R, Python, and statistical modeling, with a proven track record of analyzing large, diverse genomic datasets. I am interested in integrating machine learning approaches into genetic epidemiology studies, aiming to identify causal genetic variants and predict phenotypes more accurately.

II. Education

- 2025 University of Maryland School of Medicine – Baltimore, MD

PhD Epidemiology and Human Genetics - Molecular Epidemiology track

- 2017 Cayetano Heredia University – Lima, Peru.

MS Biochemistry and Molecular Biology

- 2013 National University of San Marcos – Lima, Peru.

BSc. Genetics and Biotechnology

III. Post Graduate Education and Training

- Jun 2022 Oxford Statistical Genomics Summer School

- Jul 2019 – Jun 2020 Fogarty Northern Pacific Global Health Scholar. University of Washington, Seattle.

- Jul 2019 – Sept 2019 Global Health Project Management. University of Washington, Seattle

- Jul 2016 Summer Institute in Statistical Genetics: 1) Advanced quantitative genetics; 2) Mixed models in quantitative genetics; 3) Statistical and quantitative genetics of disease; 4) Integrative genomics. University of Washington, Seattle.

- April 2014 – June 2014 Data analysis for Genomics. HarvardX (Harvard University on edX)

- Feb 2014 X Summer course in bioinformatics: An introduction to the use of R and Bioconductor. Universidade de São Paulo – Riberão Preto, Brazil.

IV. Key Skills

- Programming and Computational Tools: R, Python, Linux (Bash, Awk, sed), High-Performance Computing (SGE and SLURM), GitHub, Tidyverse, Conda, Jupyter.
- Genetic Epidemiology: GWAS, Polygenic Risk Score, BLUP, heritability estimation, phenotype simulation.
- Genomic data analysis: Variant calling, phasing, imputation, global and local ancestry.
- Coalescence-based sequence simulation (Msprime), coalescence modeling and Ancestral Recombination Graphs (tsinfer, relate).
- Applied Statistics: Cross-Validation, Bootstrapping, Permutation, Simulation, Benchmarking, Generalized Linear Mixed Models, Survival Analysis, Dimensionality Reduction, clustering and prediction methods.
- Scientific communication: manuscript writing, collaborative research
- Interdisciplinary teamwork and project management.

V. Professional Experience

a. University of Maryland School of Medicine – Baltimore, MD

- Developed the Coefficient of Genealogical Similarity (GeSi), a novel measure of genetic similarity grounded in coalescent theory that enables unified modeling of relatedness and population structure.
- Conducted comprehensive benchmarking of GeSi against standard methods (GRMs and PCA) in GWAS analyses and phenotypic prediction using the BLUP framework.
- Presented preliminary findings about GeSi at the American Society of Human Genetics 2024 annual meeting, earning the Reviewer's Choice Abstract recognition.
- Performed functional annotation of multi-ancestry GWAS meta-analysis results, contributing to the identification of novel loci associated with Parkinson's disease and emphasizing the value of diverse datasets in genetic epidemiology.
- Conducted extensive demographic modeling to infer parameters such as population growth, migration, and population splits, contributing to the development of a speciation model for the cattle parasite *Theileria parva*. Used advanced tools, including DaDi (2D Site Frequency Spectrum analysis) and PSMC (Pairwise Sequentially Markovian Coalescent), to analyze evolutionary patterns.

b. Ediciones SM S.A.C

Scientific reviewer | January – April 2019.

- Reviewed and identified scientific inaccuracies in high school science and technology textbooks, providing factually accurate corrections and clarifications to ensure content accuracy.

c. Neurogenetics Research Center, National Institute of Neurological Sciences – Lima, Peru

Research Associate | 2018 – 2020

- Directed the design, implementation, and oversight of human genetics research projects on monogenic and multifactorial disorders, ensuring compliance with IRB requirements. Coordinated data exchange with partner institutions, and prepared regular IRB reports to maintain research accountability and ethical compliance.
- Led the development and design of the first genome-wide association study (GWAS) of ischemic stroke in the Peruvian population as part of the Northern Pacific Global Health (NPGH) Fogarty Fellowship, addressing gaps in Latin American representation in GWAS cohorts.
- Mentored students doing research in human Genetics. Main supervisor of Diana Cubas, BSc., who defended her thesis comparing the accuracy of mixed linear models and survival models at predicting the age of onset of Huntington’s disease.

d. Universidad Peruana Cayetano Heredia – Lima, Peru

Lecturer | 2017 – 2018

- I taught courses in “Introductory Genetics” and “Molecular Genetics” for the School of Sciences, “Animal Breeding” for the School of Veterinary Sciences, and “Molecular Biology” for the School of Medicine.
- Developed and delivered course materials, integrating applied concepts in genetics and molecular biology to diverse student audiences.

e. Universidad Peruana Cayetano Heredia – Lima, Peru

Master’s student and Teaching Assistant, Molecular Biology | 2015 – 2017

- Conducted research on the genetics of alpaca fiber traits, identifying a locus under recent selective sweep strongly associated with fiber morphology (odds ratio = 5.7), providing new insights into the genetic basis of this economically important trait.
- Investigated the genetic determinants of alpaca fiber type (suri vs. huacaya), supporting the role of artificial selection in shaping fiber characteristics.
- Assisted in teaching “Introductory Genetics” and “Molecular Genetics” for the School of Sciences, “Animal Breeding” for the School of Veterinary Sciences, and “Molecular Biology” for the School of Medicine, contributing to course delivery and student learning.

VI. Selected Projects

- Coefficient of Genealogical Similarity (GeSi): Developed a novel measure of genetic similarity based on coalescent theory, addressing the limitations of conventional metrics like the

kinship coefficient, offering a unified framework for modeling relatedness and population structure.

- **Demographic Modeling of *Theileria parva*:** Developed a speciation model of *Theileria parva* using 2D-SFS (DaDi) and PSMC analyses. This entailed implementing and testing over 30 demographic models in DaDi, formally testing their goodness of fit and comparing them using Likelihood Ratio Tests. This analysis was complemented with estimation of their population effective size trajectory and estimation of cross-coalescence rate.
- **GLAD Project:** Performed pre-imputation quality control of genotype data, phasing, imputation, and post-imputation quality control for a dataset of 53,000+ Latin Americans. I also described patterns of population structure and calculated local and global ancestry.
- **Parkinson's GWAS Meta-Analysis:** Performed functional annotation of GWAS results, contributing to a multi-ancestry study that identified novel loci associated with Parkinson's disease.
- **Alpaca Fiber Genetics:** Uncovered the genetic basis of a commercially important trait of alpaca fiber, identifying a locus under a selective sweep that is strongly associated with hair morphology.

VII. Publications

- Borda, V., Loesch, D. P., Guo, B., Laboulaye, R., **Veliz-Otani, D.**, et al., (2024). Genetics of Latin American Diversity Project: Insights into population genetics and association studies in admixed groups in the Americas. *Cell genomics*.
- Kim, J. J., Vitale, D., **Veliz-Otani, D.**, et al., (2024). Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. *Nature genetics*, 56(1), 27-36.
- Cubas-Montecino D, Cornejo-Olivas M, Mazzetti P, Veliz-Otani D. Prediction of the age of onset of Huntington Disease using a Mixed Linear Model in a Peruvian Cohort. [cited 2024 Nov 27]; Available from: <https://ehdn.org/wp-content/uploads/2021/10/F04.pdf>
- Véliz-Otani, D., Cubas-Montecino, D., Milla-Neyra, K ... & Cornejo-Olivas, M. (2021). Response to ATXN10 Microsatellite Distribution in a Peruvian Amerindian Population. *The Cerebellum*, 20, 946-947.
- Véliz-Otani D, Inca-Martinez M, Bampi GB, Ortega O, Jardim LB, Saraiva-Pereira ML, et al. ATXN10 Microsatellite Distribution in a Peruvian Amerindian Population. *The Cerebellum*. 2019 Oct;18(5):841–8.
- Véliz Otani DM. Búsqueda de genes asociados a rasgos seleccionados para la producción de fibra de alpaca. 2017 [cited 2024 Nov 27]; Master's thesis; Available from: https://www.lareferencia.info/vufind/Record/PE_66b4172e87e33a5c9c1e0d6c3790b6f6

- Kay C, Tirado-Hurtado I, Cornejo-Olivas M, Collins JA, Wright G, Inca-Martinez M, et al. The targetable A1 Huntington disease haplotype has distinct Amerindian and European origins in Latin America. *Eur J Hum Genet.* 2017;25(3):332–40.
- Cornejo-Olivas M, Espinoza-Huertas K, Velit-Salazar MR, Veliz-Otani D, Tirado-Hurtado I, Inca-Martinez M, et al. Neurogenetics in Peru: clinical, scientific and ethical perspectives. *J Community Genet.* 2015 Jul;6(3):251–7.

VIII. Participation in Conferences

- Unifying population structure and relatedness analysis through a coalescent approach. Oral Presentation. Modeling and Theory in Population Biology Workshop. National Institute for Theory and Mathematics in Biology. June 2nd – June 6th. Chicago, Illinois.
- A unified genealogical measure of genetic similarity accounts for relatedness and admixture in GWAS. Poster presented at the 2024 American Society of Human Genetics Annual Meeting (ASHG 2024), Denver, Colorado.
- From Buffalo to Cattle: The Unidirectional Migration of *Theileria parva* and the Emergence of East Coast Fever. Poster presented at 2023 the Society for Molecular Biology and Evolution Annual Meeting (SMBE 2023), Ferrara, Italy.
- Ancestry-matching cases to controls from the GLAD database to accelerate genetic research in under-represented populations. Poster presented at the 2021 American Society of Human Genetics Annual Meeting (ASHG 2021), Virtual Meeting.

IX. Awards and Recognition

- Reviewer’s Choice Abstract – ASHG (2024)
- Northern Pacific Global Health Fellow (2019–2020)
- FONDECYT-CONCYTEC Full-tuition Scholarship to pursue a master’s in molecular biology (Apr 2015 – Mar 2017)

Abstract

Title of Dissertation: Unifying Population Structure and Relatedness Analysis through a Coalescent Approach

Diego Veliz-Otani, Doctor of Philosophy, 2025

Dissertation Directed by: Timothy D. O'Connor, Ph.D., Associate Professor, University of Maryland, Baltimore School of Medicine.

Genetic similarity in genome-wide association studies (GWAS) is typically partitioned into recent kinship, modeled by a genetic relationship matrix (GRM), and distant ancestry, corrected by principal components (PCs). In this dissertation, I argue that this partitioned model is a methodological practice built on a typically implicit causal framework that conflates population structure with confounding. This work deconstructs this standard approach and proposes a unified genetic model as a formal baseline.

To this end, I make two contributions. First, I introduce the Coefficient of Genealogical Similarity (GeSi), a measure of relatedness derived from coalescent theory that captures the full continuum of shared genealogy. This leads to a classification of genetic relationship matrices (GRMs) into genealogically “full” or “shallow” matrices. Empirical tests demonstrate that full GRMs are sufficient to model the genetic covariance from population structure in the absence of confounding. This reframes the role of PCs as proxies for unmeasured confounders correlated with ancestry, rather than as a necessary correction for population structure itself.

Second, I develop `phenocause`, an R package for simulating phenotypes under complex genetic and non-genetic causal models. This tool addresses a critical methodological gap by enabling the simulation of specific genetic and non-genetic

confounding scenarios which are necessary to test the assumptions of GWAS models. Together, these contributions provide a theoretical basis and a practical tool to move the field beyond correcting for inflation of test statistic and towards understanding the mechanisms that give rise to such inflation.

Unifying Population Structure and Relatedness Analysis through a Coalescent Approach

By
Diego Martin Veliz-Otani

Dissertation submitted to the Faculty of the School of Graduate Studies
of the University of Maryland, Baltimore in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Copyright 2025 By Diego Veliz-Otani
All rights Reserved

Table of Contents

1.	Introduction and Literature Review	1
1.1.	Background.....	1
1.1.1.	Classical Quantitative Genetics: Phenotypic Resemblance among Relatives	1
1.1.2.	Mixed Linear Model Association Analysis: Practical Utility of Relatedness.....	2
1.1.3.	Confounding and the Lack of a Formal Causal Framework in Genetic Epidemiology 4	
1.1.4.	The Continuum of Relatedness Under Coalescent Theory and Tree-Sequence Inference 9	
1.1.5.	Phenotype Prediction via the Best Linear Unbiased Predictor Equations	12
1.2.	Specific Aims	14
1.2.1.	Aim 1: To develop the Coefficient of Genealogical Similarity, a New Coalescent- Based Genetic Similarity Statistic that Measures both Recent and Distant Relatedness.	15
1.2.2.	Aim 2: To evaluate whether GeSi Can Account for Population Structure and Familial Relationship in GWAS.....	16
1.2.3.	Aim 3: To Develop A Flexible Phenotype Simulation Tool That Can Handle Different Genetic Architectures and Confounding Mechanisms.	17
2.	Aim 1: Derivation of the Coefficient of Genealogical Similarity.....	19
2.1.	Introduction.....	19
2.2.	Subjects and Data Sources	20
2.2.1.	Joint Variant Calling of Real Whole-Genome Sequence Data	20
2.2.2.	Exploratory Simulation (Simulation Set #1).....	23
2.3.	Methods.....	24
2.3.1.	Calculation of GeSi from Real and Simulated Data	24
2.3.2.	Estimation of Other Genetic Relationship Matrices (GRMs).....	25
2.3.3.	Exploratory analysis of GeSi's basic properties	26
2.4.	Results.....	26
2.4.1.	Derivation of the Coefficient of Genealogical Similarity (GeSi)	26
2.4.1.1.	GeSi as the correlation of the number of derived alleles carried by two diploid individuals 27	

2.4.1.2.	GeSitree as the Correlation of the Deviations from the Ancestral Phenotype due to Additive Genetic Effects	31
2.4.1.3.	Genotype-based Estimators of GeSitree	37
2.4.2.	Geometric Interpretation and Extension of the GeSi Model	41
2.4.3.	Exploration of the Basic Properties of Coefficient of Genealogical Similarity.....	44
2.5.	Discussion.....	48
2.6.	Strengths and implications.....	50
2.7.	Limitations	51
2.8.	Future directions	51
2.9.	Conclusions.....	52
3.	Aim 2: Using GeSi to Account for Population Structure and Relatedness in Genome Wide Association Studies.....	54
3.1.	Introduction.....	54
3.2.	Data and Methods	55
3.2.1.	BioMe cohort	55
3.2.2.	Phenotype simulation.....	55
3.2.2.1.	Effect of the minor allele frequencies on SNP effect sizes (α parameter)	56
3.2.2.2.	Simulation of confounding effects.....	57
3.2.2.3.	Simulations under a Latin American Demographic Model - Simulation Sets #2 and #3	58
3.2.2.4.	BioMe-Based Simulations (Simulation Sets #4 and #5).....	59
3.2.3.	Calculation of GRMs and Principal Components Analysis.....	60
3.2.4.	Using GeSi to Estimate Variance Components via Mixed Linear Models.....	62
3.2.5.	Mixed linear model association testing and heritability estimation.....	64
3.2.6.	Inference of the α parameter through CV-BLUP	66
3.3.	Results.....	67
3.3.1.	Fully Informative GRMs do not Require Principal Components to Adjust for Admixture in Single-Variant Association Tests in the Absence of Confounding	67
3.3.2.	Heritability does not Require Principal Components in the Absence of Environmental Confounding.....	79
3.3.3.	Heritability Estimation in the Presence of Simulated Confounding.....	84
3.3.4.	Phenotype via the Best Linear Unbiased Predictor Equation	86

3.3.5.	Power to Detect Population Structure via Principal Components	89
3.3.6.	Prediction of Human Height in Real Data from the BioMe Cohort	90
3.3.7.	Heritability Estimation of Human Height in the BioMe Cohort.....	93
3.4.	Discussion	101
3.5.	Strengths and Implications.....	104
3.6.	Limitations	105
3.7.	Future Directions	106
3.8.	Conclusions.....	106
4.	Aim 3 - Development of a Flexible Phenotype Simulation Tool.....	107
4.1.	Introduction.....	107
4.2.	Methods.....	109
4.2.1.	Design and Implementation of the Phenocause Package.....	109
4.2.1.1.	Implementation and Dependencies	110
4.2.1.2.	Modularity.....	110
4.2.1.3.	Input and Output Formats	110
4.2.1.4.	Wrapper-Based User Interface	111
4.2.2.	Module 1: Causal Sites Sampling (sample_causal_sites).....	111
4.2.2.1.	Pre-sampling Filters	111
4.2.2.2.	Uniform and Weighted Sampling Modes.....	112
4.2.2.3.	Collider Sampling Mode.....	112
4.2.3.	Module 2: Phenotype Simulation (simulate_phenotype).....	113
4.2.3.1.	Base Genetic Model.....	113
4.2.3.2.	Confounding Models	114
4.2.3.3.	Liability Threshold Model for Binary Traits	116
4.2.3.4.	Derivation of the Confounder Parameters	117
4.2.4.	Accompanying Data and Distribution	122
4.2.4.1.	Example Genetic Data Distributed with Phenocause	122
4.2.4.2.	Metadata and LD Weights.....	123
4.2.5.	Usage Demonstration and Validation.....	124
4.2.5.1.	Example 1: Sampling Causal Sites with Varying Levels of Linkage Disequilibrium	124

4.2.5.2.	Example 2: Confounder Correlated with the Polygenic Score	125
4.2.5.3.	Example 3: Binary Trait with a Categorical Confounder in Ordinal Scale	125
4.2.5.4.	Example 4: A Factorial Experiment to Isolate Collider Bias	126
4.3.	Results.....	128
4.3.1.	Example 1: Sampling Causal Sites with Varying levels of Linkage Disequilibrium 128	
4.3.2.	Example 2: Confounder Correlated with the Polygenic Score	129
4.3.3.	Example 3: Binary Trait with a Categorical Confounder in Ordinal Scale	130
4.3.4.	Example 4: A Factorial Experiment to Isolate Collider Bias	131
4.4.	Discussion.....	135
4.5.	Strengths and implications.....	137
4.6.	Limitations and future directions	138
4.7.	Conclusions.....	139
5.	Overall Conclusions of this Study	140
5.1.	Main Findings in this Study.....	140
5.2.	Strengths and Implications of this Study	141
5.3.	Limitations	141
5.4.	Future Directions	143
5.5.	Summary.....	144
6.	References.....	145

LIST OF TABLES

Table 1. Genotype-based estimator of the GeSitree coefficient.	40
---	----

List of Figures

Figure 1. Parameterization of the GeSi model.....	29
Figure 2. Detection of population structure via the genotype estimator of the coefficient of genealogical similarity.	46
Figure 3. Distribution of the GeSi values stratified by familial relationship and population.	47
Figure 4. Accuracy of true causal effects estimation in single-variant mixed linear model association tests in the Msprime-generated Simulation Set #2 ($\alpha = 0$).	71
Figure 5. Accuracy of true causal effects estimation in single-variant mixed linear model association tests in the Msprime-generated Simulation Set #3 ($\alpha = -1$).	72
Figure 6. Type I error rate and inflation of the chi-square statistic in Simulated Set #2 ($\alpha = 0$).	74
Figure 7. Type I error rate and inflation of the chi-square statistic in Simulated Set #3 ($\alpha = -1$).	75
Figure 8. Statistical power to detect a true causal variant in simulated set #2 ($\alpha = 0$).	77
Figure 9. Statistical power to detect a true causal variant in simulated set #3 ($\alpha = -1$).	78
Figure 10. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = 0$ (Simulation Set #2.)	80
Figure 11. Heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = 0$	81
Figure 12. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = -1$ (Simulation Set #3.)	82
Figure 13. Distribution of heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = -1$	83
Figure 14. Square error and heritability estimate in simulations with and without confounding.	85
Figure 15. Accuracy of the predicted genetic value (g) and phenotype (y) via BLUP using different mixed linear model specifications in Simulated Set #4.	88
Figure 16. Accuracy of the height prediction via BLUP in real data from the BiMe cohort.	92
Figure 17. Effect of the α parameter on the squared error of the heritability estimates.	95

Figure 18. Selection of the α parameter for heritability estimation of human height.....	96
Figure 19. Average heritability estimates of human height.	97
Figure 20. Group-specific heritability estimates of human height based on the α value that minimizes the AIC.	99
Figure 21. Group-specific heritability estimates of human height keeping α fixed at -1.	100
Figure 22. Example 1: Comparison of the LD-dependent and the uniform sampling of causal sites.	129
Figure 23. Example 2: Simulation of a confounder effect correlated with the polygenic score.	130
Figure 24. Example 3: Simulation of a binary trait with a categorical confounder in ordinal Scale.	131
Figure 25. Test statistic inflation when the causal variants are correlated with principal components.	134
Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components.	135

List of Abbreviations

- AFR: African or African American
- AIC: Akaike Information Criterion
- ARG: Ancestral Recombination Graph
- ASMC: Ascertained Sequentially Markovian Coalescent
- BLUP: Best Linear Unbiased Predictor
- BQSR: Base Quality Score Recalibration
- CV: Cross-Validation
- DP: Depth
- DTWF: Discrete-Time Wright-Fisher
- EUR: European American
- GDS: Genomic Data Structure
- GeSi: Coefficient of Genealogical Similarity
- GQ: Genotype Quality
- GRM: Genetic Relationship Matrix
- GWAS: Genome-Wide Association Studies
- HGDP: Human Genomes Diversity Project
- HMM: Hidden Markov Model
- IBD: Identical by Descent
- JC69: Jukes and Cantor (1969) model
- 1KGP: 1000 Genomes Project
- LAT: Latino or Latin American
- LD: Linkage Disequilibrium

- MAF: Minor Allele Frequency
- MAC: Minor Allele Count
- MCMC: Markov-Chain Monte Carlo
- MLM: Mixed Linear Model
- MXB: Mexican Biobank
- PC: Principal Component
- PCA: Principal Components Analysis
- PCoA: Principal Coordinates Analysis
- PD: Public Data
- PGP: Peruvian Genome Project
- PGS: Polygenic Score
- REML: Restricted Maximum Likelihood
- RMSE: Root of the Mean Squared Error
- RS: Randomly Selected sites
- SGDP: Simons Genome Diversity Project
- SNP: Single Nucleotide Polymorphism
- SRA: Sequence Read Archive
- WGS: Whole-Genome Sequence

1. Introduction and Literature Review

1.1. Background

1.1.1. Classical Quantitative Genetics: Phenotypic Resemblance among Relatives

The kinship coefficient is a fundamental concept in classical quantitative genetics, formally defined as the probability that two alleles, each randomly sampled from one of two individuals, are identical by descent (IBD). Two alleles are said to be IBD if they can be traced back to a common ancestor without mutating (1). The kinship coefficient $\phi_{(i,j)}$ between two individuals i, j thus reflects the expected proportion of their alleles that are IBD and can be used to quantify recent genetic similarity in a population.

Under additive genetic models, the phenotypic covariance of a trait between individuals i and j is given by $Cov(y_i, y_j) = 2\phi_{(i,j)}\sigma_A^2$, where σ_A^2 is the additive genetic variance component (2). Additional genetic components, such as dominance variance, can be included but are frequently omitted in large-scale analyses of complex traits. Historically, these methods relied on pedigree records, truncating distant ancestral lines beyond a baseline generation (3). This practice meant that only recent genealogical connections were considered, treating all more remote relationships as effectively zero kinship. Estimating the additive genetic variance is useful in practice because it allows us to quantify the proportion of the phenotypic variance due to genetic variation via the heritability parameters, defined as $h^2 := \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_e^2)}$, where σ_e^2 is the residual (or environmental) variance.

In practice, complete pedigrees are usually unavailable, so correlation-based estimators of the kinship coefficient are used instead. The genomic relationship matrices (GRMs)

employed in GCTA (4) or PC-Relate (5) approximate $\phi_{(i,j)}$ by measuring how frequently two individuals co-vary in their allele counts across the genome. For instance, GCTA's estimator weights each SNP by its allele frequency in the sample and computes a centered, scaled covariance of genotypes (5). PC-Relate refines this approach by adjusting for population structure by estimating individual-specific mean allele frequencies by regressing the genotypes on the first few principal components (5). Despite these improvements, the underlying perspective remains that pairwise phenotypic resemblance is captured primarily through more contemporaneous genetic variation, captured by the GRM, rather than deep ancestral coalescent events.

One consequence of focusing on recent relatedness is that, by design, any potential phenotypic similarity caused by sharing older causal alleles is left unmodeled. While such distant shared ancestry can influence trait distributions (1), classical quantitative genetics and modern correlation-based methods often treat it as background population structure rather than as a factor that needs to be regressed out of the phenotype covariance equation. This conceptual delimitation between “structure” and “relatedness” underlies many current approaches to association testing, heritability estimation, and phenotype prediction (see section 2.1.2).

1.1.2. Mixed Linear Model Association Analysis: Practical Utility of Relatedness

Mixed linear models (MLMs) provide a unified framework for association analyses where observations are not independent due to, for instance, relatedness (6). MLMs predict the total genetic values as random effects whose covariance can be modeled using the quantitative genetics principles outlined in the previous section. In typical notation, write

$Y = X\beta + Zu + e$, where Y is a $n \times 1$ vector of phenotypes, X is a $n \times p$ matrix of fixed-effect covariates (e.g. age, sex, etc.), β is the corresponding $p \times 1$ vector of fixed effects (regression coefficients), Z an incidence matrix (typically a $n \times n$ identity matrix) linking individuals to the random genetic effects u , and e is the vector of residuals (6).

Under this model, the random effects u are assumed to follow a multivariate normal distribution with mean 0 and variance-covariance matrix $G = A\sigma_A^2$, where A is the $n \times n$ matrix of coefficients of relationship, which equals twice the matrix of kinship coefficients. The residuals follow a distribution $e \sim N(0, I_n\sigma_e^2)$, although more complex structures can be adopted if needed (e.g., heteroskedasticity or population-specific residuals (7)). This setup allows one to decompose the total phenotypic variance into a genetic component σ_A^2 and a residual component σ_e^2 , enabling estimating heritability estimation (3,6). Modern approaches replace the pedigree-based coefficients of relationship A with a GRM calculated from genotype data (5,7).

Maximum likelihood or restricted maximum likelihood (REML) methods are typically used to estimate σ_A^2 , σ_e^2 , and fixed-effect coefficients β . By modeling genetic similarity in the random effect, MLMs adjust for the non-independence of individuals due to shared genetic factors. This is crucial in genome-wide association studies (GWAS), where failing to account for relatedness can inflate test statistics or bias effect-size estimates (8). Moreover, the MLM framework naturally incorporates multiple variance components, allowing, for example, different random effects to model environmental clustering.

When individuals come from a heterogeneous population, it is common to add principal components (PCs) as fixed effects in X to control for major axes of population structure

(9). This partitioning of genetic similarity into distant and recent relatedness has proven effective in practice. However, distinguishing ancestry from relatedness is artificial, since all haplotypes ultimately coalesce to a common ancestor. This observation motivates exploring methods that can unify recent familial relationships with distant ancestry via coalescent-based approaches.

1.1.3. Confounding and the Lack of a Formal Causal Framework in Genetic

Epidemiology

In genetic epidemiology, population structure is typically treated as a confounder because it can inflate the distribution of the test statistic (10). Shared ancestry creates genetic covariance between a non-causal marker and unlinked causal loci, which can induce spurious genotype-phenotype associations and an excess of low p-values across the genome. This inflation is a well-documented phenomenon, and methods such as Genomic Control were developed to measure and correct for its genome-wide effects (10). Failing to account for population structure can lead to an increased rate of false positives (8).

The inflation of the test statistics can be further exacerbated when non-genetic factors are correlated with ancestry. Formally, confounding arises when a variable that is associated with the exposure of interest has a causal effect on the studied outcome through a path that does not depend on the exposure being tested (11). In genetic association studies, formal confounding would arise if an environmental or social variable is both causal for the phenotype and correlated with allele frequency. This phenomenon induces a non-causal association between a genetic variant and the trait of interest. An additional complication is ascertainment bias, in which the sampling process creates an artificial association

between phenotype and ancestry, for instance, by over-representing certain ethnic groups among cases or controls in binary-trait studies (8,12,13).

The standard approach to control for both sources of test statistic inflation is to include principal components (PCs) as fixed-effect covariates in a mixed linear model (14). The rationale is that the top PCs serve as proxies for the major axes of population structure (9). By including them as covariates in the model, any portion of the phenotype that is correlated with this structure, whether arising from genetic covariance or from environmental factors, is regressed out, thus correcting the association test for the variant of interest. This PC-based adjustment is often used in combination with a genetic relationship matrix (GRM), which models the phenotypic covariance arising from cryptic relatedness in the random effects component of the model (15,16).

This practical approach to controlling inflation, however, is not grounded in a formal causal model. The focus has often been on the statistical correction of population structure as a nuisance parameter, rather than on modeling the specific causal mechanism that leads to test statistic inflation. Standard methods aim to “correct for stratification” by including PCs as covariates, but it is often unclear what specific features of genetic stratification are being addressed (17). This practice treats population structure as a monolithic source of bias to be statistically removed, rather than as a biological signal of shared ancestry that can be correlated with other, true confounders (18). Consequently, the field lacks an explicit widely adopted framework for distinguishing inflation caused by true non-genetic (i.e. environmental, social, etc.) confounding from that caused by the genetic covariance inherent to structured populations (10).

This lack of formalism leads to causally imprecise interpretations of mathematically rigorous models. A prominent example is the interpretation of test statistic inflation caused by population structure. While a mathematically sound model, such as genomic control, can correctly quantify this inflation, it is often imprecisely labeled as “confounding” (10). This misattributes a true component of the genetic signal (the covariance arising from shared ancestry) as a statistical bias to be eliminated, rather than as a biological signal to be modeled and investigated. This imprecise language is not only semantic; it obscures the true causal mechanisms and can lead to flawed analytical goals. For example, another foundational method in the field, LD-score regression, explicitly groups population structure and cryptic relatedness into a single parameter labeled as “confounding bias” (19).

The partitioning of genetic effects into recent and distant relatedness is an implicit unjustified causal model. The standard mixed linear model in GWAS that includes both a genetic relationship matrix (GRM) as a random effect and principal components (PCs) as fixed effects inherently assumes that genetic effects can be separated into two different classes (4,14). This model implicitly treats recent kinship as a source of valid phenotypic covariance, while it treats distant ancestry as a source of bias to be regressed out. This partitioning is a methodological convenience that lacks formal justification from a genealogical perspective.

A partitioned causal genetic model is not supported by the high conservation of genetic effect sizes across populations and the continuous nature of genetic variation. The premise of a partitioned genetic causal model contradicts empirical evidence that the biological effects of causal variants are stable across deep ancestral genealogies. The correlation of

causal effect sizes between European and African ancestries is nearly perfect ($\rho=0.98$) in 47 out of 53 traits studied (20). Furthermore, the poor transferability of polygenic scores between populations can be largely explained by statistical differences in LD and allele frequency (21). Moreover, human genetic diversity is not organized into discrete clusters but is clinal and largely correlated with geographical distance (22–24). The perception of discrete clusters is often an artifact of biased sampling that over-represents geographic extremes and the misinterpretation of clustering software (24). Modern, diverse biobanks reveal a complex continuum of admixed ancestries that fills the gaps between previously defined reference populations, showing that participants exhibit gradients of genetic variation rather than distinct clusters (25). This evidence provides a strong biological basis against the partitioned causal genetic model that remains standard in the field.

The absence of a formal model leads to two fundamental errors in practice: the introduction of statistical bias and the misinterpretation of genetic signals. First, the standard method of adjusting for population structure can actively introduce bias. When a PC is a common effect of a true causal variant and an unlinked neutral variant, adjusting for that PC induces a spurious association at the neutral locus via collider bias (26). Second, as previously discussed, the inflation of test statistics due to the genetic covariance from population structure is often misinterpreted as a confounding bias that must be removed, rather than a true biological signal that must be properly modeled (10).

Attempts to imbue mixed linear models with a more explicit and formal causal interpretation have been largely ignored. Some frameworks have attempted to formalize the sources of confounding, explicitly distinguishing between environmental confounding

and genetic confounding (17). Causal graphs have been used to more formally disentangle direct genetic effects, indirect genetic effects from relatives, and different forms of confounding (18). Despite these important contributions, the partitioned GRM+PC approach remains the default model in the field without explicitly stating the assumed underlying causal mechanisms.

Systematic benchmarking through simulation is required to test the validity and limitations of the standard GRM+PC mixed linear model. Only in data with known causal components is it possible to separate inflation driven by genetic covariance from that caused by environmental confounding and thus quantify when the standard adjustment succeeds or fails.

While several simulation tools are available, they are not designed to formally test the specific causal assumptions underlying GWAS correction strategies. Msprime is the standard for simulating genealogies under complex demographic histories (27), and tstrait can be used to simulate phenotypes from them (28); however, tstrait is limited to simple genetic architectures and cannot implement non-genetic confounding correlated with ancestry. Forward-time simulators like SLiM offer great flexibility for modeling evolutionary processes such as selection but are computationally prohibitive for GWAS-scale datasets and lack built-in confounding modes (29). Other tools like SIMER can model non-additive genetic effects but do not allow for the simulation of confounders explicitly correlated with ancestry (30). Finally, modules in widely used packages like GCTA and LDAK can simulate phenotypes, but only under oversimplified models that do not challenge the field's conflation of structure and confounding (5,31).

No public tool yet exists to simulate phenotypes under a flexible causal model that can specify both complex genetic architectures and explicit confounding scenarios. The field therefore lacks a framework to test the conditions under which the standard GRM+PC mixed linear model is necessary or sufficient.

1.1.4. The Continuum of Relatedness Under Coalescent Theory and Tree-Sequence

Inference

Coalescent theory treats the ancestry of every pair of haplotypes as a continuous branching process in which they eventually converge on a common ancestor (32,33). With recombination, each segment of the genome follows its own local genealogy, leading to the concept of the Ancestral Recombination Graph (ARG) (34). In principle, an ARG unifies *all* ancestral relationships, from distant population splits to recent familial relationships, in a single framework. However, reconstructing ARGs is a computationally intensive task, with existing methods often constrained by trade-offs between scalability and accuracy. Several methods have been recently developed that approximate the ARG as a sequence of trees spanning the regions between recombination breakpoints.

One highly efficient approach is *tsinfer* (35), which infers a *tree sequence* by iteratively building an ancestral scaffold and then matching samples against it. The method requires fully phased, biallelic haplotype data, along with ancestral-state information at each site. The pipeline follows three main steps: 1) Generate ancestors, creating a large pool of potential ancestral haplotypes derived from the sample data (while noting breakpoints for recombination); 2) match these ancestors to each other, producing an internal ancestor tree sequence that captures shared segments among them; and 3) match the final samples to the

ancestor tree sequence, assigning each sample haplotype a path of inheritance along local genealogies. During the matching steps, tsinfer balances the cost of recombination events (i.e., breakpoints in local genealogies) against the cost of multiple mutations per site, allowing some flexibility for sequencing errors or back mutations. Tsinfer assumes a single effective population size (N_e) for the entire sample and does not explicitly model subpopulations or admixture events. Tsinfer uses a novel compact encoding that allows for ARGs to be efficiently stored as a *tree sequence* (36).

Relate is another method for reconstructing local genealogies, designed to infer genome-wide genealogies for thousands of samples with high computational efficiency (37). Unlike tsinfer, which iteratively builds an ancestral scaffold, Relate constructs genealogical trees backward in time, estimating both the sequence of coalescent events and their associated times. This process begins with a position-specific distance matrix that quantifies the genetic similarity between haplotypes using a hidden Markov model (HMM) to account for local patterns of mutation and recombination. From this matrix, Relate iteratively constructs rooted binary trees that capture the genealogical relationships within each genomic region. Recombination events are inferred as changes in tree topology across adjacent regions.

To estimate coalescence times, Relate maps mutations onto tree branches and uses a Markov Chain Monte Carlo (MCMC) approach under a coalescent prior. This allows Relate to model historical changes in effective population size directly from the data. The method assumes panmixia within labelled subpopulations, enabling it to account for population structure by estimating cross-coalescence rates and demographic splits between

groups. Relate’s runtime scales linearly with genome size but quadratically with sample size, limiting its application to datasets with a few thousand individuals. Additionally, the accuracy of inferred genealogies may decrease in the presence of admixture or recent relatedness, as the method does not explicitly model these properties.

ARG-Needle provides a scalable approach for reconstructing ARGs in biobank-scale datasets (38). Unlike tsinfer’s haplotype-matching scaffold or Relate’s coalescent merging, ARG-Needle uses an iterative “threading” strategy to incorporate samples into the ARG while maintaining computational efficiency (39). It begins with genotype hashing to identify candidate relatives, followed by the Ascertained Sequentially Markovian Coalescent (ASMC) algorithm to estimate pairwise coalescence times (40). The sample is then threaded into the graph, connecting it to inferred ancestors while accounting for recombination events. A final normalization step ensures consistency across genealogical times.

In the UK Biobank dataset, ARG-Needle enabled the detection of ultra-rare variants with minor allele frequencies as low as 0.0007%, many of which were enriched for loss-of-function mutations (38). By modeling genealogical relationships, ARG-Needle provides an implicit framework for imputing unobserved variants and integrates them into a linear mixed model to enhance rare variant association testing. This approach addresses the limitations of traditional imputation by relying on genealogical structure instead of reference panels, uncovering novel genotype-phenotype associations and offering critical insights into the genetic architecture of complex traits.

The tree-sequence structure used by all these inference methods can in principle represent the continuous nature of relatedness by incorporating both recent and distant coalescent events into a unified genealogy. However, to achieve tractability, they each rely on assumptions (e.g. panmixia, simple or piecewise-constant N_e , biallelic sites with known ancestral alleles) that can break down in large, admixed cohorts or in the presence of close relatives. Identifying a method that exploits the full continuum of genealogical relationships able to account for recent relatedness and without partitioning the sample into homogeneous discrete subpopulations remains an open challenge.

1.1.5. Phenotype Prediction via the Best Linear Unbiased Predictor Equations

Best linear unbiased prediction (BLUP) is a method developed as part of the Mixed Linear Model framework developed by Henderson to predict total genetic effects by modeling the phenotypic covariance via the pairwise kinship coefficients (6). Using the MLM notation given before:

$$Y = X\beta + Zu + e$$

The vector of total genetic effects \hat{u} can be predicted via the equation:

$$\hat{u} = GZ^TV^{-1}(Y - X\hat{\beta})$$

Where $G = A\sigma_A^2$ is variance-covariance matrix of total additive genetic effects, and $V^{-1} = (G + R)^{-1}$, and $R = I_n\sigma_e^2$ is the variance-covariance matrix of residuals. Importantly, BLUP can predict the total genetic value of individuals without known phenotype, provided the pairwise relatedness values are known. Let n be the size of the training

sample, and m that of the test sample. Then, the matrices used in the BLUP equations are $G_{(m+n) \times (m+n)}$, $Z_{(m+n) \times n}^T$, and $V_{(n \times n)}^{-1}$ (3).

It should be noted, however, that predicting phenotypes via the BLUP equations requires making assumptions about the genetic architecture of the trait. For instance, the classical pedigree-based BLUP implicitly assumes that the phenotypic covariance is proportional to the amount of genome shared IBD, regardless of the nature of the mutations carried through each genomic segment. GCTA, in contrast, assumes that all the SNPs used to calculate the GRM explain the same amount of phenotypic variance, which implies that the effect size of any SNP j is inversely proportional to $\sqrt{2p_j(1-p_j)}$, where p_j is the minor allele frequency (MAF) of SNP j (5). Furthermore, because GCTA does not account for LD, causal variants located in genomic regions with high LD are repeatedly tagged by many more markers than causal variants in regions with low LD. Association between the distribution of SNP effect sizes and the distribution of background local LD would represent a direct violation of the GCTA model. LDAK is a genetic relationship matrix that generalizes the GCTA model by introducing two additional parameters that reflect the dependence of the distribution of causal effects on the MAF and local LD patterns (31). How violations of the GRM and mixed linear model assumptions affect the accuracy of BLUP-based predictions remains an open question.

The small effective population sizes and traditionally well-characterized pedigrees of domestic organisms have facilitated its application in animal and plant breeding (41). However, BLUP has been less successful when applied to datasets of unrelated human subjects (42). Human populations exhibit larger effective population sizes and lower levels

of linkage disequilibrium that can limit the amount of variation captured by a GRM (41). Furthermore, GWAS typically use principal components (PCs) to correct for population stratification (8). However, the PCA is often calculated by decomposing the GRM (5). Thus, PCs may absorb part of the genetic variance that a GRM would otherwise attribute to random genetic effects, thereby reducing the estimated predictive accuracy of BLUP. The extent to which this phenomenon affects predictions in admixed or structured cohorts is also an open question.

1.2. Specific Aims

Relatedness can be used to model the phenotypic covariance and thus predict the total additive genetic value and account for background genetic similarity among subjects in genome-wide association studies (GWAS)(6,43). In GWAS, the genetic similarity is typically partitioned into recent (familial relationships) and distant (shared ancestry) relatedness (4,14). Accounting for ancestry is meant to control for ascertainment bias and confounding, while accounting for recent relatedness ensures that the residuals remain independent (6,8,44,45).

In this work, I used a coalescent approach to develop the coefficient of genealogical similarity (GeSi), a measure of overall genetic similarity that represents both familial relationships and shared ancestry. I derive GeSi by modeling the distribution of the number of shared derived alleles given a realized genealogy. Next, I prove that GeSi can be calculated directly from genotype data without inferring the genealogy. I show that the eigen-decomposition of GeSi reflects population ancestry. I define deep and shallow genetic relationship matrices (GRMs) as those that measure the full genealogy and recent

relationships only, respectively. I argue that population structure and confounding are different phenomena and show that the phenotypic covariance due to both recent and ancestral relationships can be modeled via mixed linear models (MLMs) using GeSi or other full GRMs without using principal components (PCs), but that shallow GRMs do require principal components to account for population structure. I also show that, when environmental confounding is not correlated with the principal components, adding PCs to a MLM with a full GRM offers no benefit. Finally, in order to enable further research that explicitly distinguishes between the causal mechanisms of population structure and those of environmental confounding, I developed the R package *phenocause*, designed to simulate polygenic traits with complex genetic architectures and different modes of confounding.

1.2.1. Aim 1: To develop the Coefficient of Genealogical Similarity, a New

Coalescent-Based Genetic Similarity Statistic that Measures both Recent and Distant Relatedness.

Hypothesis: The genetic similarity under any level of structure and relatedness can be measured by a coalescence-based statistic. This statistic can be used to model pairwise the phenotypic correlation.

Approach: a) I will model the genetic similarity between two individuals as a function of the branch lengths of a coalescence tree (36). I will call the resulting statistic the *coefficient of genealogical similarity*, abbreviated as GeSi. b) I will use Msprime (27) and a demographic model of Latin America (46) to simulate a diverse panel of unrelated and related individuals. c) I will evaluate the distribution of the GeSi coefficient under different

levels of population structure and familial relationship to determine whether it captures both recent and distant relatedness. I will assess whether the PCoA decomposition of the GeSi matrix reveals population clustering in a diverse Latin American sample with high levels of indigenous American ancestry. d) I will develop a genotype-based estimator of GeSi that does not require inference of the ARG.

Outcome: GeSi, a new genealogy-based coefficient that measures overall genetic similarity given the full genealogy of a sample set.

1.2.2. Aim 2: To evaluate whether GeSi Can Account for Population Structure and Familial Relationship in GWAS.

Hypothesis: GeSi can correct for population structure and familial relationship and allows to estimate the narrow-sense heritability (h^2) in GWAS.

Approach: a) I will simulate a quantitative trait with different levels of heritability, polygenicity, and LDK's α parameter (31). b) I will calculate the pairwise GeSi matrix, as well as a GRM and principal components (PCs). c) Next, I will compare the performance of mixed linear model analysis GWAS including either a GeSi matrix, or a GRM and ten PCs. Specifically, I will calculate the accuracy and bias of the effect size estimates, as well as the statistical power, type-2 error rate, and the genomic inflation. d) I will compare the accuracy and bias of the heritability estimates using GeSi and GRM plus PCs. e) Finally, I will run a GWAS of height, body mass index, coronary artery disease and peripheral artery disease using the TOPMed (47) BioMe cohort (DbGAP accession number: phs001644, N = 12,054).

Outcome: Full characterization of the statistical performance of GeSi and kinship matrices as measures of genetic similarity in Mixed Linear Model Association Analysis GWAS.

1.2.3. Aim 3: To Develop A Flexible Phenotype Simulation Tool That Can Handle Different Genetic Architectures and Confounding Mechanisms.

Hypothesis: A flexible simulation framework can contribute to demonstrating that population structure by itself is not a confounder.

Approach: Develop an R package that implements a pipeline to simulate polygenic traits in two steps: first, the causal sites are sampled from a user-provided file containing the genotype data; followed by the simulation itself of the causal components. The sampling can be done with uniform probability or using sampling weights based on linkage disequilibrium (LD) scores or any other user-provided weights data. Alternatively, sites that are either correlated or uncorrelated with recent relatedness (e.g. population structure) can be oversampled. The genetic architecture of the trait is further governed by the number of causal sites, the heritability and the dependence of the distribution of causal effects on the minor allele frequency. Additional confounding components can be added to a baseline genetic effects-only model. The confounding effects can be defined by categorical variables in nominal (e.g. population labels, zip codes, etc.) or ordinal scale (socioeconomic status, educational attainment), or by quantitative variables correlated with the polygenic score or specific ancestry components. The user has control over how much phenotypic variance is explained by the confounding variables, and the correlation between confounding and genetic effects.

Outcome: A publicly available and documented software tool for advanced phenotypic simulation.

2. Aim 1: Derivation of the Coefficient of Genealogical Similarity

2.1. Introduction

A unified, continuous model of genetic relatedness is supported by the high conservation of genetic effect sizes across populations and the continuous nature of genetic variation. The premise of a unified genetic component is strengthened by empirical evidence that the biological effects of causal variants are stable across deep ancestral genealogies. The correlation of causal effect sizes between European and African ancestries is nearly perfect ($\rho=0.98$) in 47 out of 53 traits studied (20). Furthermore, the poor transferability of polygenic scores between populations can be largely explained by statistical differences in LD and allele frequency (21). Moreover, human genetic diversity is not organized into discrete clusters but is clinal and largely correlated with geographical distance (10–12). The perception of discrete clusters is often an artifact of biased sampling that over-represents geographic extremes and the misinterpretation of clustering software (12). Modern, diverse biobanks reveal a complex continuum of admixed ancestries that fills the gaps between previously defined reference populations, showing that participants exhibit gradients of genetic variation rather than distinct clusters (13). Thus, the conservation of the underlying biology across populations and the continuous nature of genetic variation across populations provides a strong biological basis for a continuous causal genetic model. Such a model could serve as a null hypothesis whose rejection would require further investigation to determine which specific assumptions were violated. In this aim, I derive the coefficient of Genealogical Similarity (GeSi), a unified measure of genetic similarity

that is applicable regardless of population structure and familial relationship. Importantly, I explicitly state the assumptions required for it to be valid.

2.2. Subjects and Data Sources

I used both real and simulated genotype data to assess the basic properties of the coefficient of Genealogical Similarity (GeSi). The data encompassed varying degrees of relatedness and heterogeneity.

2.2.1. Joint Variant Calling of Real Whole-Genome Sequence Data

I ran a variant-calling pipeline on high-coverage whole genome sequence (WGS) data from the Peruvian Genome Project (PGP), complemented with selected subsets of public data (PD) in order to maximize the genetic diversity and the number of callable sites. I call this dataset PGP+PD. The public datasets were the 1000 Genome Project (1KGP), the Human Genomes Diversity Project (HGDP), and the Simons Genome Diversity Project (SGDP). The total sample size was 1,702, distributed in the following way:

- PGP (n=147):
 - Coastal Peruvian (n=46): Trujillo (n=16) and Moches (n=30).
 - Andean Peruvian (n=73): Chopccas (n=30), Cusco (n=16), and Uros (27).
 - Amazonian Peruvian (n=28): Iquitos (n=16), Matzes (n=12).
- 1KGP (n=1417):
 - Gambian in Western Division-Mandinka (GWD, n = 112).
 - Luhya in Webuye, Kenya (LWK, n = 98).
 - Yoruba in Ibadan, Nigeria (YRI, n = 108).

- Colombian in Medellin (CLM, n = 93).
 - Mexican ancestry in Los Angeles (MXL, n = 64).
 - Peruvian in Lima (PEL, n = 121).
 - Puerto Rican in Puerto Rico (PUR, n = 103).
 - Han Chinese in Beijing, China (CHB, n = 102).
 - Han Chinese in South China (CHS, n = 105).
 - Japanese in Tokyo (JPT, n = 104).
 - Utah residents with ancestry from Northern and Western Europe (CEU, n = 99).
 - Finnish in Finland (FIN, n = 97).
 - Iberian populations in Spain (IBS, n = 105).
 - Tuscan in Italy (TSI, n = 106).
- HGDP (n = 115):
- Africa: Bantu in Kenya (n=10), Bantu in South Africa (n=3), Bantu Ovambo (n=1), Biaka (n=22), Mbuti (n=10).
 - Americas: Karitiana (n=9), Maya (n=19), Pima (n=12), Surui (n=6).
 - East Asia: Yakut (n=23).
- SGDP (n = 23):
- Africa: Igbo (n=2), Kongo (n=1), Lemande (n=2), Ju|'hoan North (n=2).
 - Americas: Chipewyan (n=2), Cree (n=2), Nahua (n=2).
 - South Asia: Kashmiri Pandit (n=1), Kharia (n=1), Kurumba (n=1), Mala (n=1), Onge (n=2).
 - East Asia: Sherpa (n=2), Tibetan (n=2).

A joint variant-calling pipeline was implemented to process the samples from all four sources. The HGDP FASTQ files were obtained from the BioProject PRJEB6463 at the NCBI's Sequence Read Archive (SRA) using the fasterq-dump tool. Raw 30x WGS 1KGP FASTQ files were retrieved via IBM Aspera FASP from the ENA public run directories (/vol1/run/) corresponding to BioProject PRJEB31736/ERP114329.

The FASTQ files were then quality-filtered using fastp (48) (version 0.23.2) in default mode to remove low-quality bases, short reads, and reads with too many uncalled bases (Ns). Filtered reads were then aligned to the GRCh38 reference genome using bwa-mem (49) (version 0.7.17-r1188) with read group header line information incorporated via the bwa -R parameter. For samples with multiple FASTQ sets due to multiple sequencing runs, aligned BAM files were merged using samtools merge (50) (version 1.11) before further processing.

Joint genotyping was performed using an adapted pipeline based on the UK Biobank (51) and NYGC (52) workflows. Base quality score recalibration (BQSR) was done using GATK with recognized known sites from dbSNP v.138 and Mills Gold Standard (53) mapped to the GRCh38 reference human genome downloaded from the Broad Institute's public GATK resource bundle (54). Picard (55) (v.2.25.3) was used to ensure the correct association of paired-end reads, and samtools for sorting and marking duplicates. GATK's (v.4.2.6.1) HaplotypeCaller was run in GVCF mode for each sample, producing a single gVCF with genotype likelihood information for every site. gVCF files were generated with GATK's GenomicsDBImport and GenotypeGVCFs modules, splitting each chromosome into 2Mb intervals (keeping shorter segments at chromosome ends), with 1000 bp of

interval padding. Genotyping was done in batches of 50 individuals. Failed interval runs were retried by increasing memory or using a different GATK version, following UK Biobank WGS pipeline recommendations.

For variant recalibration, GATK VariantRecalibrator and ApplyVQSR were run with dbSNP, HapMap, and 1KGP high-confidence files for SNPs and Mills and dbSNP files for INDELS. The truth sensitivity thresholds were set to 99.8% for SNPs and 99% for INDELS. Finally, bcftools (50) (v.1.15.1) was used to keep only PASS variants and to convert genotypes with depth (DP) lower than 10 and genotype quality (GQ) lower than 20 to missing data.

2.2.2. Exploratory Simulation (Simulation Set #1)

Msprime (27,56) was used to simulate genealogical trees and whole-genome sequence data using a demographic model of Latin American and reference continental populations (46). Specifically, the backward-in-time Discrete-Time Wright-Fisher model (57) (DTWF) was used to generate data resembling eight 1KGP populations: Chinese in Beijing (CHB), Colombians in Medellin (CLM), Iberians in Spain (IBS), a hypothetical Ancestral Indigenous Mexicans (MXB), Mexicans in Los Angeles (MXL), Peruvians in Lima (PEL), Puerto Ricans (PUR), and Yoruba in Ibadan, Nigeria (YRI). Each population consisted of 60 individuals, including 30 unrelated samples, five full-sib pairs, five half-sib pairs, and five first-cousin pairs. Three human-sized chromosomes (chr20-chr22) were simulated using the corresponding GRCh38 recombination maps, a mutation rate of 1.25×10^{-8} per site per generation, and the Jukes and Cantor (JC69) nucleotide substitution model (58). Phenotypes were not simulated for this dataset.

2.3. Methods

2.3.1. Calculation of GeSi from Real and Simulated Data

The derivation of the GeSi equations is presented in the results section. Here, I describe how GeSi was estimated from the genotype and genealogical data. GeSi can be calculated using either centered or non-centered genotype data, and allows for different allele frequency-dependent scaling through the parameter α (see derivation in the results section). I refer to the original definition of GeSi, based on genealogical trees, as $GeSi_{tree}$, and to its genotype-based estimator as \widehat{GeSi}_{tree} . This estimator uses non-centered genotype data and sets the scaling parameter to $\alpha = 0$. Throughout this manuscript, the non-centered genotype data will be used only with $\alpha = 0$; every other value of α will be used only with centered genotype data, and the resulting GeSi statistic will be referred to as $GeSi_{\alpha}$ or simply GeSi if the context is clear.

$GeSi_{tree}$ was defined based on a single genealogical tree (see derivation in the results section), however, the genealogy of DNA sequences that have undergone recombination cannot be represented by a single tree. Thus, I calculated the genome-wide $GeSi_{tree}$ value as the average of all trees along the genome weighted by the tree span, i.e. the length of the genomic region covered by a tree. \widehat{GeSi}_{tree} , the genotype-based estimator of $GeSi_{tree}$, was calculated using all biallelic SNPs in whole-genome sequence (WGS) data. Later in aim 2, I test the effects of different ways of processing the genotype data (e.g. by removing sites in linkage disequilibrium).

The genome-wide $GeSi_{tree}$ values were calculated in python, by looping through the trees simulated via Msprime. Its genotype-based estimator, \widehat{GeSi}_{tree} , was calculated in R by

efficiently reading small site-wise chunks of genotype data stored in GDS (Genomic Data Structure) format using the function `seqBlockApply` from the `SeqArray` package. Missing genotypes were mean-imputed as they were read in small chunks, which kept the memory requirements low. Each chromosome was processed separately, leading to chromosome-specific GeSi matrices, which were then averaged into a genome-wide GeSi matrix using either the sequence length (\mathbf{GeSi}_{tree}) or number of segregating sites ($\widehat{\mathbf{GeSi}}_{tree}$) as chromosome weights.

Finally, $\mathbf{GeSi}_{\alpha=0}$ was calculated following the same pipeline that was used for $\widehat{\mathbf{GeSi}}_{tree}$, except that the genotype data was centered.

2.3.2. Estimation of Other Genetic Relationship Matrices (GRMs)

The GCTA and PC-Relate matrices were calculated using to serve as references against whom to compare GeSi. Sites with minor allele frequency (MAF) lower than 0.01 or missingness rate above 0.02 were filtered out. Samples with missingness rate above 0.02 were also discarded. Sites in high linkage disequilibrium were pruned using PLINK (59) v.1.9 with the option `--indep-pairwise` with a window size set to 400 SNPs with steps of 50 SNPs, and `r2` threshold of 0.1.

The GCTA matrix was calculated using the `gcta` with the `--grm-gz` option (5). Estimating the PC-Relate (4) GRM required a series of iterative steps. An initial GRM was calculated via the KING-robust method (60) as implemented in `SNPRelate::snpGdsIBDKING` (i.e. the `snpGdsIBDKING` function in the R package `SNPRelate` (61)). Next, the principal components were calculated via the PCAiR method implemented in `GENESIS::pcair` (7)

using the KING-robust matrix as input, with a kinship threshold of $(1/2)^{4.5}$ and divergence threshold of $-(1/2)^{4.5}$, corresponding to the maximum kinship coefficient value for third-degree relatives (62). Next, the PC-Relate kinship matrix was calculated via the GENESIS::pcrelate function, using the first four principal components calculated with PCAiR in the previous step as covariates. A second and final round of PCAiR was run, using the PC-Relate kinship estimates as input and keeping the kinship threshold at $(1/2)^{4.5}$ and the divergence threshold at $-(1/2)^{4.5}$. The output of PCAiR included a list of unrelated and related individuals based on the PC-Relate kinship estimates and the specified kinship threshold.

2.3.3. Exploratory analysis of GeSi's basic properties

The concordance between \mathbf{GeSi}_{tree} and $\widehat{\mathbf{GeSi}}_{tree}$ was assessed by calculating the coefficient of the Pearson's correlation coefficient, by plotting them against each other, and by visually comparing heatmaps of the distribution of values calculated by the two approaches. After validating $\widehat{\mathbf{GeSi}}_{tree}$ as an estimator of \mathbf{GeSi}_{tree} , the distribution of $\widehat{\mathbf{GeSi}}_{tree}$ values for related and unrelated pairs of subjects was visualized through violin plots stratified by ancestry pairs.

2.4. Results

2.4.1. Derivation of the Coefficient of Genealogical Similarity (GeSi)

In this section, I use tree branch statistics to express the correlation in the number of derived alleles carried by two individuals. I then demonstrate that this branch statistics-based coefficient represents the pairwise phenotypic correlation between individuals under the

assumption that $\alpha = 0$, i.e., that the causal variants' effect sizes are independent of their minor allele frequency. Then, I demonstrate that the pairwise cosine similarity of the genotype vectors is an estimator of this branch-based statistic, and thus of the Pearson correlation coefficient of the phenotype values.

2.4.1.1. GeSi as the correlation of the number of derived alleles carried by two diploid individuals

To model the correlation between the number of mutations carried by two diploid individuals, I first model the covariance in the number of derived alleles carried by two haplotypes given their realized genealogy, and the variance in the number of mutations carried by any haplotype. I then use these components to model the correlation between the number of derived alleles carried by two diploid individuals.

– Covariance of the Number of Derived Alleles Carried by Two Haplotypes

Let i and j be two nodes representing the genealogy of any two haplotypes (**Figure 1a**), and let N_i, N_j be two random variables (r.v.) representing the number of mutations they have accumulated since they diverged from the ancestral sequence represented by the root of the tree. Let $T_{ij}^{(s)}$ be the time between the root of the tree and the time when the genealogies of i and j diverged and let $T_{ij}^{(u)}$ be the time since their genealogies diverged until present time. Finally, let $T^{(R)}$ denote the time from present time to the root of the tree. I call $T_{i,j}^{(s)}$ the shared divergence time of haplotypes i, j , and $T_{ij}^{(u)}$ their unshared divergence time. Let $N_{i,j}^{(s)}$ be a random variable (r.v.) representing the number of mutations that

occurred during time $T_{ij}^{(s)}$; and $N_i^{(u)}$ and $N_j^{(u)}$ two r.v. representing the number of mutations that occurred during time $T_{ij}^{(u)}$ on the haplotypes represented by nodes i and j . Define:

$$N_i = N_{ij}^{(s)} + N_i^{(u)}$$

$$N_j = N_{ij}^{(s)} + N_j^{(u)}$$

Then:

$$Cov(N_i, N_j) = Cov\left[\left(N_{ij}^{(s)} + N_i^{(u)}\right), \left(N_{ij}^{(s)} + N_j^{(u)}\right)\right]$$

I assume that the number of mutations arising on a haplotype follows a Poisson distribution given by the length L of the sequence, the mutation rate ν , and the divergence time. Thus:

$$N_i \sim \text{Poisson}(\lambda_{ij}^{(s)} + \lambda_{ij}^{(u)})$$

$$N_j \sim \text{Poisson}(\lambda_{ij}^{(s)} + \lambda_{ij}^{(u)})$$

Where $\lambda_{ij}^{(s)} = \nu L T_{ij}^{(s)}$ and $\lambda_{ij}^{(u)} = \nu L T_{ij}^{(u)}$

The number of mutations occurring on haplotypes i and j after their genealogies diverged are independent of each other. Thus, the covariance in the number of mutations carried by any two haplotypes is given by $Cov(N_i, N_j) = Cov\left[N_{ij}^{(s)}, N_{ij}^{(s)}\right] = Var\left(N_{ij}^{(s)}\right)$. Because $N_{ij}^{(s)}$ follows a Poisson distribution, $Var\left(N_{ij}^{(s)}\right) = \lambda_{ij}^{(s)}$, and:

$$Cov(N_i, N_j) = \nu L T_{ij}^{(s)} \quad \text{Equation (1)}$$

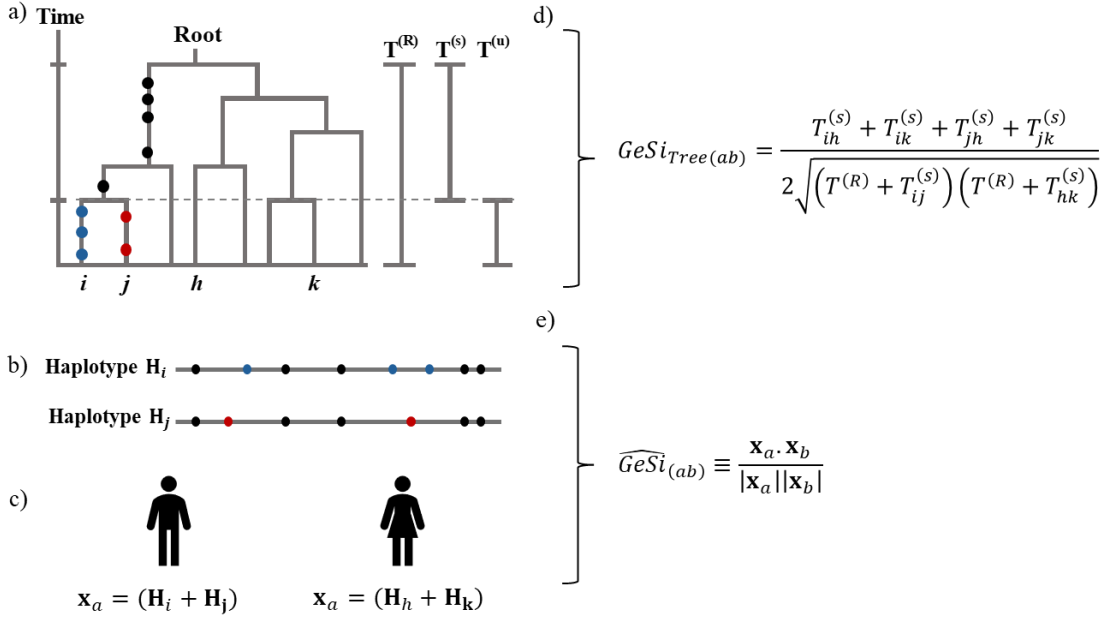


Figure 1. Parameterization of the GeSi model . **a)** Genealogy of four arbitrary haplotypes i, j, h and k . The circles on the tree represent the mutations carried by haplotypes i and j as they occurred across time; mutations on other haplotypes are not shown. $T^{(s)}$: Shared divergence time between haplotypes i and j . $T^{(u)}$: Unshared divergence time between haplotypes i and j . $T^{(R)}$: Time from the present to the root of the tree. **b)** Schematic showing the distribution of the mutations across either or both haplotypes i, j depending on when they occurred. The mutations marked with a black circle are shared by both haplotypes. The mutations marked with a blue or red circle occurred after their genealogies diverged and are exclusively found on either haplotype. **c)** Genotype vectors \mathbf{x} expressed as the sum of the haplotype vectors \mathbf{H} . The genotype vectors of two subjects carrying the pair of haplotypes (i, j) and (h, k) are indexed by the subscripts a and b , respectively. Thus, the genotype vectors are the sum of the haplotype vectors: $\mathbf{x}_a = (\mathbf{H}_i + \mathbf{H}_j)$ and $\mathbf{x}_b = (\mathbf{H}_h + \mathbf{H}_k)$. **d)** and **e)** connection between the time parameters and the GeSi coefficient and its estimator, respectively.

– Covariance of the number of mutations carried by two diploypes

Let $\{ij\}$ and $\{hk\}$ denote two diploypes formed by the pairs of haplotypes i, j and h, k , respectively. The number of mutations carried by them is given by $N_{\{ij\}} = N_i + N_j$, and $N_{\{hk\}} = N_h + N_k$. Then the covariance of the number of mutations carried by them is:

$$\text{Cov}(N_{\{ij\}}, N_{\{jk\}}) = \text{Cov}[(N_i + N_j), (N_h + N_k)]$$

$$= \text{Cov}(N_i, N_h) + \text{cov}(N_i, N_k) + \text{Cov}(N_j, N_h) + \text{Cov}(N_j, N_k)$$

Using equation (**Error! Reference source not found.**), this expression reduces to:

$$\text{Cov}(N_{\{i,j\}}, N_{\{h,k\}}) = vLT_{ih}^{(s)} + vLT_{ik}^{(s)} + vLT_{jh}^{(s)} + vLT_{jk}^{(s)} \quad \text{Equation (2)}$$

– Correlation of the Number of Mutations Carried by Two Diploid Individuals

By definition:

$$\text{Cor}(N_{\{i,j\}}, N_{\{h,k\}}) = \frac{\text{Cov}(N_{\{ih\}}, N_{\{hk\}})}{\sqrt{\text{Var}(N_{\{ij\}}) \times \text{Var}(N_{\{hk\}})}} \quad \text{Equation (3)}$$

To obtain the denominator, note that $N_{\{ij\}} = N_i + N_j$, and $N_{\{hk\}} = N_h + N_k$. Note that

$T_{ii}^{(s)} = T^{(R)}$ for any haplotype i . Thus:

$$\begin{aligned} \text{Var}(N_{\{i,j\}}) &= \text{Var}(N_i) + \text{Var}(N_j) + 2\text{Cov}(N_i, N_j) \\ &= 2vL(T^{(R)} + T_{ij}^{(s)}) \\ &= 2vL(T_R + T_{hk}^{(s)}) \end{aligned}$$

And analogously:

$$\text{Var}(N_{\{h,k\}}) = 2vL(T_R + T_{hk}^{(s)}) \quad \text{Equation (4)}$$

Replace Equations (2) and **Error! Reference source not found.**(4) into Equation (3):

$$\text{GeSi}_{Tree} := \text{Cor}(N_{\{i,j\}}, N_{\{h,k\}})$$

$$\begin{aligned}
&= \frac{vL(T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)})}{2vL\sqrt{(T^{(R)} + T_{ij}^{(s)})(T^{(R)} + T_{hk}^{(s)})}} \\
&= \frac{T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)}}{2\sqrt{(T^{(R)} + T_{ij}^{(s)})(T^{(R)} + T_{hk}^{(s)})}} \quad \text{Equation (5)}
\end{aligned}$$

A disadvantage of $GeSi_{Tree}$ is that it requires knowing the true genealogy. Later I demonstrate how to calculate it from genotype data without reconstructing the tree.

2.4.1.2. $GeSi_{tree}$ as the Correlation of the Deviations from the Ancestral Phenotype due to Additive Genetic Effects

I now demonstrate that $GeSi_{tree}$ also represents the correlation of the deviation of genetics effects from that of a hypothetical ancestor. To model the deviation from the ancestral phenotype due to total genetic effects, focus on the genetic variants with a causal effect on a trait. Thus, the parameters N_i , N_j , N_h , and N_k now represent the number of causal variants carried by the respective haplotypes, and the λ parameters denote the rate at which causal mutations arise. It is assumed that the number of accumulated causal alleles is linear with time, and that there is no correlation between the genetic and non-genetic causal components of the phenotype. For now, it is also assumed that the effect sizes of all variants are sampled from the same distribution and are independent of the minor allele frequency, similar to the LDAK model with $\alpha = 0$.

– **Variance of the Phenotypic Deviation**

First, express the phenotypes as deviations from the phenotype of the ancestral haplotype.

Let $Y_{\{i,j\}}$ and $Y_{\{h,k\}}$ denote the phenotypes of two individuals carrying the diplotypes $\{i,j\}$ and $\{h,k\}$, respectively. Also define d_i as a deviation from the ancestral phenotypic value caused by any haplotype i . Then, the phenotypes can be expressed as:

$$Y_{\{ij\}} = y_0 + d_i + d_j + e_{\{ij\}}$$

$$Y_{\{hk\}} = y_0 + d_h + d_k + e_{\{hk\}}$$

Where y_0 is the phenotypic value of a hypothetical individual carrying two copies of the ancestral haplotype. The goal is to model the correlation between $Y_{\{i,j\}}$ and $Y_{\{h,k\}}$ in terms of the divergence.

Assume an additive model and define d_i , for any haplotype i , as the sum of the causal effect sizes “a”:

$$d_i = \sum_{m=1}^{N_i} a_m \quad \text{Equation (6)}$$

Assume the effect sizes follow some distribution with mean μ_a , and variance σ_a^2 and apply the law of total variance to find the variance of d in Equation (6) **Error! Reference source not found.**

$$\text{Var}(d_i) = E[\text{Var}(d_i|N_i)] + \text{Var}(E[d_i|N_i])$$

Simplify the first term on the right-hand side by assuming that the effect sizes of two variants on a haplotype are uncorrelated:

$$\begin{aligned}
 \text{Var}(d_i) &= E[\sigma_a^2 N_i] + \text{Var}(\mu_a N_i) \\
 &= \sigma_a^2 \lambda_i + \mu_a^2 \lambda_i = \lambda_i (\sigma_a^2 + \mu_a^2) \\
 &= vL(\sigma_a^2 + \mu_a^2) T^{(R)}
 \end{aligned}$$

This expression holds true for any time interval during which the number of mutations is linear with time. Thus, in general:

$$\text{Var}(d) = vL(\sigma_a^2 + \mu_a^2) T \quad \text{Equation (7)}$$

Where T is any arbitrary time interval where this assumption holds true, and d is the phenotypic deviation caused by the mutations that occurred during that time.

– Covariance of the Phenotypic Deviation Terms

The next step is to model the covariance between the phenotypic divergence caused by two haplotypes i and j . Use the same approach as in section 2.4.1.1 and express each deviation term as the sum of shared and unshared deviations between any pair of haplotypes i and j :

$$d_i = d_{ij}^{(s)} + d_i^{(u)} = \sum_{m=1}^{N_{ij}^{(s)}} a_m + \sum_{q=1}^{N_i^{(u)}} a_q$$

$$d_j = d_{ij}^{(s)} + d_j^{(u)} = \sum_{m=1}^{N_{ij}^{(s)}} a_m + \sum_{q=1}^{N_j^{(u)}} a_q$$

Where the $d_{ij}^{(s)}$ represents the shared phenotypic divergence due shared mutations $N_{ij}^{(s)}$ between haplotypes i and j , and $d_i^{(u)}$ and $d_j^{(u)}$ represent the unshared phenotypic divergences of haplotypes i and j due to unshared mutations $N_i^{(u)}$ and $N_j^{(u)}$, respectively.

The covariance between d_i and d_j is given by:

$$\begin{aligned}
Cov(d_i, d_j) &= Cov \left[\left(\sum_{m=1}^{N_{ij}^{(s)}} a_m + \sum_{q=1}^{N_i^{(u)}} a_q \right), \left(\sum_{m=1}^{N_{ij}^{(s)}} a_m + \sum_{q=1}^{N_j^{(u)}} a_q \right) \right] \\
&= Var \left(\sum_{m=1}^{N_{ij}^{(s)}} a_m \right) + Cov \left[\left(\sum_{m=1}^{N_{ij}^{(s)}} a_m \right), \left(\sum_{q=1}^{N_j^{(u)}} a_q \right) \right] + \\
&\quad Cov \left[\left(\sum_{q=1}^{N_i^{(u)}} a_q \right), \left(\sum_{m=1}^{N_{ij}^{(s)}} a_m \right) \right] + Cov \left[\left(\sum_{q=1}^{N_i^{(u)}} a_q \right), \left(\sum_{q=1}^{N_j^{(u)}} a_q \right) \right] \\
&= Var \left(\sum_{m=1}^{N_{ij}^{(s)}} a_m \right) \\
&= Var(d_{i,j}^{(s)}) \tag{Equation (8)}
\end{aligned}$$

The covariance terms were cancelled under the assumption that the phenotypic deviation of two haplotypes from the ancestral phenotype is independent of each other once their genealogies have diverged. By replacing Equation (7) into Equation (8) and using the appropriate time parameter, obtain:

$$Cov(d_i, d_j) = vL(\sigma_a^2 + \mu_a^2)T_{i,j}^{(s)} \tag{Equation (9)}$$

Thus, the covariance of phenotypic deviations is equal to the variance of the shared deviation.

– **Correlation of the Phenotype of two Diploid Individuals**

To calculate the correlation, first calculate the variance and the covariance. Start by modeling the variance of the phenotype of a diploid individual. Remember the definition of the phenotypic deviations due to genetic effects:

$$d_{\{ij\}} = y_0 + d_i + d_j + e_{\{i,j\}}$$

Then:

$$\begin{aligned} \text{Var}(d_{\{ij\}}) &= \text{Var}(d_i + d_j) \\ &= \text{Var}(d_i) + \text{Var}(d_j) + 2\text{Cov}(d_i, d_j) \end{aligned}$$

Plug in the results from Equation (7) and Equation (9), to see that the variance is given by:

$$\begin{aligned} \text{Var}(d_{\{ij\}}) &= 2\nu L T^{(R)} (\sigma_a^2 + \mu_a^2) + 2\nu L T_{ij}^{(S)} (\sigma_a^2 + \mu_a^2) \\ &= 2\nu L (\sigma_a^2 + \mu_a^2) (T^{(R)} + T_{ij}^{(S)}) \end{aligned} \quad \text{Equation (10)}$$

And the covariance of $d_{\{i,j\}}$ and $d_{\{h,k\}}$ is given by:

$$\begin{aligned} \text{Cov}(d_{\{ij\}}, d_{\{hk\}}) &= \text{Cov}(y_0 + d_i + d_j + e_{\{i,j\}}, y_0 + d_h + d_k + e_{\{h,k\}}) \\ &= \text{Cov}(d_i, d_h) + \text{Cov}(d_i, d_k) + \text{Cov}(d_j, d_h) + \text{Cov}(d_j, d_k) \\ &= \nu L (\sigma_a^2 + \mu_a^2) (T_{ih}^{(S)} + T_{ik}^{(S)} + T_{jh}^{(S)} + T_{jk}^{(S)}) \end{aligned} \quad \text{Equation (11)}$$

Finally, to obtain the correlation, standardize the covariance by the square root of the product of the variances:

$$\begin{aligned}
Cor(d_{\{ih\}}, d_{\{hk\}}) &= \frac{Cov(d_{\{ij\}}, d_{\{hk\}})}{\sqrt{Var(d_{\{ij\}})Var(d_{\{hk\}})}} \\
&= \frac{vL(\sigma_a^2 + \mu_a^2)(T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)})}{\sqrt{(2vL(\sigma_a^2 + \mu_a^2)(T^{(R)} + T_{ij}^{(s)}))(2vL(\sigma_a^2 + \mu_a^2)(T^{(R)} + T_{hk}^{(s)})}} \\
&= \frac{vL(\sigma_a^2 + \mu_a^2)(T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)})}{2vL(\sigma_a^2 + \mu_a^2)\sqrt{(T^{(R)} + T_{ij}^{(s)})(T^{(R)} + T_{hk}^{(s)})}} \\
&= \frac{(T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)})}{2\sqrt{(T^{(R)} + T_{ij}^{(s)})(T^{(R)} + T_{hk}^{(s)})}}
\end{aligned}$$

The expression above is the same as the definition of $GeSi_{Tree}$ given in Equation (5). Thus, $GeSi_{Tree}$ also represents the correlation of the deviation from the expected ancestral phenotype under the assumptions that the number of causal mutations carried by a haplotype is linear with time and that all causal alleles' effect sizes are sampled from the same distribution:

$$\begin{aligned}
Cor(d_{\{i,h\}}, d_{\{h,k\}}) &\equiv GeSi_{Tree} \\
&= \frac{T_{ih}^{(s)} + T_{ik}^{(s)} + T_{jh}^{(s)} + T_{jk}^{(s)}}{2\sqrt{(T^{(R)} + T_{ij}^{(s)})(T^{(R)} + T_{hk}^{(s)})}} \quad \text{Equation (12)}
\end{aligned}$$

2.4.1.3. Genotype-based Estimators of $GeSi_{tree}$

The goal now is to demonstrate that GeSi can be calculated directly from genotype data without inferring the genealogy of the samples.

First, note that the derivation of GeSi (Equations (5) and (12)), relied on a Poisson approximation of the number of shared mutations given the shared divergence time. Also note that each of the right-hand terms of Equation (11) represents the lambda parameter of such Poisson distributions:

$$\begin{aligned} Cov(N_{\{i,j\}}, N_{\{h,k\}}) &= vLT_{ih}^{(s)} + vLT_{ik}^{(s)} + vLT_{jh}^{(s)} + vLT_{jk}^{(s)} \\ &= \lambda_{ih}^{(s)} + \lambda_{ik}^{(s)} + \lambda_{jh}^{(s)} + \lambda_{jk}^{(s)} \end{aligned}$$

For a Poisson process, the observed mutation counts are both the maximum likelihood and the method-of-moments estimators of the lambda parameters. Thus, the realized number of mutations can be replaced in to obtain an estimator:

$$\widehat{Cov}(N_{\{i,j\}}, N_{\{h,k\}}) = N_{ih}^{*(s)} + N_{ik}^{*(s)} + N_{jh}^{*(s)} + N_{jk}^{*(s)} \quad \text{Equation (13)}$$

Where the $N^{*(s)}$ parameters have the same interpretation as the $N^{(s)}$ parameters before, except that they represent the realized counts of derived alleles rather than random variables.

The advantage of Equation (13) is that it can be calculated from whole genome sequence data without knowing the genealogical tree, provided the reference allele is set to the ancestral allele, and the alternate allele to the derived allele, and the analysis is limited to

biallelic sites. To see this, first express each of the $N^{*(S)}$ parameters as a sum of indicator variables:

$$N_{ih}^{*(S)} = \sum_{m=1}^p I_{ih,m}^*$$

Where m is the position along a sequence of length p , and I^* is an indicator variable that takes a value of 1 if both haplotypes i and h carry the derived allele at position m and takes the value 0 otherwise. Importantly, this expression holds true even under linkage disequilibrium (within the genomic region spanned by a single tree) because of the linearity of the summation operator. Apply the same expression to the other terms in Equation (13) and obtain:

$$\begin{aligned} \widehat{Cov}(N_{\{ij\}}, N_{\{hk\}}) &= \sum_{m=1}^p I_{ih,m}^* + \sum_{m=1}^p I_{ik,m}^* + \sum_{m=1}^p I_{jh,m}^* + \sum_{m=1}^p I_{jk,m}^* \\ &= \sum_{m=1}^p S_m^* \end{aligned} \quad \text{Equation (14)}$$

Where:

$$S_m^* := I_{ih,m}^* + I_{ik,m}^* + I_{jh,m}^* + I_{jk,m}^* \quad \text{Equation (15)}$$

S_m^* in Equation (15) is the number of shared derived alleles at a single site m . To facilitate calculation directly from genotype data, I introduce matrix notation and indexing by diploid subject and site instead of by haplotype. Let $\mathbf{X} \in \{0,1,2\}^{n \times p}$ denote the genotype matrix for n subjects across p sites, where entries $x_{a,m}$ represent the number of derived alleles

for diploid subject a at site m . Let subjects $a, b \in \{1, 2, \dots, n\}$ carry the diplotypes $\{ij\}$ and $\{hk\}$, respectively (see **Figure 1**). The realized total number of derived alleles carried by them is given by $N_{\{ij\}}^* = \sum_{m=1}^p x_{am}$, and $N_{\{hk\}}^* = \sum_{m=1}^p x_{bm}$. In **Table 1**, I tabulate both the pairwise number of shared derived alleles and the product of the genotypes coded as number of derived allele dosages and demonstrate that the left-hand side of equation (15) is numerically equivalent to $x_{am}x_{bm}$. Substitute this identity into Equation (14) and obtain:

$$\begin{aligned}
\widehat{Cov}(N_{\{ij\}}, N_{\{hk\}}) &= \sum_{m=1}^p S_m^* \\
&= \sum_{m=1}^p x_{am}x_{bm} \\
&= \mathbf{x}_a \cdot \mathbf{x}_b
\end{aligned}
\tag{Equation (16)}$$

Table 1. Genotype-based estimator of the $GeSi_{tree}$ coefficient. Tabulating the right-hand side of the values obtained by applying Equations (15) and (16) to a single site m , shows that both equations always yield the same value for biallelic sites in which the ancestral allele is set to the reference allele.

Subject a		Subject b		Eq. (15).	Eq. (16)
Status	Genotype = $g_{ij,m}$	Status	Genotype = $g_{hk,m}$	$S_m^* =$	
	x_{am}		x_{bm}	$I_{ih,m}^* + I_{ik,m}^* + I_{jh,m}^* + I_{jk,m}^* + x_{am} \cdot x_{bm}$	
Hom. Anc.	Anc/Anc	Hom.			
	0	Anc.	Anc/Anc	0	0
Hom. Anc.	Anc/Anc	Het.	Anc/Der	1	0
Hom. Anc.	Anc/Anc	Hom. Der.	Der/Der	2	0
Het.	Anc/Der	Het.	Anc/Der	1	1
Het.	Anc/Der	Hom. Der.	Der/Der	2	2
Hom. derived	Der/Der	Hom. Der.	Der/Der	2	4

Hom: Homozygous. Het: Heterozygous. Anc: Ancestral allele. Der: Derived allele. Eq.: Equation.

Where $\mathbf{x}_a, \mathbf{x}_b \in \{0,1,2\}^p$ are the genotype vectors for subjects a and b .

In Equation (5**Error! Reference source not found.**), $GeSi_{tree}$ was defined as the correlation of the number of derived alleles between two diploid individuals, and in equation (12) it was demonstrated that $GeSi_{tree}$ also represents phenotype correlation between two diploid individuals under specific assumptions stated before. The goal now is to estimate the $n \times n$ matrix \widehat{GeSi}_{tree} of estimated $GeSi_{tree}$ coefficients directly from the genotype matrix \mathbf{X} . The strategy is to first obtain the variance-covariance matrix $\widehat{\mathbf{C}}$, whose entry (ab) estimates the covariance term in the numerator of Equation (12). Then use the diagonal elements of $\widehat{\mathbf{C}}$, which estimate the variance terms in the denominator of Equation (12), to standardize the matrix. Equation (16**Error! Reference source not found.**) established that the (a, b) -th entry of the covariance matrix $\widehat{\mathbf{C}}$ is $\mathbf{x}_a \cdot \mathbf{x}_b$. Therefore, the full variance-covariance matrix is:

$$\widehat{\mathbf{C}} = \mathbf{X}\mathbf{X}^T$$

Let $\mathbf{D} = \text{diag}(\widehat{\mathbf{C}})$ be diagonal matrix of variances. The \widehat{GeSi}_{tree} matrix is then obtained by standardizing $\widehat{\mathbf{C}}$:

$$\begin{aligned} \widehat{GeSi}_{tree} &= \mathbf{D}^{-1/2} \widehat{\mathbf{C}} \mathbf{D}^{-1/2} \\ &= \mathbf{D}^{-1/2} \mathbf{X}\mathbf{X}^T \mathbf{D}^{-1/2} \end{aligned} \quad \text{Equation (17)}$$

2.4.2. Geometric Interpretation and Extension of the GeSi Model

The estimator \widehat{GeSi}_{tree} in Equation (17) represents a matrix of cosine similarities between the raw genotype vectors:

$$\widehat{GeSi}_{tree(ab)} \equiv \frac{\mathbf{x}_a \cdot \mathbf{x}_b}{|\mathbf{x}_a| |\mathbf{x}_b|} \quad \text{Equation (18)}$$

Thus, $\widehat{GeSi}_{tree(ab)}$ measures the cosine of the angle between the raw genotype vectors originating from the zero vector (i.e. the root's genotype vector). Thus, the genotype cosine similarity is an estimator of the Pearson correlation of the deviations from the ancestral phenotype due to genetic effects. Note that this differs from standard GRM methods, which use a weighted average of per-site genotype covariances and thus require LD-pruned genotype data.

The original derivation of GeSi, $GeSi_{tree}$, and its estimator \widehat{GeSi}_{tree} , modeled the Pearson correlation of the deviations from the ancestor's total genetic effects ($d_{\{ij\}} = g_a$) due to the total divergence time from the present to the root of the tree. Importantly, it relied on a Poisson distribution to model the variance that would be created if all possible lineages were followed during a given amount of time. However, study samples are not representative of all potential lineages and thus a sample variance will not match the variance parameters used by $GeSi_{tree}$.

To address this limitation and focus on the variance observed in an ascertained sample, I propose generalizing this geometric interpretation. I hypothesize that the genotype cosine similarity can model the Pearson correlation of genetic effects relative to reference points other than the ancestral origin. A natural reference for an ascertained sample is its centroid, defined by the mean genotype vector $\bar{\mathbf{x}}$. Importantly, because the phenotype is assumed to follow a purely additive model, centering the genotype matrix on the sample mean is mathematically equivalent to analyzing deviations from the sample's average genetic effect.

Let n denote the number of subjects in a very large set of samples representative of all potential lineages descending from the root of a tree. Let $S = \{1, \dots, n_S\}$ with $n_S < n$ denote the set of all ascertained subjects included in a study. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the raw genotype matrix of a very large sample representative of all potential lineages deriving from the root of a tree. Let $\mathbf{X}_S \in$

$\mathbb{R}^{n_s \times p}$ denote the genotype matrix of the ascertained subjects. Let $\bar{\mathbf{x}}_{asc} \in \mathbb{R}^p$ denote the vector of mean genotypes in the ascertained set. Let $\mathbf{W} = \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}_{asc}^T \in \mathbb{R}^{n \times p}$ be the full genotype matrix centered on the ascertained samples' mean genotype, and $\mathbf{W}_S \in \mathbb{R}^{n_s \times p}$ a subset of \mathbf{W} containing only the ascertained subjects. Let \mathbf{w}_a and \mathbf{w}_b , where $a, b \in \{1, \dots, n\}$, denote the a -th and b -th row of \mathbf{W} , respectively. Then the angle, and therefore the cosine similarity, between the displacement vectors $(\mathbf{x}_a - \bar{\mathbf{x}}_{asc})$ and $(\mathbf{x}_b - \bar{\mathbf{x}}_{asc})$ originating from the ascertained samples' centroid $\bar{\mathbf{x}}_{asc}$ in the original space \mathbf{X} are identical to the angle and cosine similarity between the vectors \mathbf{w}_a and \mathbf{w}_b originating from the origin in the centered space. Notably, $\widehat{\text{GeSi}}_{tree} = \mathbf{D}^{-1/2} \mathbf{X} \mathbf{X}^T \mathbf{D}^{-1/2}$, which represents the cosine similarity in the original space \mathbf{X} , is an estimator of the Pearson correlation of the total genetic effects measured as deviations from the ancestral phenotype at the root of the tree. Therefore, I hypothesize that $\widehat{\text{GeSi}}_{asc}$, the cosine similarity between the centered vectors in the \mathbf{W} space, are estimators of the Pearson correlation of genetic effects of the ascertained samples:

$$\widehat{\text{GeSi}}_{asc} = \mathbf{D}_W^{-1/2} \mathbf{W} \mathbf{W}^T \mathbf{D}_W^{-1/2} \quad \text{Equation (19)}$$

Where $\mathbf{D}_W = \text{diag}(\mathbf{W} \mathbf{W}^T)$. The GeSi coefficient for any two samples depends only on those samples' genotype vectors. Thus, GeSi can be calculated just for the ascertained samples by restricting the genotype matrix to contain only the ascertained samples:

$$\widehat{\text{GeSi}}_{asc,S} = \mathbf{D}_{W_S}^{-1/2} \mathbf{W}_S \mathbf{W}_S^T \mathbf{D}_{W_S}^{-1/2} \quad \text{Equation (20)}$$

Where $\widehat{\text{GeSi}}_{asc,S} \in \mathbb{R}^{n_s \times n_s}$ is an estimator of the deviations from the ascertained samples' mean due to genetic effects. The inherent dependence on the sample mean implies that the variance components are sample-specific with potential implications on the transferability of heritability components estimates across divergent populations.

I further extend GeSi to allow for differential contributions of different sites based on, e.g., minor allele frequency, such as is the case for the LDAK model. Let f_m denote the scaling factor for site m , where $m \in \{1, \dots, p\}$. To obtain an LDAK-like scaling, set $q_m = [2f_m(1 - f_m)]^{\alpha/2}$, where f_m is the alternate allele frequency at site m in the ascertained sample, and α an arbitrary scalar. Let $\mathbf{Q} = \text{diag}(q_1, \dots, q_p)$ be a diagonal matrix containing the scaling factors. Then the matrix of scaled genotypes is $\mathbf{Z}_S = \mathbf{W}_S \mathbf{Q}$. I hypothesize that if these scaling factors reflect the true causal relationship between genotype and phenotype, then the cosine similarity calculated in this scaled space is an estimator of the Pearson correlation of the genetic effects deviations from the ascertained samples' mean.

$$\widehat{\text{GeSi}}_{asc,S,\alpha} = \mathbf{D}_{\mathbf{Z}_S}^{-1/2} \mathbf{Z}_S \mathbf{Z}_S^T \mathbf{D}_{\mathbf{Z}_S}^{-1/2} \quad \text{Equation (21)}$$

Because in practice only the ascertained sample is accessible, and because typically only the GeSi values calculated from genotype data are of interest, for the rest of this manuscript I drop the subscripts related to the ascertainment process and refer to the estimator in Equation (21) as GeSi_α . For the specific cases where GeSi has been calculated from genealogical trees using the original definition in Equations (5) or (12), I refer to the estimated value as GeSi_{tree} .

2.4.3. Exploration of the Basic Properties of Coefficient of Genealogical Similarity

In order to validate GeSi's genotype based estimator, the GeSi values calculated from the true genealogical trees following Equation (5) were plotted against those obtained from the non-centered genotype-based estimator calculated according to Equation (17). The results, shown in **Figure 2a**, reveal a high concordance between both estimates ($r^2 > 0.999$). The heatmaps of both sets of values clustered by population labels also reveal a similar population structure patterns

(**Figure 2b** and **c**) further validating the use of the genotype-based estimator. Thus, for the rest of this document all GeSi values are estimated via a genotype-based estimator, unless otherwise specified.

GeSi's genotype-based estimator constitutes a cosine similarity (see Equation (18), therefore, I applied an arc-cosine transformation to the matrix of pairwise GeSi values and obtained a matrix of angles between the genotypes vectors that could be interpreted as a measure of distance. I used this distance matrix to run a principal coordinates analysis (PCoA) and plotted the first two Principal Coordinates (**Figure 2d**), which showed that the samples clustered by population of origin, with in turn clustered roughly together based on geographic proximity. In contrast, the results of a Principal Components Analysis (PCA) carried out with Plink offered little resolution (**Figure 2e**).

Next, I visualized the distribution of $GeSi_{tree}$ values stratified by population and familial relationship level (**Figure 3a**), and confirmed that GeSi is not only indicative of population structure, as shown in **Figure 2**, but also of recent relatedness. Furthermore, the distribution of GeSi values between intra-population pairs of unrelated subjects differed across populations (**Figure 3a**), highlighting that the basal genetic similarity between “unrelated” individuals is different, in concordance with the definition of $GeSi_{tree}$, which is inversely proportional to the genome-wide average coalescence time. Finally, I compared, via boxplots, the distribution of the mean-centered genotype-based estimate ($GeSi_{\alpha=0}$) against that of the kinship estimates obtained via widely used methods (**Figure 3b**). As expected, PC-Relate values were highly concordant with the expected kinship coefficients, which measure only recent relatedness, across all four familial relationship levels. In contrast, GeSi and GCTA yielded values that exceeded the expected value

for first cousins and half-siblings, and exhibited higher variance than PC-Relate among unrelated subjects. KING yielded highly negative values for unrelated pairs, but yielded accurate kinship estimates for all three tested relationship levels.

It should be noted that, by design, GeSi does not attempt to estimate the kinship coefficient, which is a measure of recent relatedness only. Instead, GeSi attempts to quantify the total relatedness across the full genealogy. Therefore, it is not surprising that the GeSi values did not coincide with the expected kinship estimates.

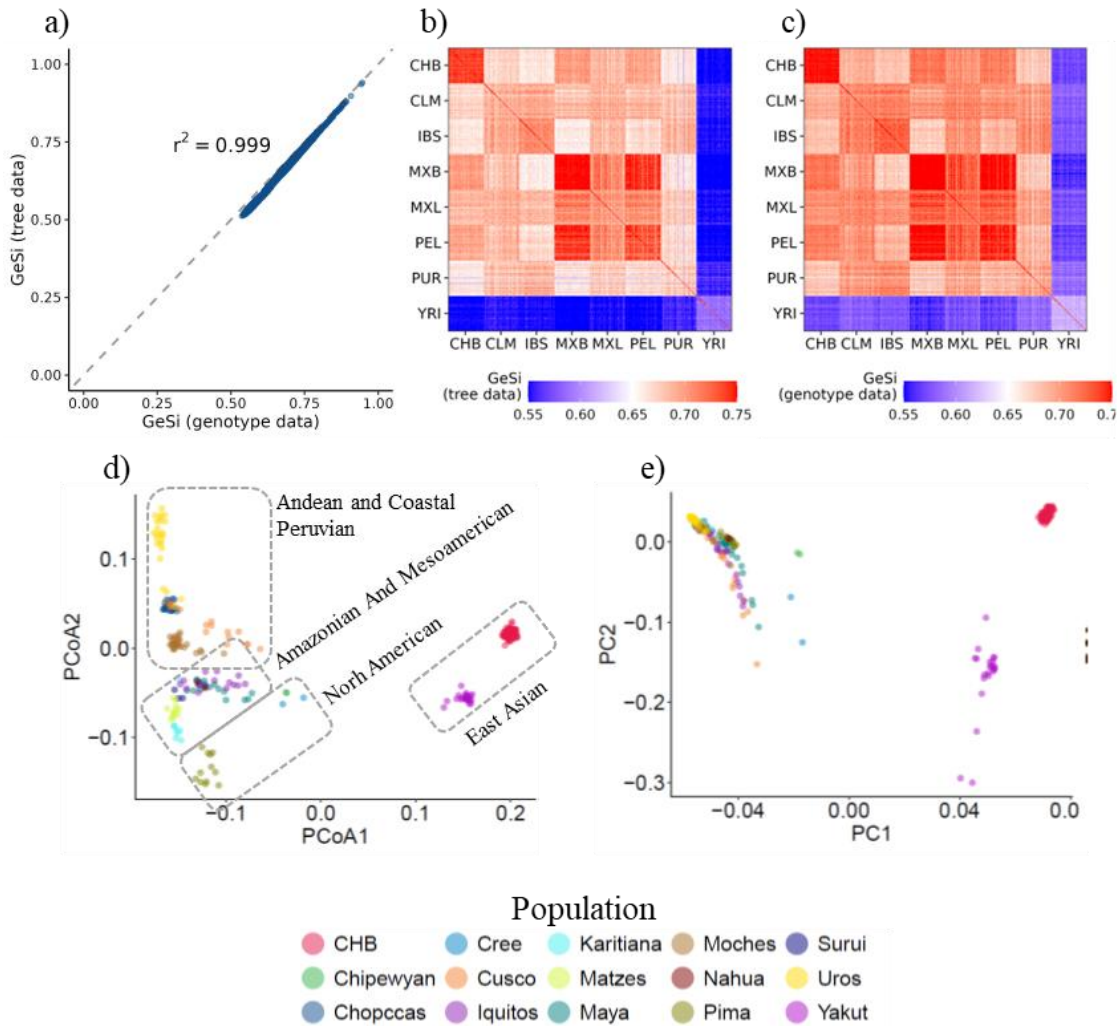


Figure 2. Detection of population structure via the genotype estimator of the coefficient of genealogical similarity. **a)** Comparison of the tree-based and genotype-based GeSi

estimates shows a high correlation between both values. **b)** Heatmaps of the GeSi values, clustered by populations labels, calculated from the true genealogical trees and **c)** from genotype data in Simulation set #1 (n=480, see methods). Both estimates reveal patterns of population structure corresponding with the simulated population labels. **d)** Principal Coordinate Analysis (PCoA) of the genotype-based GeSi values transformed into angles via an arc-cosine transformation. **e)** First two principal components estimated via Plink. r^2 : Squared Pearson's correlation coefficient.

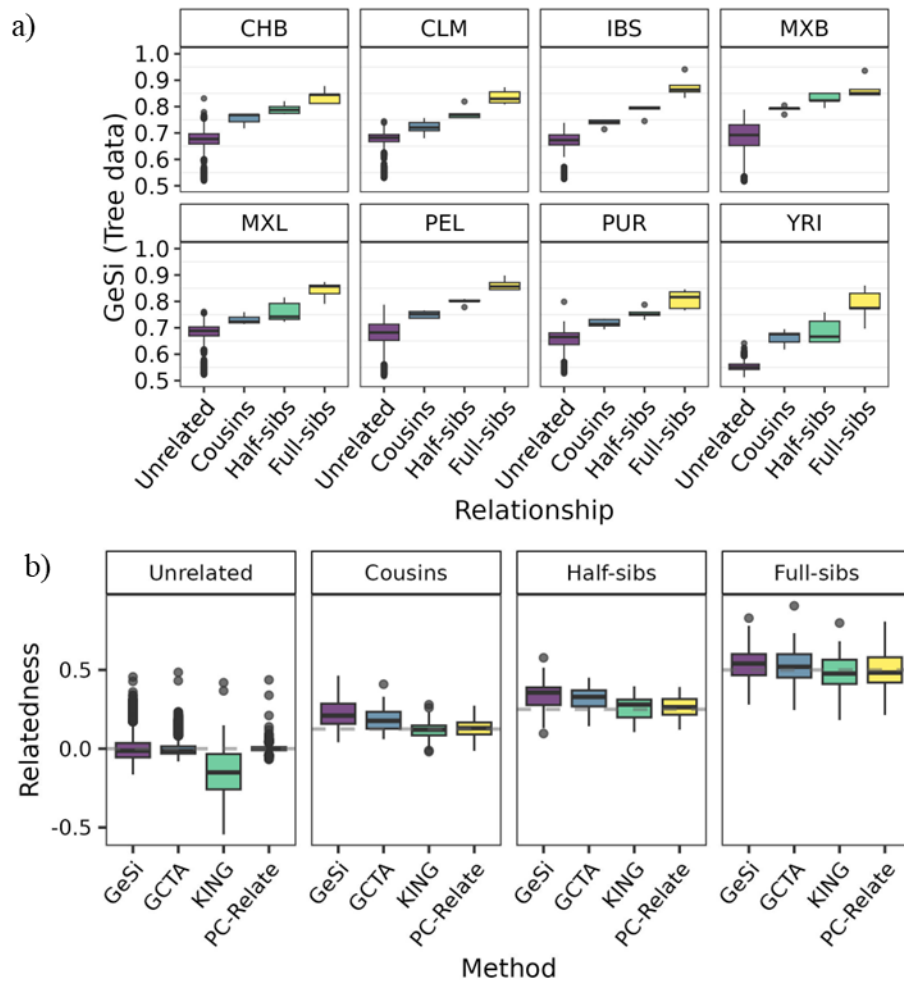


Figure 3. Distribution of the GeSi values stratified by familial relationship and population. **a)** Boxplot of the tree-based GeSi values of same-population pairs stratified by familial relationship and population of origin. **b)** Comparison of the mean-centered genotype-based GeSi estimates against other widely used relatedness measures stratified by familial relationship.

2.5. Discussion

This work introduces $GeSi_{tree}$ a new metric derived from coalescent theory that models genetic relatedness as a continuous measure of shared genealogical history. $GeSi_{tree}$ is derived from first principles by modeling the correlation of mutation counts as a function of shared branch lengths in a coalescent tree. Under a model where causal variants accumulate linearly with time and their effect sizes are independent of allele frequency, $GeSi_{tree}$ is equivalent to the Pearson correlation of the total additive genetic effects between individuals. This provides a direct, formal connection between evolutionary times and the phenotypic covariance.

The cosine similarity of the genotype vectors encoding number of derived alleles is an accurate estimator of the theoretical, tree-based definition of GeSi. Empirical validation using simulated data shows a very high concordance between the tree-based definition and its genotype-based estimator ($r^2 > 0.999$; see **Figure 2a**). This estimator successfully resolves population structure in a principal coordinates analysis (**Figure 2d**) and, unlike other relatedness measures, by design quantifies the variable background similarity among individuals traditionally classified as “unrelated” (**Figure 3a**).

A comparison of GeSi and other GRMs reveals a fundamental difference in the genealogical information they contain. Methods like PC-Relate (4) and KING-robust (60) accurately recover the expected kinship coefficients for close relatives, however, PC-Relate removes most of the background similarity among “unrelated” subjects, whilst KING-robust yields highly negative unintended kinship estimates (**Figure 3b**). Because the kinship coefficient is, by design, a measure of familial relationship, this confirms that these methods accurately estimate recent relatedness but fail to properly quantify the background similarity due to deeper shared ancestry. In contrast, both

GeSi and GCTA yield relatedness estimates that would be typically considered as biased, particularly when compared to the kinship coefficient as the inference target. However, at least in the case of GeSi, this is by design: GeSi was derived with the explicit purpose of quantifying the total relatedness, which includes recent familial relationship and deep ancestral connections. This distinction motivates a new classification of genetic relationship matrices into “shallow GRMs”, which estimate recent relatedness only, and “full GRMs”, which attempt to capture the entire continuum of genealogical history. In Aim 2, I explore the downstream effects of this distinction between full and shallow GRMs.

The empirical distribution of GCTA values and its mathematical relationship with both GeSi and principal components support its classification as a full GRM. While GCTA is often used with LD-pruned data, its mathematical definition (a scaled covariance matrix) is directly proportional to the numerator of the GeSi estimator, which represents the sum of the shared divergence time between the haplotypes of two diploid individuals. Furthermore, the eigenvectors of the GCTA matrix are asymptotically equivalent to the full set of principal components calculated from the same genotype data (5), and principal components are themselves projections of the underlying matrix of pairwise times to the most recent common ancestor (63). Together, these facts suggest that GCTA, like GeSi, contains information about the full genealogy, a conclusion that will be tested in Aim 2.

The field’s use of these two different classes of GRMs has not been guided by an explicit causal model. Shallow GRMs are used to estimate recent kinship, which is treated as a valid source of phenotypic covariance to be modeled as a random effect. In the same model, distant relatedness, captured by principal components, is treated as a source of bias to be regressed out as a fixed effect

(4,14). This partitioned approach is a direct consequence of the generalized conflation of population structure with confounding, which causes the genetic covariance caused by distant ancestry to be treated as a statistical bias to be removed instead of modeled and investigated. The lack of a unifying causal theory has allowed these two classes of GRMs, which capture different portions of the genealogy, to be used together without a formal justification for why or when such a partition is appropriate.

This work proposes a new null genetic causal model, where phenotypic covariance is proportional to the full continuum of genealogical similarity. This model states that the genetic component of a complex trait is the product of a single, continuous biological process. Under this model, the appropriate measure of genetic similarity should be therefore a “full GRM”, like GeSi, which captures the total shared ancestry from recent kinship to deep coalescent events. This null model for the genetic component is not a claim that environmental confounding does not exist; it is a claim that the genetic signal itself is not a source of bias and does not require partitioning unless there is evidence to do so. This framework is supported by empirical evidence showing that the biological effects of causal variants are highly conserved across continental ancestries (20) and that the poor transferability of polygenic scores is largely explained by differences in LD and allele frequency, not by different underlying biology (21).

2.6. Strengths and implications

In this work, I have proposed that studying the causal mechanisms of complex phenotypes should make the assumptions explicit. Particularly, I present GeSi as a potential null genetic causal model with specific assumptions that must be met for it to hold. The primary strength of this framework is not its universal correctness, but its explicitly stated assumptions and boundary conditions.

Unlike the standard partitioned model, whose implicit assumptions are rarely articulated or tested, this unified model provides a clear basis for studying the genetic architecture of a trait. By defining the specific conditions under which the model is valid, it provides a rigorous foundation for the systematic study of how phenomena like assortative mating, selection, or environmental confounding impact a trait, and motivates the development of analysis methods and simulation tools to test these specific scenarios.

2.7. Limitations

The validity of this model rests on explicit assumptions about a trait's genetic architecture, providing clear conditions under which it is expected to hold or fail. First, it assumes that the number of causal variants that have accumulated on a haplotype is linear with time, meaning that ancient and recent branches of the genealogy contribute to the phenotype in proportion to their length. Second, it assumes that non-genetic causal components of the phenotype are uncorrelated with the genetic component across all genealogical timescales. Violations of the first assumption can occur through non-neutral evolutionary processes, such as assortative mating or selection, that create a non-uniform distribution of causal alleles along branches of the tree (18). Violations of the second assumption occur in the presence of true environmental confounding, where a non-genetic causal factor is correlated with ancestry or the polygenic score (64).

2.8. Future directions

The distinction between full and shallow GRMs raises the need to systematically benchmark how each class of matrix performs when the assumptions of its underlying causal model are violated. This requires a simulation framework capable of explicitly generating data under scenarios with

more complex genetic architectures, such as those influenced by assortative mating or selection, and in the presence of precisely defined environmental confounders. Such a framework, which will be developed in Aim 3, would allow for a rigorous evaluation of when the additional genealogical information in a full GRM is necessary, and under what specific confounding scenarios the inclusion of PCs as covariates is necessary.

More broadly, this work highlights the need to move beyond statistical correction and develop formal methods to test the true causal architecture of complex traits in real data. This requires tackling two separate challenges. First, methods must be developed to test for the presence of true, non-genetic confounding by disentangling the correlation of an environmental factor with ancestry from the genetic covariance of the trait itself. Second, it remains to be evaluated whether the concept of genetic confounding (17) can be stated from an evolutionary point of view as a question of non-linearity between phenotypic divergence and genealogical time, and, if that's the case, whether that linearity can be tested. Developing such hypothesis tests is a significant challenge, but it is the necessary next step in moving the field from correcting for statistical inflation to a more rigorous, mechanistic understanding of the sources of inflation.

2.9. Conclusions

This work provides a new, theoretically grounded framework for measuring genetic relatedness and re-evaluates the role of population structure in genetic association studies. The coefficient of genealogical similarity, GeSi, is derived from coalescent theory and successfully captures the full continuum of shared ancestry in a single parameter. This approach reveals a distinction between “full” relationship matrices like GeSi and GCTA, which contain the complete genealogical history

of a sample, and “shallow” matrices like PC-Relate, which are designed to capture recent kinship only.

Finally, this work challenges the field’s standard practice of partitioning genetic similarity and conflating population structure with confounding. It reframes population structure as a source of genetic covariance that a full GRM can correctly model if explicit conditions are met. This work provides not only a new statistic but also a more rigorous causal model, shifting the burden of proof to justify the partitioning of genetic effects rather than assuming it as the default.

3. Aim 2: Using GeSi to Account for Population Structure and Relatedness in Genome Wide Association Studies

3.1. Introduction

Aim 1 established the theoretical distinction between “full” GRMs, which capture the entire continuum of shared ancestry, and “shallow” GRMs, which capture only recent relatedness. This distinction leads to a single, falsifiable hypothesis: in the absence of confounding from non-genetic factors, the genetic covariance from population structure is fully modeled by a full GRM, rendering the inclusion of principal components redundant. The standard partitioned model, which uses a GRM for relatedness and PCs for structure, is therefore only necessary if its two components model different phenomena. If, however, both recent and distant relatedness are part of the same continuous biological process, a single, genealogically complete GRM should be sufficient to model the genetic covariance, as long as the assumptions of the GRM are met.

In this aim, I test that central hypothesis through a series of simulation and real-data experiments that systematically compare the performance of full and shallow GRMs, both with and without PC adjustment, in standard genetic analyses. The experiments are designed to move from idealized scenarios, where the true genetic architecture is known and non-genetic confounders are absent, to the complexity of real human data where the causal mechanisms are unknown. This step-wise approach allows for the formal disentanglement of the statistical signatures of genetic covariance from the effects of confounding by non-genetic factors.

3.2. Data and Methods

For this Aim, I used real data from the BioMe cohort and five different sets of simulated data.

3.2.1. BioMe cohort

The BioMe dataset originated from the BioMe cohort (dbGAP study accession: phs001644) in the TOPMed project Freeze 9 dataset (47). This dataset consisted of phenotype and whole-genome sequence (WGS) data available for 12,054 subjects. Samples missing basic covariates (population label, sex and age) or height data, and potential duplicate entries (Kinship coefficient ≥ 0.49) were filtered out. Only subjects described as African or African Americans (AFR, $n = 2,975$), European Americans (EUR, $n = 2,442$), and Latino or Latin American (LAT, $n = 4,468$) were kept, making a total of 9,885 subjects in the final dataset. Subjects described as East Asians, South Asians or Indigenous Americans were filtered out due to the small sample sizes of those groups. Using PC-Relate (see methods) and a kinship coefficient threshold of $(1/2)^{4.5}$, corresponding to the lower-boundary of third-degree relatives (62), I identified a set of 9,190 unrelated subjects and 695 subjects related to them, which were described as AFR ($n = 206$), EUR ($n = 28$) and LAT ($n = 461$).

3.2.2. Phenotype simulation

I generated multiple simulated data sets to test different scenarios. The Simulation Set #1 generated for and described in aim 1 was not used in Aim 2 and is thus no further discussed. Simulated Sets #2 and #3 consisted of both simulated genotype and phenotype data; and Sets #4-#5 consisted of phenotypes simulated from the real BioMe genotype data. I describe first the general strategy for phenotype simulation, and then mention which strategy was used for each Simulated Set.

The genetic architecture was defined by the heritability, number of causal sites, dependence of the effect size on the minor allele frequency (LDAK's α parameter (31)), and effect of the background linkage disequilibrium on the distribution of causal SNPs. Additionally, an environmental confounder component defined by the population label was also included in Simulated Set #4. All phenotype simulations were done in R v.4.3.3.

I simulated quantitative phenotypes under a polygenic additive genetic model. For each simulation, I randomly assigned 1,000 biallelic sites as causal sites. The causal sites were sampled either with uniform probability, or with sampling weight depending on their LD scores, as will be explained later. The effect sizes were drawn from a normal distribution with mean 0 and unit variance and were then scaled according to their minor allele frequency (MAF) via the α parameter as described in the next section.

For individual i , the total genetic effects were calculated as $g_i = \sum_j X_{ij}\beta_j$, where X_{ij} represents the genotype for individual i at causal site j , and β_j is the effect size for causal site j . The narrow-sense heritability (h^2) was set to 0.5 by adding an environmental component e_i drawn from a normal distribution with mean zero and variance adjusted to achieve the target heritability according to the equation $\sigma_e^2 = \frac{\sigma_a^2(1-h^2)}{h^2}$.

The final phenotype was calculated as $y_i = g_i + e_i$.

3.2.2.1. Effect of the minor allele frequencies on SNP effect sizes (α parameter)

I used Doug Speed's LDAK (31) approach to allow SNP effect sizes to vary with allele frequency.

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the unscaled uncentered genotype matrix of n subjects at p sites and let x_{ij}

denote the genotype of subject i at site j , where $x_{ij} \in \{0,1,2\}$ denotes the number of copies of the alternate allele. Let f_j denote the alternate allele frequency at site j , then the matrix of mean-centered genotypes $\mathbf{X}^{(c)}$ has elements $x_{ij}^{(c)} = (x_{ij} - 2f_j)$, and the scaled genotypes matrix $\mathbf{Z} \in \mathbb{R}^{n \times p}$ has elements $z_{ij} = x_{ij}^{(c)} \times [2f(1-f)]^{-\alpha/2}$. Let β_j denote the effect size of variant j on a trait in the raw genotype scale. Then, if the effect sizes have constant variance in the scaled genotype space, the variance of the raw effect sizes is $Var(\beta_j) = [2f(1-f)]^\alpha \times constant$, and $E[h^2] \propto [2f(1-f)]^{1+\alpha}$ is the expected heritability explained by site j . The LDAK model reduces to the GCTA model by setting $\alpha = -1$.

To use the LDAK model, I first simulated the effects sizes $\beta_j^{(raw)}$ from a standard normal distribution and assigned them to the randomly selected sites. Then the final effect sizes were obtained with $\beta_j = \beta_j^{(raw)} \times [2f_j(1-f_j)]^{-\alpha/2}$.

3.2.2.2. Simulation of confounding effects

The confounding effects were simulated to correlate with population structure, specifically related to the three population labels European Americans (EUR), African or African Americans (AFR), and Latino or Latin Americans (LAT). An incidence matrix $\mathbf{Q} \in \mathbb{R}^{n \times 3}$ was constructed, mapping each of the n subjects to their respective population. The vector $\mathbf{b}^T = [b_{EUR}, b_{LAT}, b_{AFR}]$ contains the effect sizes associated with each population. Then, I defined the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{b} + \mathbf{e}$, where \mathbf{X} is the genotype matrix of causal sites and $\boldsymbol{\beta}$ a vector of their coefficients.

I followed the next procedure to introduce a specific directional confounding pattern that correlated with the amount of both European and African ancestry:

1. The effect $b_{raw(LAT)} = b_{(LAT)}$ was set to zero, $b_{raw(EUR)}$ was sampled from a truncated negative standard normal distribution, and $b_{raw(AFR)}$ from a truncated positive standard normal distribution. Then I defined the vector $\mathbf{b}_{raw}^T = [b_{raw(EUR)}, b_{raw(LAT)}, b_{raw(AFR)}]$. I set $b_{(LAT)}$ to zero because LAT had intermediate levels of both European and African ancestry compared to the groups labeled as EUR and AFR.
2. A scaling factor ω was calculated to ensure that the genetic and confounding effects explained the same amount of phenotypic variance. This factor was calculated by $\omega = \left(\frac{\text{Var}(\mathbf{X}\boldsymbol{\beta})}{\text{Var}(\mathbf{Q}\mathbf{b}_{raw})} \right)^{1/2}$. Where $\sigma_a^2 := \widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta})$ is the variance of the total additive genetic effects, and $\text{Var}(\mathbf{Q}\mathbf{b}_{raw})$ is the variance contributed by the confounding effects using the raw coefficients \mathbf{b}_{raw} .
3. The final confounding effect vector \mathbf{b} was obtained by multiplying the raw coefficient vector by the scaling factor: $\mathbf{b} = \omega\mathbf{b}_{raw}$.

This procedure yielded correlation values between genetic and confounding effects that ranged between -0.18 to -0.09. The residual error term \mathbf{e} was drawn independently from a normal distribution with mean zero and variance σ_e^2 . The residual variance parameter was defined as before, $\sigma_e^2 = \frac{\sigma_a^2(1-h^2)}{h^2}$, where h^2 is the pre-specified narrow-sense heritability.

3.2.2.3. Simulations under a Latin American Demographic Model - Simulation Sets #2 and #3

I modified the previously described demographic model of Latin America (46), by adding a new artificial population (“LAT”, short for Latino and Latin Americans) resulting from an admixture event three generations ago consisting of 25% CLM, 25% MXL, 25% PEL, and 25% PUR. The

admixture event was added three generations in the past so that current samples could draw their grandparents from any of the contributing populations. Similarly to Simulation Set #1 (see aim 1), I used the DTWF and JC69 models, and mutation rate of 1.25×10^{-8} , but generated ten chromosomes (chr13-chr22) using the GRCh38 recombination maps. I generated 5,000 samples under four different scenarios:

- No recent admixture and no recent relatedness: 5,000 unrelated samples from IBS.
- Admixture without recent relatedness: 5,000 unrelated samples from LAT.
- No recent admixture with recent relatedness: 600 full-sib pairs, 600 half-sib pairs, 600 first-cousin pairs, and 1,400 unrelated subjects from IBS.
- Admixture with recent relatedness: 600 full-sib pairs, 600 half-sib pairs, 600 first-cousin pairs, and 1,400 unrelated subjects from LAT.

For both Simulation Set #2 and Set #3, I simulated a polygenic trait with $h^2 = 0.5$ determined by 1,000 sites sampled at random with uniform probability. The causal sites' effect sizes were generated using LDAK's $\alpha = 0$, consistent with the original *GeSi_{tree}* model, in Simulation Set #2, and with $\alpha = -1$ in Simulation Set #3, which corresponds to the GCTA model. Both simulation sets consisted of 10 replicates per scenario.

3.2.2.4. BioMe-Based Simulations (Simulation Sets #4 and #5)

I also ran simulations using real genotypes from the BioMe WGS data, assigning simulated phenotypes under various scenarios:

- **Simulation Set #4:** Phenotypes simulated with $\alpha = -1$, causal sites sampled with uniform probability and no confounding.

- **Simulation Set #5:** Phenotypes simulated with $\alpha = -1$, causal SNPs sampled with uniform probability, and environmental confounding weakly or moderately correlated with genetic ancestry.

1,000 causal variants were randomly selected, and the heritability set to 0.5 in the EUR group. All simulations were replicated 10 times. The purpose of defining a population-specific heritability was to observe the changes in heritability that can occur simply due to population structure without invoking any complex mechanism such as epistasis, gene-by-environment interaction, etc.

3.2.3. Calculation of GRMs and Principal Components Analysis

The derivation of the GeSi equations was presented in the results section of aim 1. Here I describe how GeSi and other GRMs were calculated from the genotype for simulated and real data used in in this aim.

The derivation of GeSi did not require that the sites are in linkage equilibrium. However, the definition of GeSi was limited to single trees, and the effect of averaging contiguous trees was unknown. In order to assess whether genotype correlation across sites could have any impact downstream in the analysis, I generated three different genotype filesets for each dataset: a) all biallelic sites in whole-genome sequence (WGS) data, b) only sites that were roughly in linkage equilibrium (LD-pruned), or c) using sites selected at random (RS) with uniform probability.

The “LD-pruned” file sets were generated using PLINK(59) v.1.9 with the option `--indep-pairwise`. For the Simulation Sets #2 and #3 (Msprime-generated), the window size was set to 400 SNPs with steps of 50 SNPs, and r^2 threshold of 0.1. For the BioMe data, the window size was 1000 sites with steps of 100 sites, and r^2 threshold of 0.07. The sites in the “RS” file set were selected

with uniform probability from the biallelic sites in the WGS data, ensuring the number of sites per chromosome was proportional to the total WGS sites count per chromosome. The total number of RS sites across all autosomes was the same as the total number of sites in the corresponding LD-pruned set.

The GeSi and LDAK matrices were calculated using each file set in the BioMe and Simulated Sets #2-6 data, with α values ranging from -2 to 0 in steps of 0.25. The LDAK matrix with $\alpha = -1$ corresponds to the GCTA model. The PC-Relate matrix was also calculated, but only with $\alpha = -1$ because its implementation in the GENESIS does not allow for different α values. Furthermore, because of the high computational runtime of PC-Relate, it was calculated using LD-pruned data, only.

For the LD-pruned and RS filesets, a single GRM corresponding to all autosomes was computed for each α . For the full WGS data, GRMs were first calculated independently for each autosome and subsequently averaged in R, weighting each chromosome-specific GRM by its respective site count, to generate the final genome-wide GRM for each α .

The GeSi matrices were not calculated directly from the genotype data. By definition, GeSi's genotype-based estimators (see Equation (21)) is equivalent to a standardized LDAK matrix. Therefore, the LDAK matrices were first calculated using the highly efficient c++ toolkit ldak v.6.1 (31). GeSi was then calculated in R in a single step after loading the LDAK matrix from disk:

```
## R code:
ldak <- data.table::fread("ldak_grm.tsv.gz", header =
TRUE)%>% as.matrix()
gesi <- ldak/sqrt(tcrossprod(diag(ldak)))
```

For the LD-pruned and RS filesets, PCs were calculated directly with LDAK via the `--pca` option keeping 20 PCs. For the WGS data, PCs were obtained via Eigen decomposition (eigen function in R) of each final averaged GRM (one GRM per α value per data type per dataset). For each decomposition, the top 20 eigenvectors were retained as the PCs.

The PC-Relate matrices were calculated in the same way as in aim 1. Briefly, I calculated an initial GRM using the KING-robust method (60) via `SNPRelate::snpgdsIBDKING`. Next, I calculated the principal components via the PCAiR method implemented in `GENESIS::pcair` using the KING-robust matrix as input, with a kinship threshold of $(1/2)^{4.5}$ and divergence threshold of $-(1/2)^{4.5}$, corresponding to the maximum kinship coefficient value for third-degree relatives (62). Next, I calculated the PC-Relate kinship matrix via the `GENESIS::pcrelate` function, using the first four principal components calculated with PCAiR in the previous step as covariates. I ran a second and final round of PCAiR using the PC-Relate kinship estimates as input and keeping the kinship threshold at $(1/2)^{4.5}$ and the divergence threshold at $-(1/2)^{4.5}$. The output of PCAiR included a list of unrelated and related individuals based on the PC-Relate kinship estimates and the specified kinship threshold.

3.2.4. Using GeSi to Estimate Variance Components via Mixed Linear Models

Standard GRM methods assume that a single genetic variance component exists and is estimable through mixed linear models (MLMs). However, the generative model in Equation (10) showed that the variance of a subject's genetic effects depends on its amount of autozygosity. If this model held true, it would mean that MLMs cannot estimate a single genetic variance parameter because such parameter does not exist. Instead, MLMs would estimate an average of the pairwise product of genetic effects' standard deviations:

$$\text{Cor}(g_i, g_j) = \text{GeSi}_{tree}$$

$$\frac{\text{Cov}(g_i, g_j)}{\sigma_{g_i} \sigma_{g_j}} = \text{GeSi}_{tree}$$

$$\text{Cov}(g_i, g_j) = \text{GeSi}_{tree} \times [\sigma_{g_i} \sigma_{g_j}]$$

Thus, to use $\widehat{\text{GeSi}}_\alpha$ in a mixed linear model, it is necessary to assume that a single genetic variance component exists, in which case:

$$\widehat{\text{Cov}}(g'_i, g'_j) = \widehat{\text{GeSi}}_\alpha \times \sigma_g^2$$

Where $\widehat{\text{GeSi}}_\alpha$ is calculated from mean-centered genotype data in order to measure the phenotype deviations from the mean instead of from the ancestral phenotype. This is justified by the observation that:

- $y_i = y_o + \mathbf{x}_i \cdot \boldsymbol{\beta} + e_i \equiv \bar{y} + \mathbf{x}_i^{(c)} \cdot \boldsymbol{\beta} + e_i$
- $g_i := (\mathbf{x}_i \cdot \boldsymbol{\beta})$
- $g_i^{(c)} := (\mathbf{x}_i^{(c)} \cdot \boldsymbol{\beta})$

Where y_o is the ancestral phenotype, $\mathbf{x}_i \in \mathbb{R}^{p \times 1}$ is the genotype vector of subject i encoding the number of derived alleles at p sites, $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is the vector of causal sites' effect sizes, \bar{y} is the average phenotype value in the sample, and $\mathbf{x}_i^{(c)}$ is the mean-centered genotype vector of subject i .

Where, g_i is the polygenic score of subject i , and $g_i^{(c)}$ its polygenic score calculated as deviations from the sample mean. In other words, after centering the genotype matrix, the phenotype deviations are measured with respect to the sample mean.

3.2.5. Mixed linear model association testing and heritability estimation

I used mixed linear models (MLMs) for both genome-wide association studies (GWAS) and heritability estimation. The general inference and prediction model was $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, where \mathbf{Y} is the phenotype vector of length n , $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix for fixed effects, $\boldsymbol{\beta}$ is the vector of fixed effect sizes of length p , \mathbf{Z} is the design matrix for random effects, \mathbf{u} is the vector of random genetic effects, and \mathbf{e} is the residual error vector. I assumed that $e \sim N(0, \mathbf{I}\sigma_e^2)$ and $\mathbf{u} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{K})$, where \mathbf{K} is a genetic relationship matrix, σ_a^2 is the additive genetic variance, and σ_e^2 is the residual variance.

Genome-wide association tests (GWAS) were run on data from Simulation Sets #2 and #3 and the real data from the BioMe cohort using the GENESIS package for R in a two-step procedure. In the first step, the null model was fit using the function `GENESIS::fitNullModel`, with all the fixed-effect covariates and random effect matrices but excluding single genetic variants. This first step outputs the variance components estimates required to calculate the heritability and make phenotype predictions (see next section). In the second step, the single-variant association tests was run using the null model from the previous step through the function `GENESIS::assocTestSingle` with the Gaussian family distribution (61). The trait with real data analyzed in the BioMe data was height.

To assess the impact of ancestry correction, all models were run with either zero or ten principal components (PCs) as fixed-effect covariates. MLMs of Simulated Sets #2 and #3 did not include any covariates other than PCs. MLMs of the BioMe-based simulations (Set #2 and #3) and of the real phenotype data from the BioMe cohort included sex, age, age² and population label as covariates.

The residuals were assumed to be homoscedastic in all MLMs of Simulated Sets #2-5. MLMs of the real height data in the BioMe cohort used a fully-adjusted two-stage residual rank normalization (65) (options `two.stage=TRUE` and `norm.option="by.group"` in the `GENESIS::assocTestSingle` function), with a different residual variance component estimated for each sex-by-population group, specified via the option `group.var="ancestry_sex"`.

The heritability was estimated in all datasets with the variance components output by the `GENESIS::fitNullModel` function. The single-variant association tests were run only in Simulated Sets #2-3. I used the Simulated Sets #2-3 to systematically assess:

- Type I error rate: The proportion of truly null SNPs declared significant under a nominal p-value threshold ($p < 0.05$) or genome-wide significance ($p < 5 \times 10^{-8}$).
- Statistical power: The proportion of truly causal SNPs achieving significance, particularly relevant when comparing how different methods handle admixture or cryptic relatedness.
- Genomic inflation: Calculated as the genomic control lambda (λ_{GC}) statistic.
- Accuracy of effect size estimates: Measured the mean squared error of the estimates effect sizes, and by the squared Pearson correlation (r^2) and the coefficient of determination (R^2) between true and estimated effect sizes.
- Bias of effect size estimates: Assessed by the regression coefficient of the estimated effect sizes on the true effect sizes.
- Accuracy of the heritability estimates: Mean squared error of the heritability estimates, by comparing the inferred values against the simulated parameters.

3.2.6. Inference of the α parameter through CV-BLUP

I used the Best Linear Unbiased Predictor (BLUP) equation (3,66) to predict the total genetic effects in the held-out sets of a 20-fold cross-validation (CV) analysis of Simulated Sets #4-5 and the real height data from the BioMe cohort. I repeated this analysis for each GRM calculated with α parameter values from -2 to 0 in steps of 0.25, in order to assess whether this approach could be used to find the true α value. Additionally, I recorded the restricted maximum likelihood of all models as an alternative approach to select the best α value, as has been done before (67).

All datasets were split into 20 cross-validation folds of 494 subjects each. Because the total sample size (9,885) was not a multiple of 20, I randomly excluded the same five samples from all cross-validation analyses. I fit a MLM the GENESIS::fitNullModel function on the training set of each cross-validation fold. The null models included a genetic similarity matrix (either a GRM or a GeSi matrix) and the same covariates as in the MLM association testing describe in the previous section. I obtained the variance components estimates for each fold, which was then fed into the BLUP equation:

$$\hat{\mathbf{u}} = \mathbf{GZ}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Where $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ is the vector of phenotypes of the training set ($n = 9,386$); $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix of fixed effects covariates in the training set $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p \times 1}$ is the vector of fixed effects coefficients estimated in the training set; $\hat{\mathbf{u}} \in \mathbb{R}^{(n+m) \times 1}$ is the predicted genetic effects of both the training and validation ($m = 494$) sets; $\mathbf{G} = \sigma_a^2\mathbf{K}$ is the estimated covariance matrix of the genetic effects; \mathbf{K} is the genetic similarity matrix of both training and validation sets, and $\mathbf{G}, \mathbf{K} \in \mathbb{R}^{(n+m) \times (n+m)}$. The

incidence matrix $\mathbf{Z} \in \mathbb{R}^{n \times (n+m)}$ maps the genetic effects to each subject, $\mathbf{V} = (\mathbf{ZGZ}^T + \mathbf{R})$, where $\mathbf{V} \in \mathbb{R}^{n \times n}$ is the total variance matrix, and $\mathbf{R} \in \mathbb{R}^{n \times n}$ is the diagonal matrix of residual variance.

I calculated the total phenotype of the validation set as the sum of estimated fixed effects and random genetic effects $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$, and then used it to estimate the root of the mean squared error (RMSE) of each fold as $RMSE_{fold} = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2$, and calculated the final RMSE as the average across all CV folds. For each data set, similarity matrix algorithm (LDAK/GCTA or GeSi), and data type (WGS, LD-pruned or RS), I found the α value that minimized the average RMSE, and calculated the Pearson correlation coefficient between the true genetic value and the predicted genetic value (simulated data only), and between the true phenotype and predicted phenotype (simulated and BioMe data).

3.3. Results

3.3.1. Fully Informative GRMs do not Require Principal Components to Adjust for Admixture in Single-Variant Association Tests in the Absence of Confounding

The purpose of this simulation was two-fold, first, to validate the generalizations of GeSi by mean-centering the genotype matrix and changing the scaling factor α to values other than zero, and second, to assess whether PCs are required to adjust for inflation due to population structure or admixture. I benchmarked the generalized GeSi estimators against standard methods in single-variant association tests using the Msprime-generated Simulation Sets #2 ($\alpha = 0$) and #3 ($\alpha = -1$).

The GeSi and GCTA/LDAK GRMs accurately estimated SNP effect sizes across all scenarios, with slightly higher accuracy in admixed (LAT) samples than in homogeneous (IBS) ones.

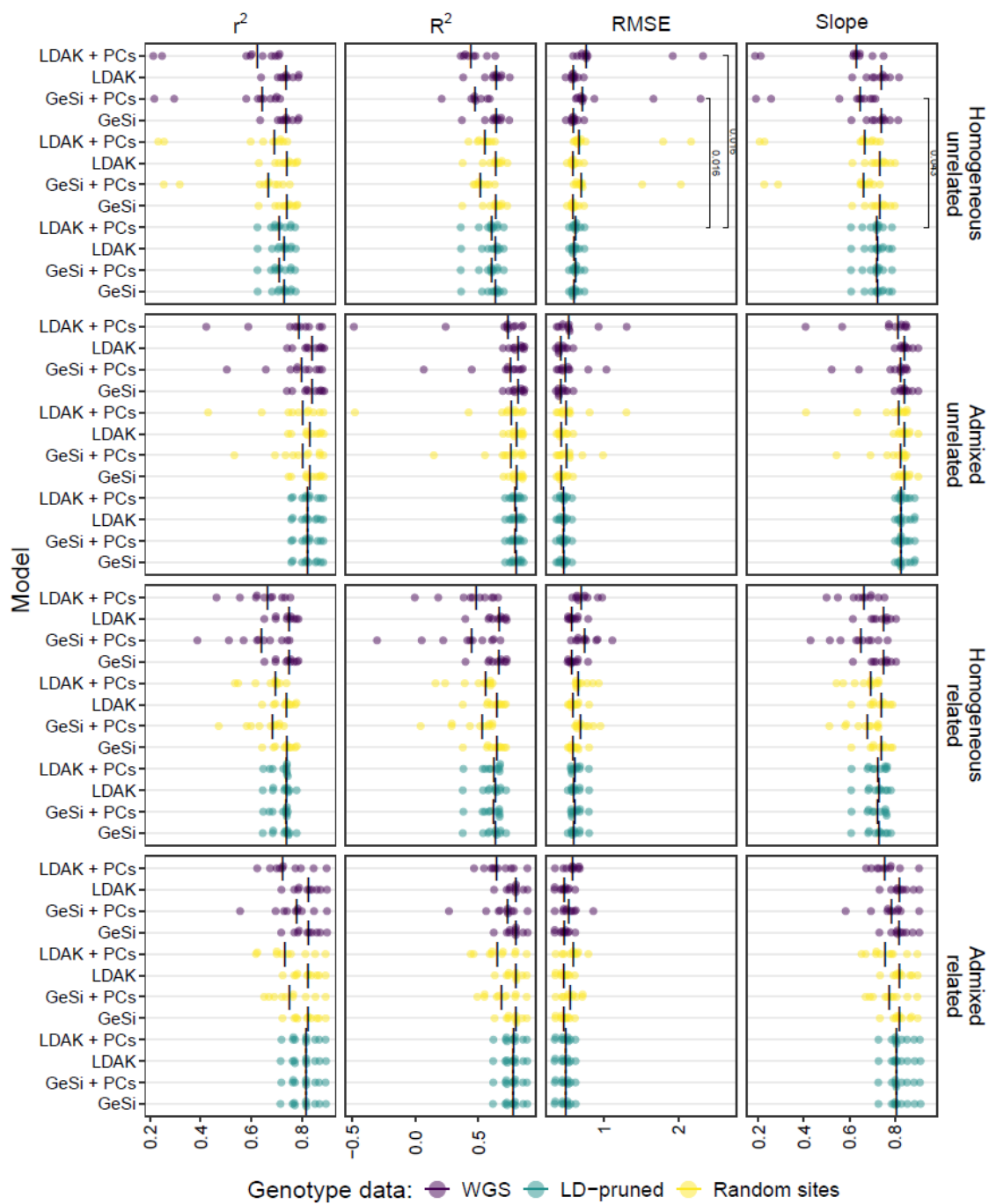
Including principal components (PCs) as covariates in models with GRMs built from dense data WGS or random sites (RS) consistently reduced the accuracy of effect size estimates by introducing a downward bias (slope falling farther away from 1) and reducing the Pearson's correlation between true and estimated effect sizes. This effect was absent for models using GRMs built from LD-pruned data (**Figure 4** and **Figure 5**). The PC-Relate GRM, which is built from LD-pruned data, had estimated the effect sizes with accuracy comparable to other LD-pruned GRMs and was also unaffected by the inclusion of PCs (**Figure 4** and **Figure 5**).

Test-statistic inflation was primarily influenced by the data type used to construct the GRM. Models using WGS-based GRMs were the most conservative (mean $\lambda_{gc} \leq 1.0$), whereas those using LD-pruned GRMs showed slight inflation (mean $\lambda_{gc} \approx 1.03 - 1.07$, see **Figure 6** and **Figure 7**)

These simulations lacked non-genetic confounding, and the inflation was observed in all four demographic and sampling scenarios. Thus, the observed inflation does not reflect uncorrected population or confounding, but rather the expected polygenic signal from non-causal variants tagging true causal variants. While PC inclusion did not affect inflation for GeSi or GCTA/LDAK models, it was essential for controlling inflation in PC-Relate models, particularly in the admixed and relatedness scenarios, where models without PCs showed significantly greater inflation (**Figure 6** and **Figure 7**).

Finally, a trade-off between inflation control and statistical power was observed. Models with LD-pruned GRMs showed slightly higher raw power to detect causal variants than models using WGS or RS-based GRMs (**Figure 8** and **Figure 9**). However, after adjusting for test-statistic inflation

via the genomic control, the statistical power was equivalent across all methods and modeling choices.



(Caption in next page)

Figure 4. Accuracy of true causal effects estimation in single-variant mixed linear model association tests in the Msprime-generated Simulation Set #2 ($\alpha = 0$). Each column shows a different accuracy measure of the estimated SNP effect sizes: squared Pearson correlation coefficient (r^2), coefficient of determination (R^2), root of the mean squared error (RMSE), and the slope of regressing the predicted effects on the true effects (Slope). The black vertical bars represent the median for each measure within each panel. Each panel represents one of the four different combinations of demographic and ascertainment scenarios that were simulated: A) Unrelated individuals from Iberians in Spain (IBS); B) Mixture of related and unrelated individuals from IBS; C) Unrelated individuals from Latino and Latin American populations (LAT; see main text and methods); and D) Mixture of unrelated and unrelated individuals from LAT. Within each panel, the vertical axis details whether the mixed linear model (MLM) used a GeSi or LDAK genetic relationship matrix, and whether principal components (PCs) were included in the model. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. A two-sided Wilcoxon rank sum test was used to compare, within each panel, all methods against the reference model (LDAK-LD-pruned + 10 PCs). Only the FDR-adjusted p-values below 0.05 are shown.

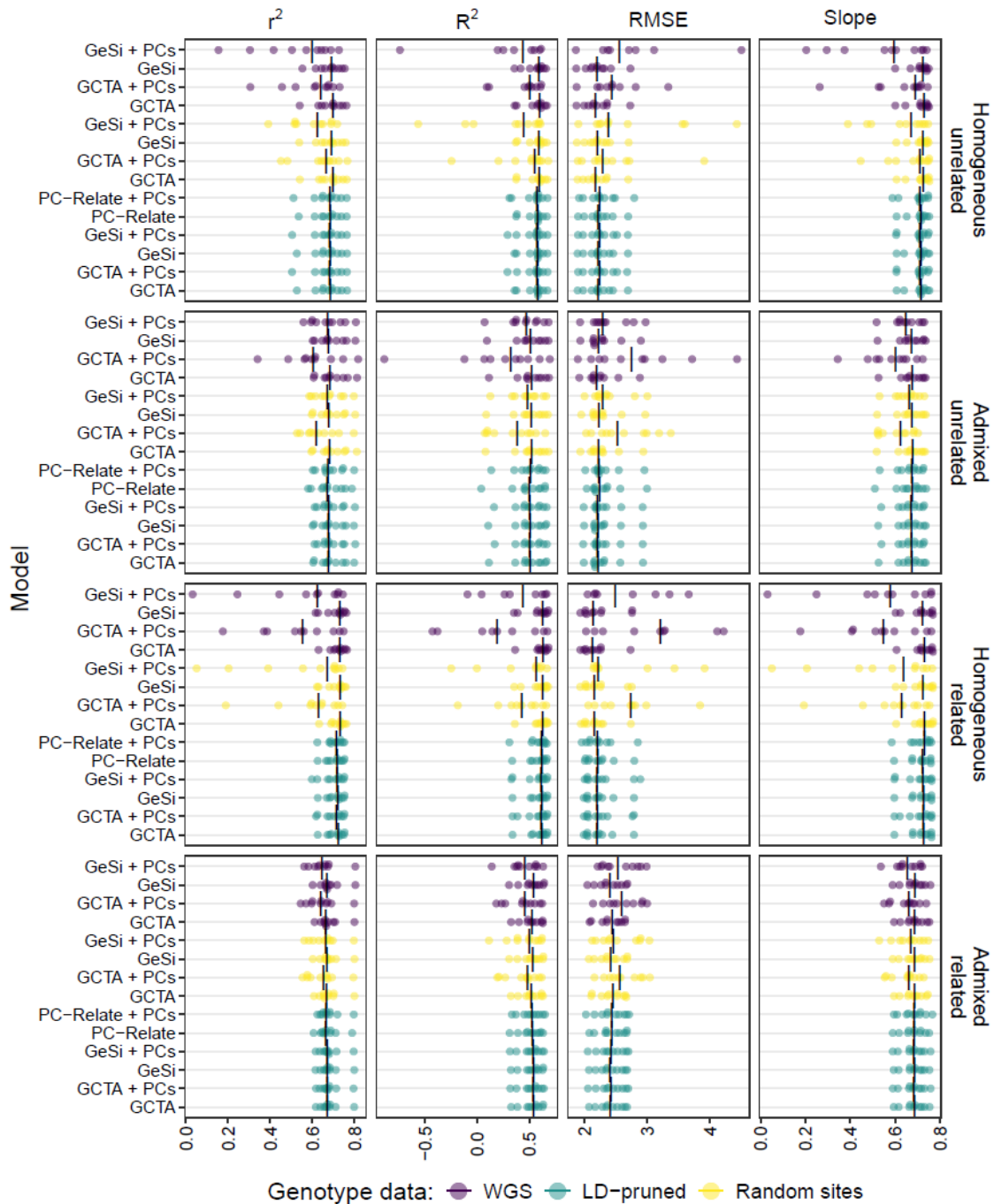
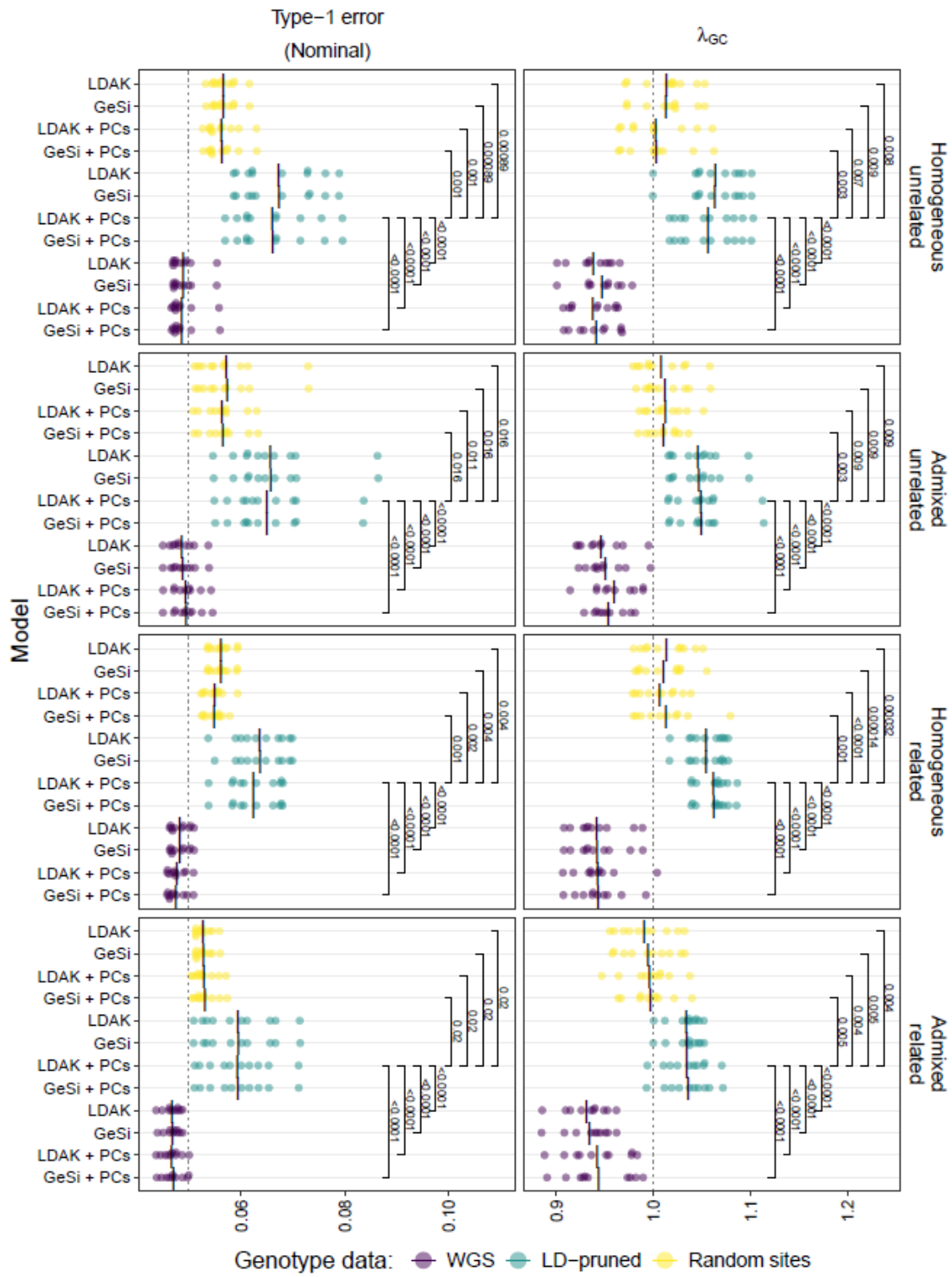


Figure 5. Accuracy of true causal effects estimation in single-variant mixed linear model association tests in the Msprime-generated Simulation Set #3 ($\alpha = -1$). Similar to Suppl. Fig. 1, but for the Simulation Set #3 ($\alpha = -1$). Note that the GCTA GRM is equivalent to LDAK with $\alpha = -1$.



(Caption in next page).

Figure 6. Type I error rate and inflation of the chi-square statistic in Simulated Set #2 ($\alpha = 0$). The first column shows the type I error rate at a significance threshold of 0.05; and the second column shows the genomic control lambda parameter, defined as the median of the chi-square statistic divided by the median of a chi-square distribution with one degree of freedom. The black vertical bars show the mean for each measure within each panel. Each row represents one of the four different combinations of demographic and ascertainment scenarios that were simulated: A) Unrelated individuals from Iberians in Spain (IBS); B) Mixture of related and unrelated individuals from IBS; C) Unrelated individuals from Latino and Latin American populations (LAT; see main text and methods); and D) Mixture of unrelated and unrelated individuals from LAT. Within each panel, the vertical axis details whether the mixed linear model (MLM) used a GeSi or LDAK genetic relationship matrix, and whether principal components (PCs) were included in the model. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. A two-sided Wilcoxon rank sum test was used to compare, within each panel, all methods against the reference model (LDAK-LD-pruned + 10 PCs). Only the FDR-adjusted p-values below 0.05 are shown.

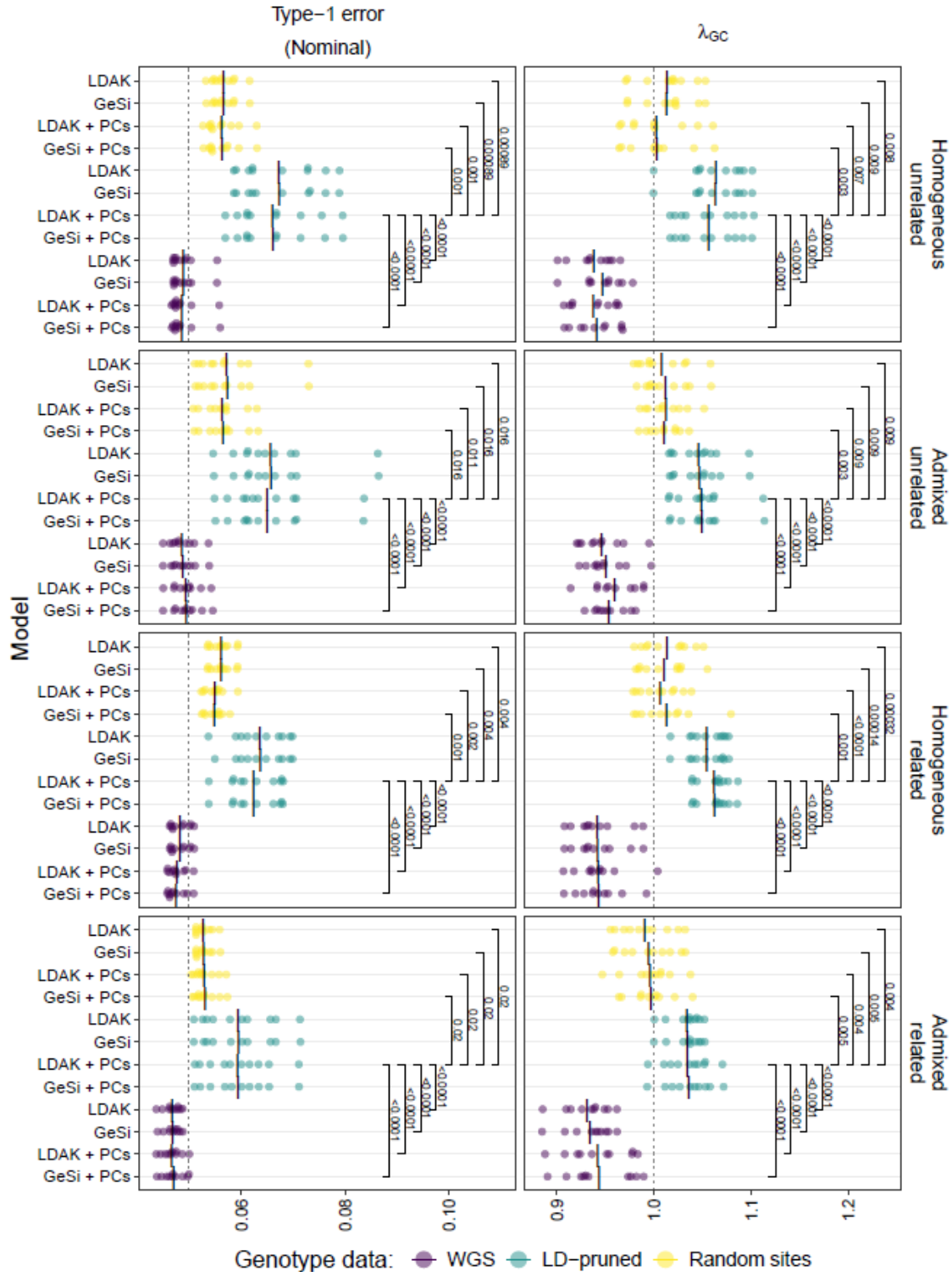
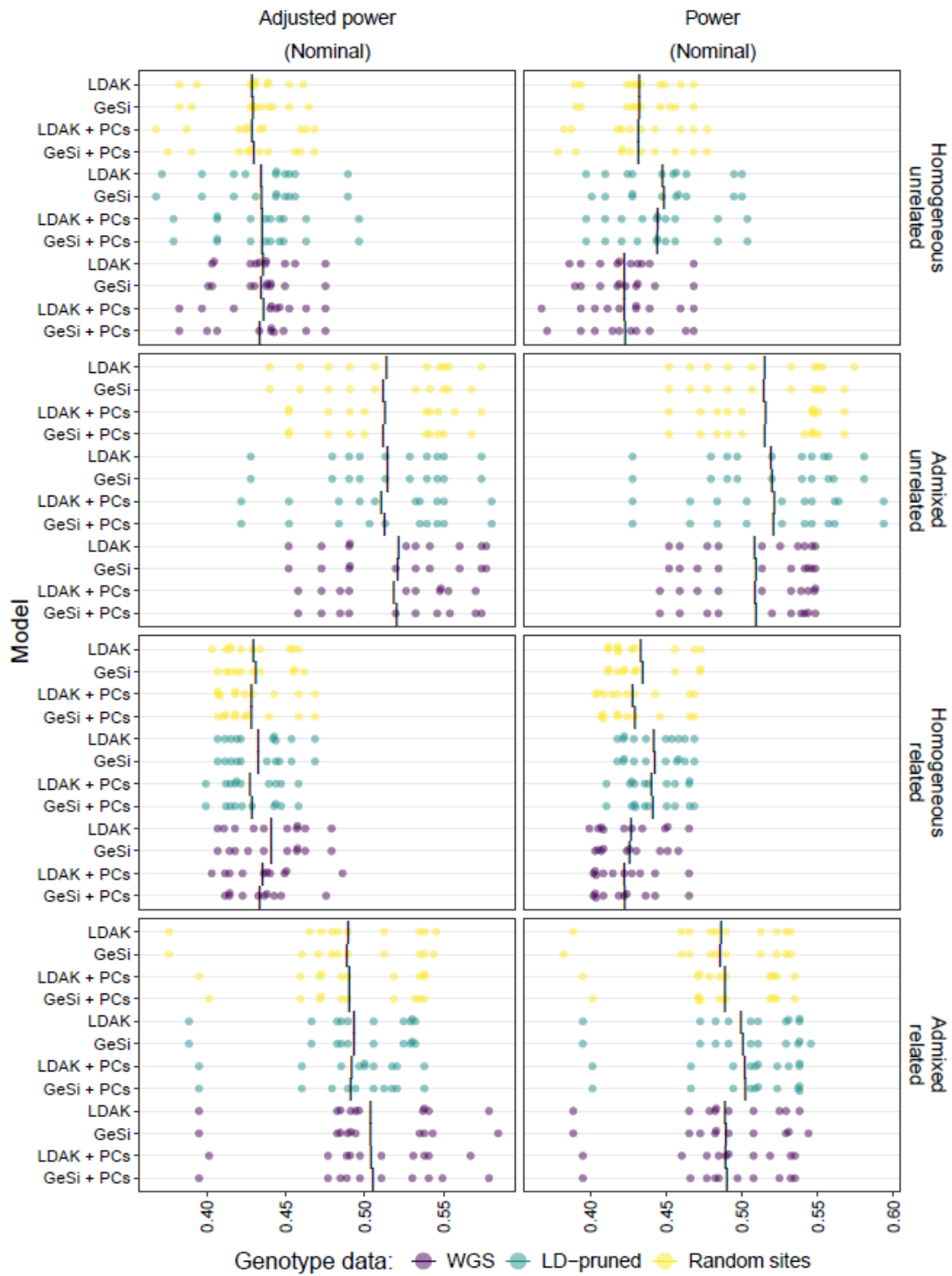


Figure 7. Type I error rate and inflation of the chi-square statistic in Simulated Set #3 ($\alpha = -1$). Similar to Suppl. Fig. 3, but for the Simulation Set #3 ($\alpha = -1$). Note that the GCTA GRM is equivalent to LDAK with $\alpha = -1$.



(Caption in next page)

Figure 8. Statistical power to detect a true causal variant in simulated set #2 ($\alpha = 0$). The first column shows the λ_{GC} -adjusted power to detect a true causal variant, and the second column shows the statistical power without adjusting for inflation. Each row represents one of the four different combinations of demographic and ascertainment scenarios that were simulated: A) Unrelated individuals from Iberians in Spain (IBS); B) Mixture of related and unrelated individuals from IBS; C) Unrelated individuals from Latino and Latin American populations (LAT; see main text and methods); and D) Mixture of unrelated and unrelated individuals from LAT. Within each panel, the vertical axis details whether the mixed linear model (MLM) used a GeSi or LDAK genetic relationship matrix, and whether principal components (PCs) were included in the model. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. A two-sided Wilcoxon rank sum test was used to compare, within each panel, all methods against the reference model (LDAK-LD-pruned + 10 PCs). However, none of the comparisons were significant and therefore are not shown (FDR-adjusted p-values > 0.05).

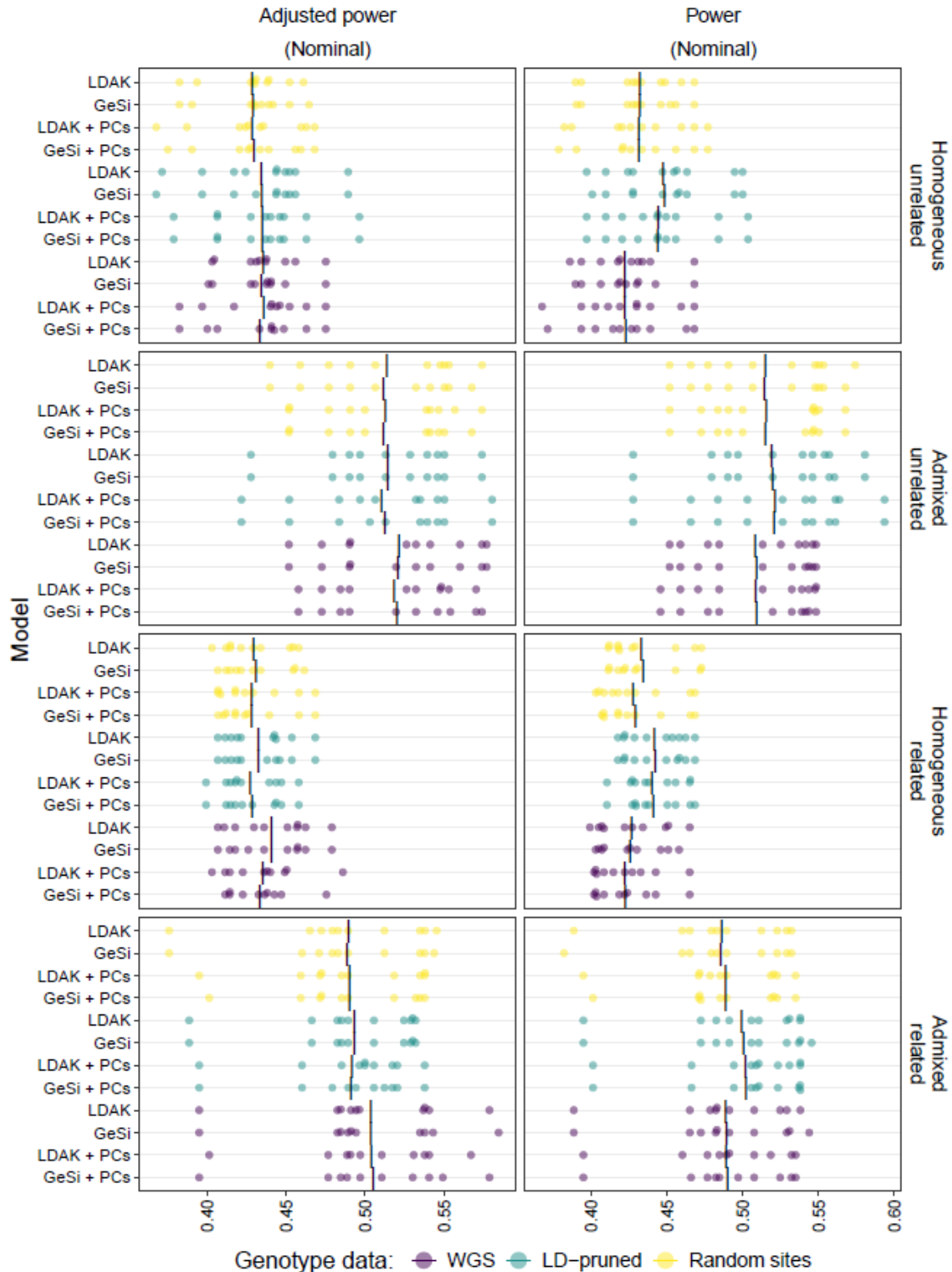


Figure 9. Statistical power to detect a true causal variant in simulated set #3 ($\alpha = -1$). Similar to Suppl. Fig. 5, but for the Simulation Set #3 ($\alpha = -1$). Note that the GCTA GRM is equivalent to LDAK with $\alpha = -1$.

3.3.2. Heritability does not Require Principal Components in the Absence of Environmental Confounding

Across both the $\alpha = 0$ (Simulation Set #2) and $\alpha = -1$ (Simulation Set #3) scenarios, heritability estimation was highly sensitive to the information density of the GRM (i.e. how the genotype data was pre-processed) but largely unaffected by the inclusion of PCs as covariates. GRMs calculated from WGS data consistently produced the most accurate and least biased heritability estimates. In contrast, GRMs built from either LD-pruned or randomly sampled (RS) sites resulted in significant downward bias (**Figure 9-Figure 13**).

The $\alpha = -1$ simulation allowed for a direct comparison with PC-Relate, which yielded the most downward-biased heritability estimates of any method tested. Even when compared to GeSi and GCTA/LDAK matrices built from the same LD-pruned data, PC-Relate's estimates were significantly lower, and including PCs as covariates did not correct this severe downward bias (**Figure 9-Figure 13**). This suggests that PC-Relate, by design, fails to capture the heritability component attributable to distant relatedness, which is removed when regressing the genotypes on the top principal components.

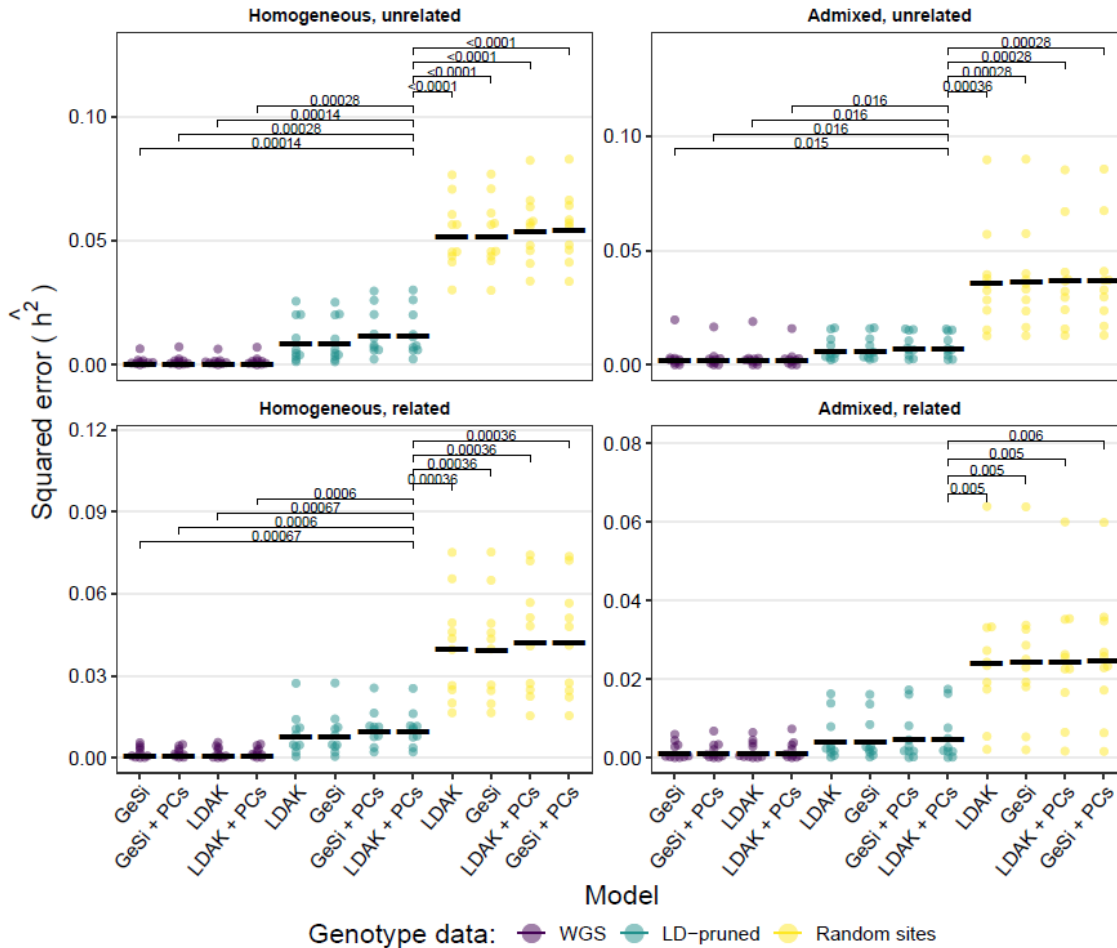


Figure 10. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = 0$ (Simulation Set #2.) The X axis details whether the mixed linear model (MLM) used a GeSi or LDAK genetic relationship matrix, and whether principal components (PCs) were included in the model. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. Each panel represents one of the four different combinations of demographic and ascertainment scenarios that were simulated: A) Unrelated individuals from Iberians in Spain (IBS); B) Mixture of related and unrelated individuals from IBS; C) Unrelated individuals from Latino and Latin American populations (LAT; see main text and methods); and D) Mixture of unrelated and unrelated individuals from LAT. A two-sided Wilcoxon rank sum test was used to compare all methods against the MLM model with 10 PCs and an LDAK matrix calculated from LD-pruned data. Models are displayed from best to worst within each panel as judged based on the mean squared error. A square bracket and FDR-adjusted p-value are shown for all comparisons with an FDR < 0.05 . Black horizontal lines: Mean squared error.

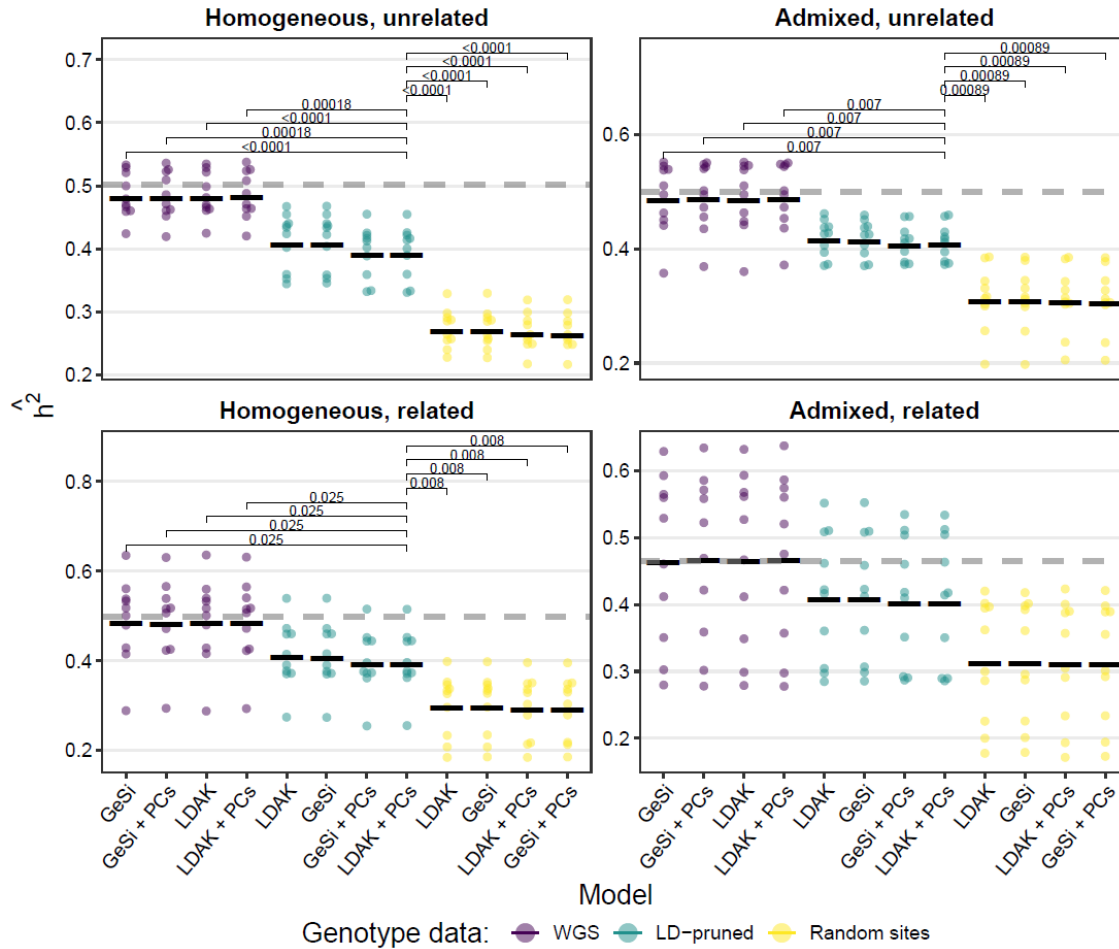


Figure 11. Heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = 0$. These heritability estimates accompany the squared errors shown in **Figure 10**. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = 0$ (Simulation Set #2.). Heritability estimates correspond to the mean-square errors shown in Figure 2. Black horizontal lines: Mean heritability estimate. Results based on Simulation Set #2.

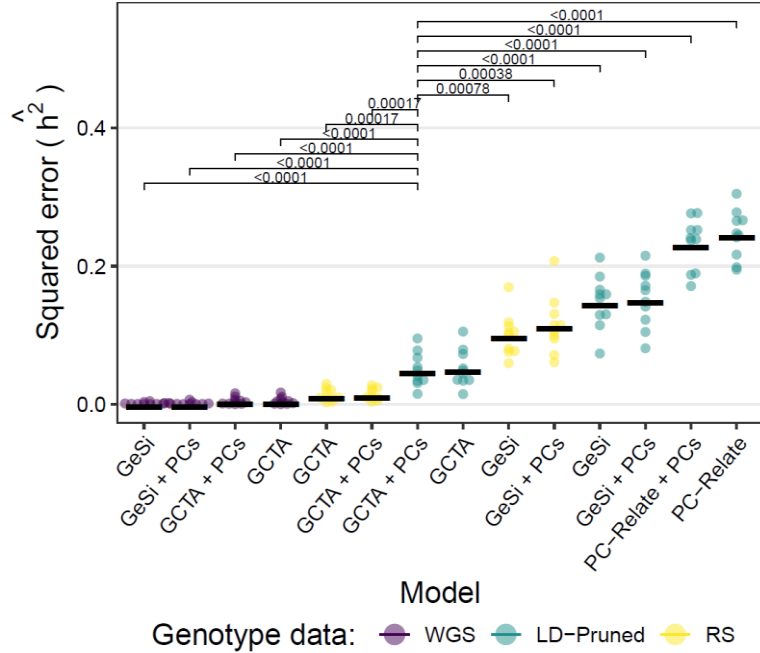


Figure 12. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = -1$ (Simulation Set #3.) The X axis details whether the mixed linear model (MLM) used a GeSi, LDAK or PC-Relate genetic relationship matrix, and whether principal components (PCs) were included in the model. All GRMs were calculated with $\alpha = -1$. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. Each panel represents one of the four different combinations of demographic and ascertainment scenarios that were simulated: A) Unrelated individuals from Iberians in Spain (IBS); B) Mixture of related and unrelated individuals from IBS; C) Unrelated individuals from Latino and Latin American populations (LAT; see main text and methods); and D) Mixture of unrelated and unrelated individuals from LAT. A two-sided Wilcoxon rank sum test was used to compare all methods against the MLM model with 10 PCs and an LDAK matrix calculated from LD-pruned data. Models are displayed from best to worst within each panel as judged based on the mean squared error. A square bracket and FDR-adjusted p-value are shown for all comparisons with an FDR <0.05 . Black horizontal lines: Mean squared error.

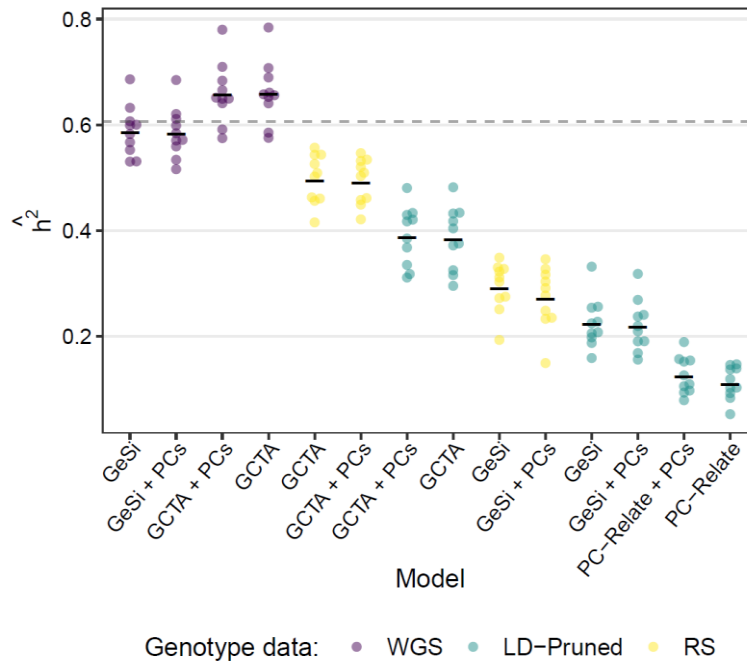


Figure 13. Distribution of heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = -1$. These results accompany the squared errors shown in **Figure 12**. Squared error of the heritability estimates under different levels of recent relatedness and admixture when the true scaling factor is $\alpha = -1$ (Simulation Set #3.). Heritability estimates correspond to the mean-square errors shown in Figure 2. Black horizontal lines: Mean heritability estimate.

3.3.3. Heritability Estimation in the Presence of Simulated Confounding

To validate our findings on more realistic genetic data and to assess the impact of confounding, I simulated phenotypes using genotypes from the BioMe cohort, first without (Simulation Set #4) and then with (Simulation Set #5) an environmental confounder correlated with population labels

The analysis of the non-confounded baseline data (Set #4) largely replicated the patterns observed in the Msprime simulations. Heritability estimates were most accurate when GRMs were calculated from WGS data, while using LD-pruned or RS data led to significant underestimation (**Figure 14a and b**). A notable difference emerged in the direction of bias for WGS-based GRMs: WGS-GCTA produced moderately upward-biased estimates, whereas WGS-GeSi was slightly downward-biased. As before, PC-Relate yielded the most severely underestimated heritability, and including PCs as covariates did not improve the accuracy for any GRM type in this non-confounded scenario (**Figure 14a and b**).

These patterns remained consistent after introducing the environmental confounder (Set #5). The relative performance of the methods and data types was unchanged, with WGS-GRMs remaining superior and PC-Relate performing the worst (**Figure 14c and d**). Crucially, even in the presence of a confounder determined by the demographic labels, the heritability estimates of the models with either a GeSi or GCTA GRM were not affected by the inclusion of PCs.

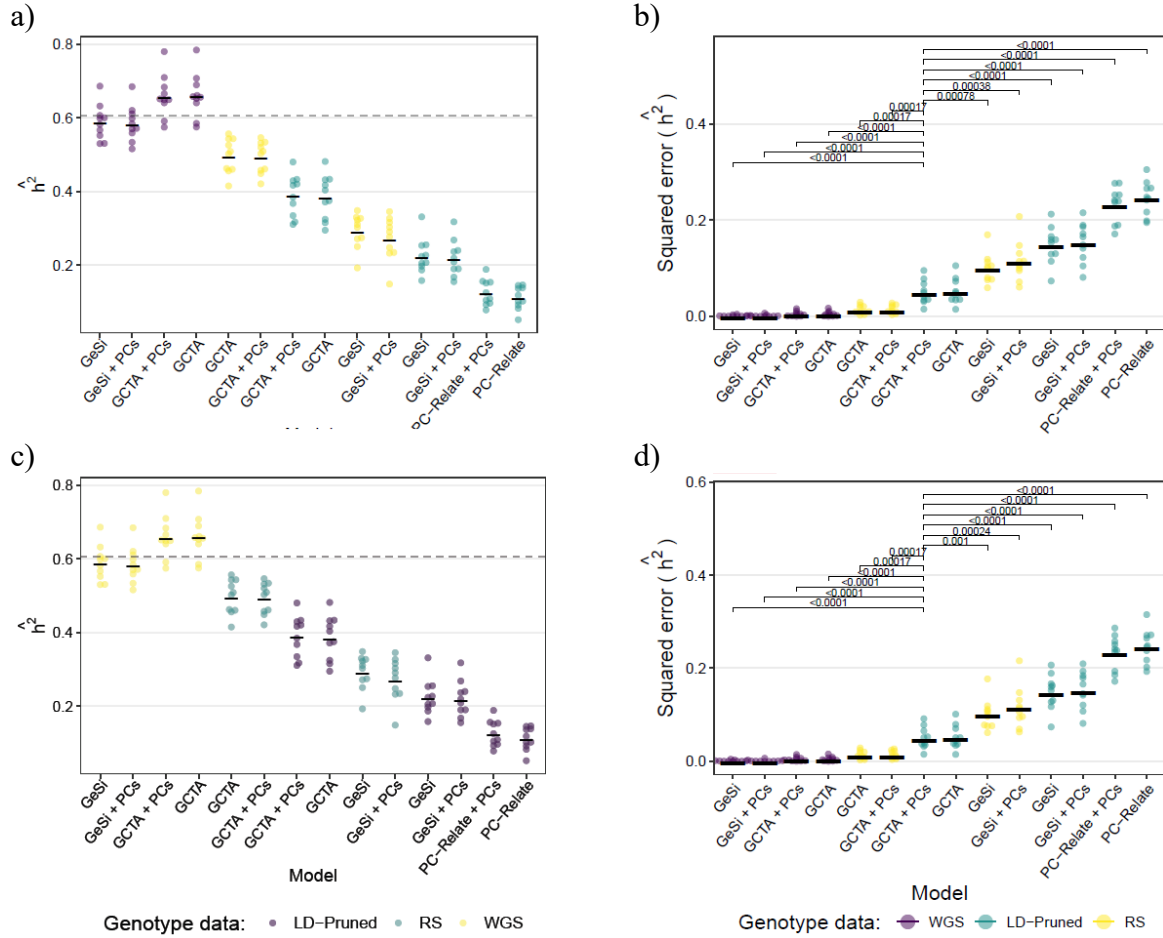


Figure 14. Square error and heritability estimate in simulations with and without confounding. Within each panel, the horizontal axis details whether the mixed linear model (MLM) used a GeSi, LDAK or PC-Relate genetic relationship matrix, and whether principal components (PCs) were included in the model. All GRMs were calculated with $\alpha = -1$. Ten replicates are shown for each model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. Panels **a** and **b**: Simulation set #4 (no confounding). Panels **c** and **d**: Simulation Set #5 (confounding determined by demographic labels). A two-sided Wilcoxon rank sum test was used to compare all models against an MLM with 10 PCs and a GCTA matrix calculated from LD-pruned data. Models are displayed from best to worst within each panel as judged based on the mean squared error. A square bracket and FDR-adjusted p-value are shown for all comparisons with an FDR < 0.05 . Black horizontal lines: Mean-squared error

3.3.4. Phenotype via the Best Linear Unbiased Predictor Equation

I used the BLUP equation to predict both the total genetic effects (g) and the phenotype (y) of the traits in the BioMe-based Simulation Sets #4 and #5 and the real height data of the BioMe cohort. The BLUP equation uses only the covariance encoded by a GRM as information to make predictions about the total polygenic scores. Therefore, I reasoned that the polygenic score prediction accuracy in a cross-validation framework could serve as an indicator of how much information is contained in a GRM. In the absence of confounding (Simulation Set #4), GeSi and the GCTA GRMs were the most accurate predictors of the true genetic value, particularly when calculated from WGS data. PC-Relate was by far the least accurate predictor of the total genetic effects g , confirming that it contains less genealogical information (**Figure 15a**). For predicting the total phenotype (y), however, the performance of PC-Relate improved to the level of other LD-pruned GRMs, but only when PCs were included as covariates (**Figure 15b**). Conversely, phenotype prediction accuracy for GeSi and GCTA models was not improved by the addition of PCs (**Figure 15b**). In fact, the accuracy of the GCTA models calculated from WGS data or random sites dropped after adding PCs to the model. Similarly, all MLMs using either a GeSi or GCTA GRM lost accuracy at predicting the genetic effects after adding PCs to the MLM (**Figure 15a**). This demonstrates that the information contained in the top PCs is essential for MLMs using a PC-Relate matrix but redundant if the MLM uses a GeSi or GCTA/LDAK matrix while there is no environmental confounding.

These findings remained virtually identical in the presence of a known environmental confounder that was included in the model (Simulation Set #5, **Figure 15c** and **Figure 15d**). The predictive accuracy for the genetic value, g , followed the same pattern, with GeSi and GCTA GRMs

outperforming PC-Relate (**Figure 15**). In this simulation, however, the strong effect of the confounder itself (which was included as a covariate in all models) drove a large portion of the phenotypic prediction accuracy (**Figure 15d**), which had the effect of largely masking the performance differences between models that were apparent in the non-confounded scenario, and in the genetic value prediction of the confounded scenario (**Figure 15c**).

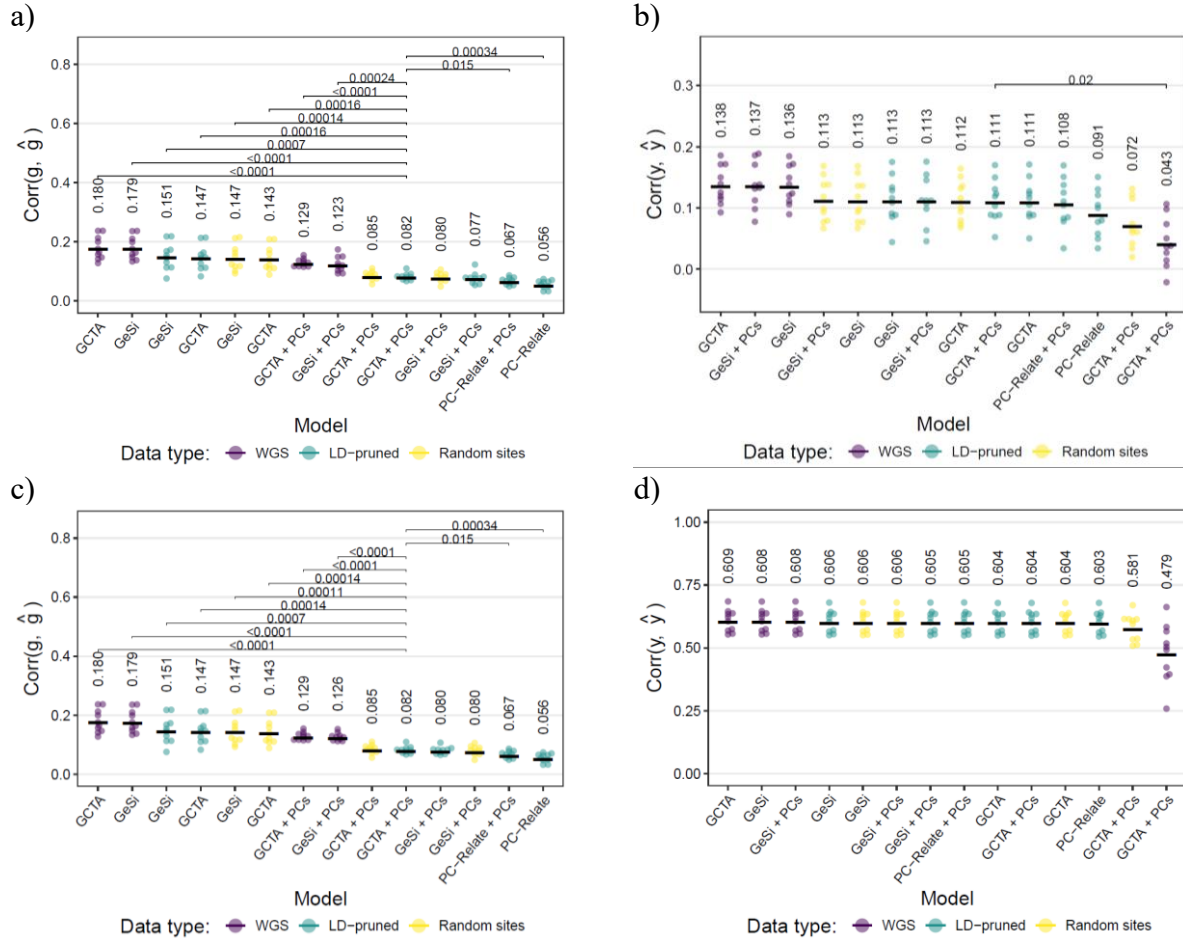


Figure 15. Accuracy of the predicted genetic value (\hat{g}) and phenotype (\hat{y}) via BLUP using different mixed linear model specifications in Simulated Set #4. Panels **a** and **b** correspond to Simulation Set #4 (no confounding), and panels **c** and **d**, to Simulation Set #5 (confounding determined by population labels). Panels **a** and **c** show the accuracy of the genetic value (g) prediction, and panels **b** and **d** the accuracy of the phenotype prediction. Each point represents the result of a cross-validation fold. The horizontal axis details whether the mixed linear model (MLM) used a GeSi, LDAK or PC-Relate genetic relationship matrix (GRM), and whether principal components (PCs) were included in the model. All GRMs were calculated with $\alpha = -1$. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. A two-sided Wilcoxon rank sum test was used to compare all models against an MLM with 10 PCs and a GCTA matrix calculated from LD-pruned data. Models are displayed from best to worst within each panel as judged based on the mean squared error. A square bracket and FDR-adjusted p-value are shown for all comparisons with an FDR < 0.05 . Black horizontal lines: Mean-squared error

3.3.5. Power to Detect Population Structure via Principal Components

A potential critique of the finding that PCs are redundant when used with full GRMs is that the sample size may be insufficient for the PCs to capture population structure. To formally address this, I tested whether the BioMe sample size provides enough statistical power based on the principal components theory developed by Patterson et al (68) and Bryc et al (69). Although defining populations in the highly admixed BioMe cohort is artificial, we can directly test the power to detect the specific structure that was explicitly simulated as the source of the confounding in Simulation Set #5.

I first calculated the differentiation between the population groupings (AFR, EUR, LAT) in the unrelated subjects of the BioMe cohort ($n = 9,190$), which yielded an F_{ST} of 0.024. Using the calculated F_{ST} to parameterize the detection threshold as $(1 + F_{ST})/2$, as proposed by Bryc et al. (69), a significance threshold of $t = 0.512$ was calculated for the normalized eigenvalues of the uncentered genotype covariance matrix. The analysis of the BioMe genotypes revealed seven normalized eigenvalues above this threshold (1986.7, 47.8, 8.1, 2.0, 1.6, 1.1, and 0.8).

This result demonstrates that the sample size provides enough statistical power to detect the genetic structure associated with the labels used to generate the confounding effects in Simulation Set #5. Therefore, the observed redundancy of PCs in the heritability and BLUP analyses is not an artifact of low power. This strengthens the conclusion that full GRMs, such as GeSi and GCTA, already contain the structural information captured by the top PCs. This is further corroborated by the consistently poor performance of the structure-corrected PC-Relate matrix when PCs are not included in the mixed linear models.

3.3.6. Prediction of Human Height in Real Data from the BioMe Cohort

To test whether PCs would also be unnecessary to predict phenotypes when there are other, unknown sources of confounding, I also used the BLUP-CV approach to predict human height in the BioMe cohort. Unlike the Simulation Set #5, where confounders were known, the true sources and mechanisms of confounding in the real BioMe data are unknown. The demographic labels were included as fixed-effect covariates in the mixed linear models, thus serving as imperfect proxies for the true, unobserved environmental and social factors that may be weakly or strongly correlated with ancestry or polygenic score. Because the true polygenic score was unknown, only the accuracy of the phenotype prediction is reported, which was assessed via the coefficient of determination (R^2) in a 20-fold cross-validation analysis. Because the true LDAK α parameter is unknown for real traits, I first compared all methods using a fixed value of $\alpha = -1$, corresponding to the GCTA and PC-Relate models, before identifying the optimal value for LDAK and GeSi matrices.

When α was fixed at -1, the results resembled those observed in our simulations without confounding. Specifically, most models using GeSi or GCTA GRMs performed similarly well, with the notable exception of the GCTA-WGS model, which lost accuracy when PCs were included in the model (**Figure 16a**). The PC-Relate model without PCs had the lowest predictive accuracy of all tested methods. However, its accuracy improved substantially, becoming comparable to other LD-pruned GRMs, after including PCs as covariates (**Figure 16a**).

These patterns were largely replicated when selecting the α value that maximized the cross-validation predictive accuracy for each method, although PC-Relate was kept fixed at $\alpha = -1$. (**Figure 16b**). In this optimal scenario, all models using either a GeSi or LDAK GRM achieved

similar predictive accuracy, with a slight, consistent but non-significant trend of WGS models to achieve higher accuracy than other models. For these methods, the inclusion of PCs as covariates had no impact on the predictive power. PC-Relate remained the only method that demonstrated a clear benefit from PC adjustment, and its model without PCs was the poorest performer overall. This confirms that even in a real-world setting with complex and unknown confounding, the information captured by the top PCs is redundant for GRMs that do not partition the genetic relatedness, but essential for partitioned GRMs like PC-Relate.

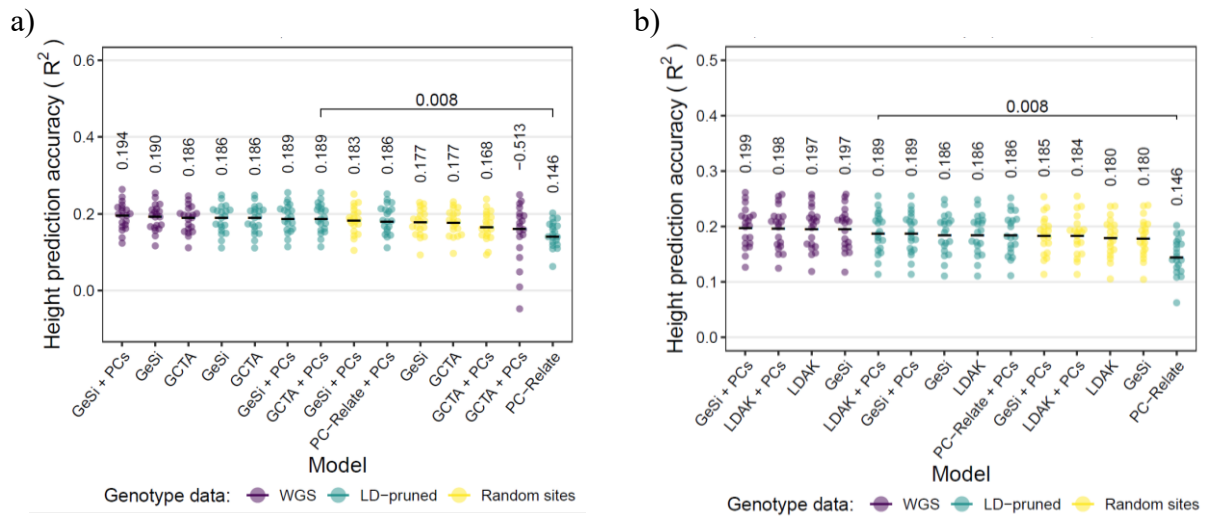


Figure 16. Accuracy of the height prediction via BLUP in real data from the BioMe cohort. **a)** Prediction accuracy keeping α fixed at -1. **b)** Prediction accuracy for the α value that maximizes the mean prediction accuracy. Each point represents the result of a cross-validation fold. The horizontal axis details whether the mixed linear model (MLM) used a GeSi, LDAK or PC-Relate genetic relationship matrix (GRM), and whether principal components (PCs) were included in the model. Each GRM was calculated from three different types of color-coded genetic data: Whole Genome Sequence (WGS) data, LD-pruned data, or randomly selected sites. A two-sided Wilcoxon rank sum test was used to compare all models against an MLM with 10 PCs and a GCTA matrix calculated from LD-pruned data. Models are displayed from best to worst within each panel as judged based on the mean squared error. A square bracket and FDR-adjusted p-value are shown for all comparisons with an FDR < 0.05. Black horizontal lines: Mean-squared error

3.3.7. Heritability Estimation of Human Height in the BioMe Cohort

I estimated the heritability (h^2) of height in the BioMe cohort to test the sensitivity of the heritability estimates to the inclusion of PCs on a real complex trait where the true genetic architecture and sources of confounding are unknown. Because the optimal α value is not known for real traits, and because its misspecification can impact the accuracy of h^2 estimates (see **Figure 17**), I performed an α scan for GeSi and LDAK models. Specifically, I selected the α value for each method that minimized the Akaike Information Criterion (AIC) of the MLM, which included sex, age, age², population label, and sex-by-population label interaction terms as covariates, and accounted for variance heteroscedasticity across sex-by-population groups. I ran all models in duplicate, with and without ten principal components as covariates. The GRMs were calculated with α values ranging between -2 and zero in increments of -0.25 (**Figure 18**).

The model with the best overall fit, as judged by the AIC, used a GeSi matrix calculated from WGS data with an $\alpha = -0.5$ and included PCs as covariates. Because the model was specified with heterogeneous residual variance among the sex-by-population groups, a different heritability estimate was obtained for each of them. **Figure 19** shows the sample size-weighted average heritability across all six groups sorted left to right from lowest to highest AIC. The group-specific estimates with the best inferred α value are presented in **Figure 20**.

Consistent with the results of Simulation Sets #2-5, models using GRMs calculated from WGS data yielded higher h^2 estimates than GRMs calculated from either LD-pruned or RS data. Models that included PCs consistently achieved a better fit than those without them (**Figure 19**). This suggests that at least one PC captured variance from an unmeasured confounder not fully accounted for by the fixed-effects covariates.

For a direct comparison with PC-Relate, I also estimated heritability with α fixed at -1 (**Figure 21**). In this analysis, the results mirrored the simulations' findings, namely, that after even after controlling for type of genotype data, PC-Relate yielded the lowest heritability estimates.

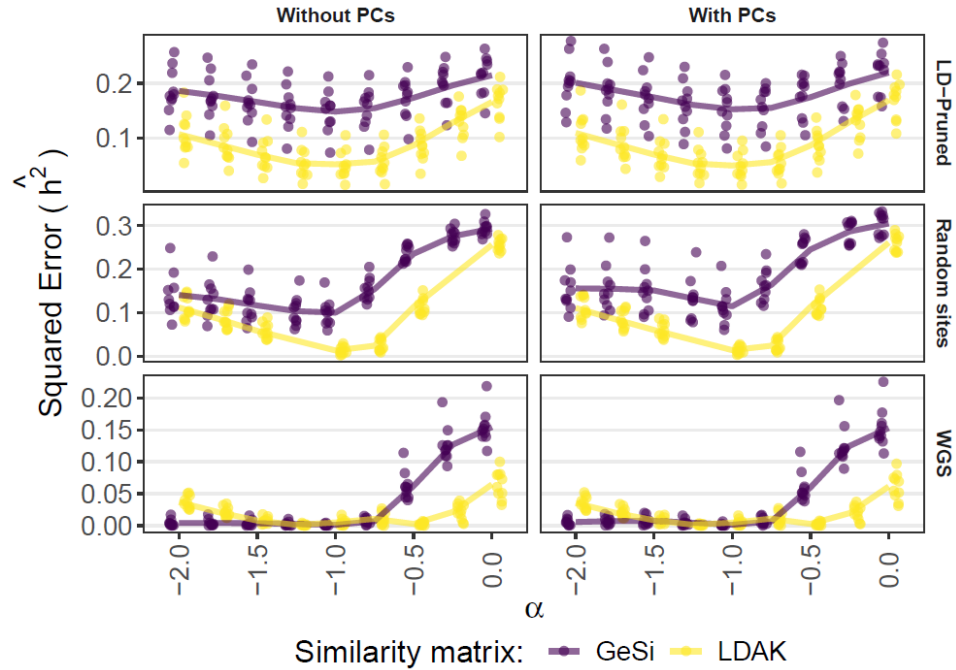


Figure 17. Effect of the α parameter on the squared error of the heritability estimates. Each panel shows the heritability estimate squared error obtained from mixed linear models fit under different conditions specified by the facet titles. Each row indicates the type of genotype data (LD-pruned, Random sites or WGS) used to calculate the GRMs in the model. Each column indicates whether the model included principal components (“Without PCs” and “With PCs”). Each dot is a replicate ($n=10$), and the solid lines connect the mean of the ten replicates of each GRM method (purple for GeSi, yellow for LDAK). The GCTA is equivalent to the LDAK model with $\alpha = -1$. Data from Simulation Set #4 (no confounding, true $\alpha = -1$).

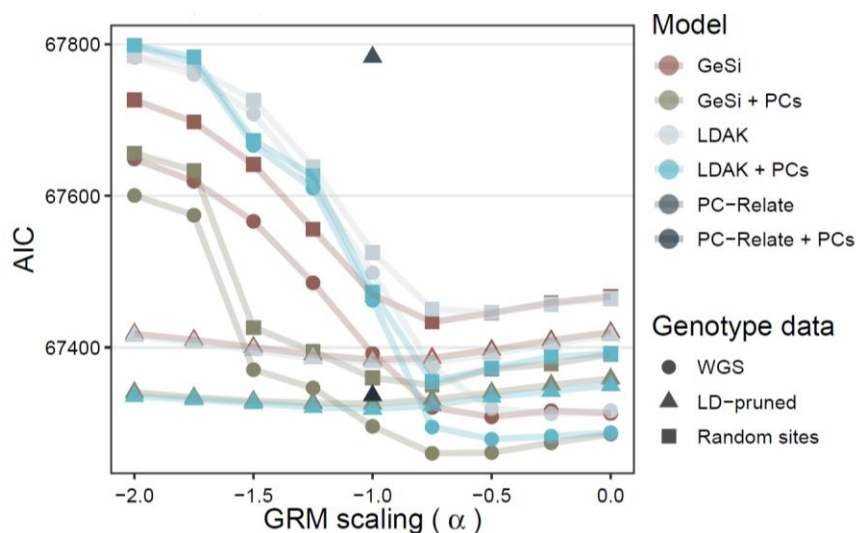


Figure 18. Selection of the α parameter for heritability estimation of human height. I defined six different lineal models based on the type of GRM they used (GeSi, LDAK/GCTA or PC-Relate) and on whether they used principal components as covariates or not. The GRM was calculated from the three different types of genotype data specified in the right-hand legend (WGS data, LD-pruned data or random sites), and with different α values, specified by the horizontal axis, ranging from -2 to 0 in increments of 0.25. I calculated the AIC for each of the mixed linear models defined by the combination of all these parameters. For each type of model and genotype data, I identified the α value that minimized the AIC of the model and used it to selected the values reported in **Figure 19**.

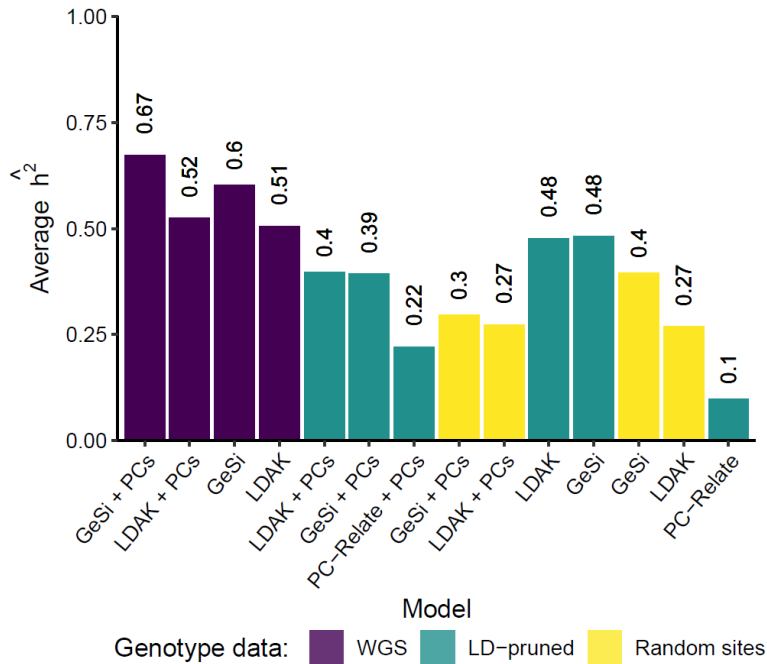
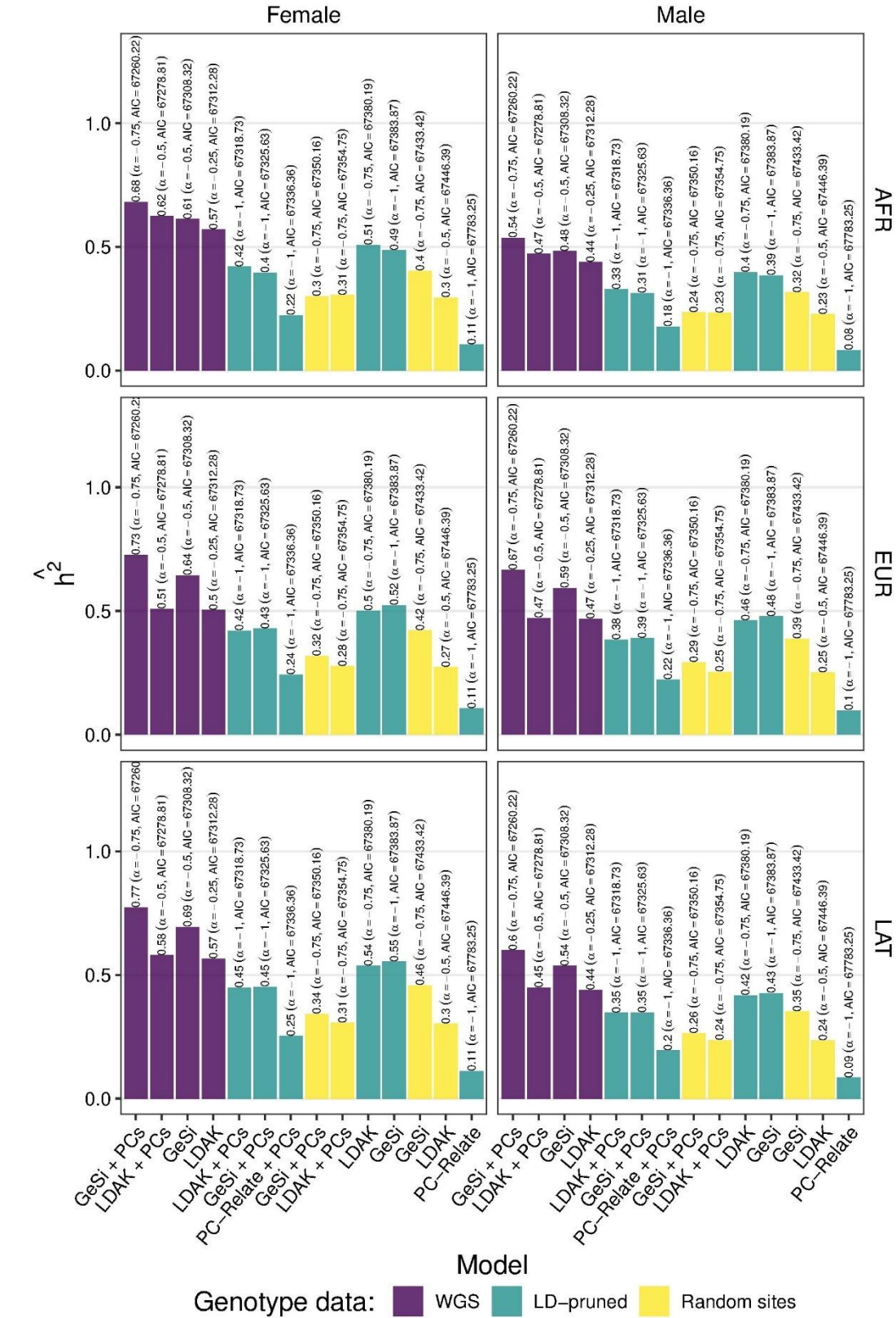


Figure 19. Average heritability estimates of human height. I identified the value of the α parameter that minimized the AIC of each model defined on the horizontal axis and data type indicated in the legend. The models are sorted from lowest to highest AIC from left to right. Heterogeneous residual variance across each of the sex-by-population groups was allowed, and thus six different heritability estimates were obtained for each model for each data type. The vertical axis shows the sample size-weighted average heritability estimate of the six sex-by-population groups. The group-specific estimates are shown in **Figure 20** (estimates with the α value that minimizes the AIC) and **Figure 21** (α fixed at -1). The selection of the α parameter value is explained in the main text and in **Figure 18**.



(Caption in next page)

Figure 20. Group-specific heritability estimates of human height based on the α value that minimizes the AIC. I identified the value of the α parameter that minimized the AIC of each model defined on the horizontal axis and data type indicated in the legend. The models are sorted from lowest to highest AIC from left to right. I allowed for heterogeneous residual variance across each of the sex-by-population groups, and thus obtained six different heritability estimates for each model for each data type. The vertical axis shows the group-specific heritability estimates for each combination of sex and population labels (AFR, EUR, LAT) specified on the facets' titles. The selection of the α parameter value is detailed in Suppl. Fig. 13.

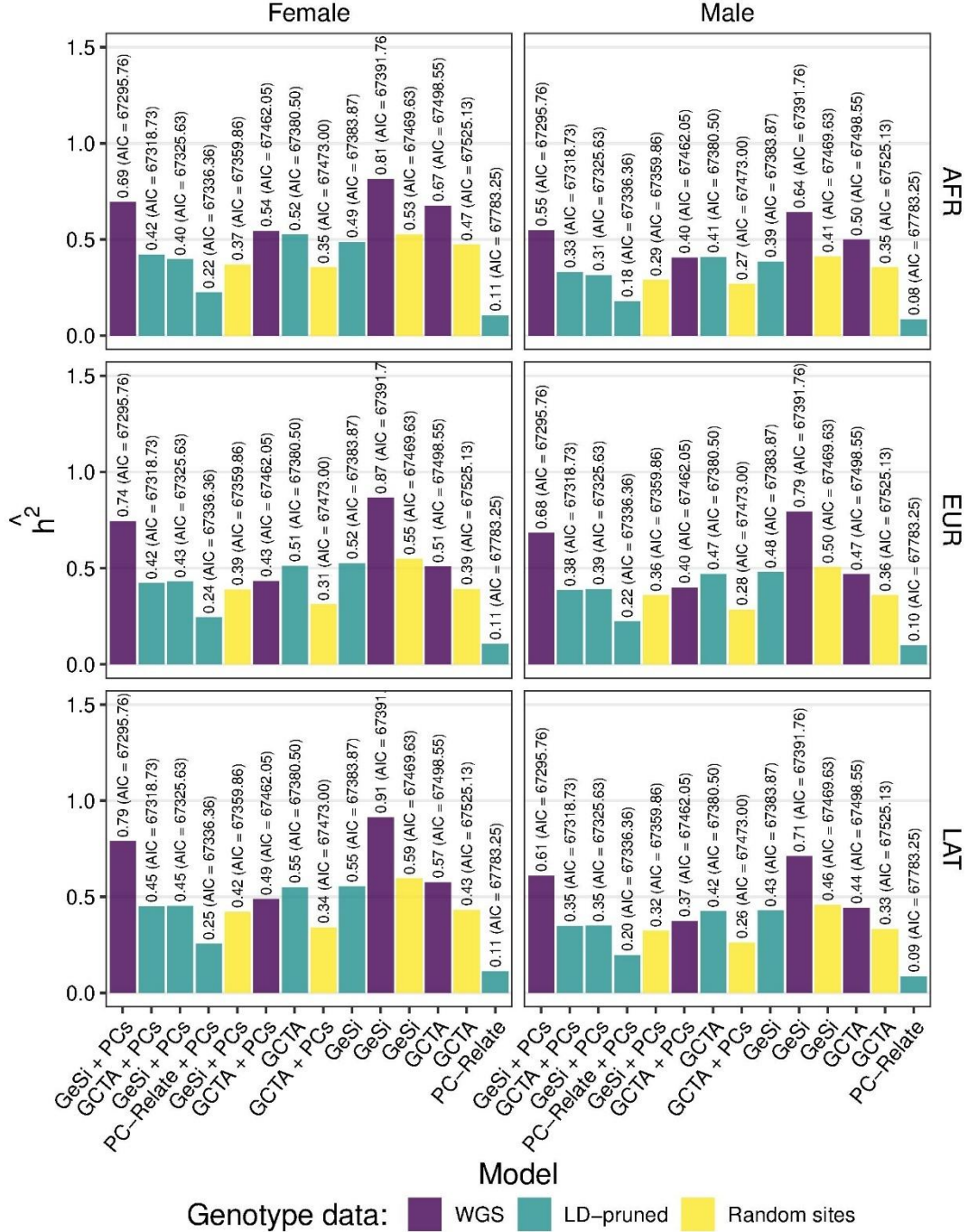


Figure 21. Group-specific heritability estimates of human height keeping α fixed at -1. Similar to Suppl. Fig. 14, except that the value of the α parameter was fixed at -1.

3.4. Discussion

The results presented here show the downstream effects of the differences in information content between full GRMs and shallow GRMs. The heritability was severely underestimated by PC-Relate, which, by design, presumably fails to capture the variance component attributable to distant relatedness (**Figure 11** and **Figure 13**). Furthermore, the BLUP analyses showed that full GRMs consistently outperformed PC-Relate at predicting an individual's total genetic value (**Figure 15a** and **c**). Moreover, neither the phenotype prediction (**Figure 15b** and **d**) nor the total genetic value prediction (**Figure 15a** and **c**) improved after adding principal components to mixed linear models containing a full GRM. Therefore, in the absence of confounding, full GRMs do not require principal components as covariates to account for population structure in mixed linear models, whereas the performance of PC-Relate, and presumably other shallow GRMs, is fundamentally dependent on the inclusion of PCs (**Figure 15**). The notion that a single GRM can model both recent and distant relatedness is further supported by recent evidence showing that genetic effect sizes for most complex traits are highly conserved across continental ancestries (20).

A potential limitation of our findings is that the redundancy of principal components when used with a full GRM could arise from insufficient statistical power to detect population structure. I formally evaluated this possibility using the theory of eigenanalysis (68,69). Although these frameworks are built on simplified models of discrete, non-hierarchical populations, they can be applied to test the power to detect the specific structure that was explicitly simulated as a confounder in Simulation Set #5. Using the provided continental labels to define populations, I calculated a between-group divergence of $F_{ST} = 0.024$. According to the theory proposed by Bryc

et al (69), this corresponds to a significance threshold of $t = 0.512$ for the normalized eigenvalues of the uncentered genotype covariance matrix. seven eigenvalues (1986.7, 47.8, 8.1, 2.0, 1.6, 1.1, and 0.8) substantially exceeded this threshold, confirming that the BioMe sample size provides sufficient power to resolve the major axes of genetic variation. Importantly, PCs are projections of the underlying matrix of pairwise time to the most recent common ancestor (TMRCA) (63). However, the eigenvectors of the GCTA GRM are equivalent to the full set of principal components (5), which further supports the notion that both the full set of PCs and the GCTA GRM contain the same information about the full genealogy.

The new framework of studying genetic relatedness as a continuum leads to a more nuanced interpretation of the role of PCs in practice. The common justification for including PCs is to correct for confounding, but this rests on the central paradox that the application of PCs often inverts the nature of the effects being modeled. Theoretical frameworks that give genetic meaning to PCs rely on unrealistic models of discrete populations (69,70), yet the complex, non-linear effects of the social environment are then modeled as a simple, continuous function of those PCs. This paradox arises from an informal but pervasive conflation of population structure with confounding itself (17). From a coalescent perspective, population structure is simply a source of genetic covariance that a full GRM appropriately models. The true confounding that biases association studies occurs when non-genetic factors, such as social constructs, cultural practices, and environmental exposures, are correlated with the patterns of genetic relatedness that PCs detect (64,71). Furthermore, adjusting for PCs can introduce collider bias, especially in admixed populations where PCs may capture multiple local genomic features rather than only genome-wide ancestry (26). Despite these theoretical limitations, PC adjustment has been, in practice, an effective strategy for mitigating inflation due to confounding.

This is consistent with our findings from the analysis of real height data. Unlike in our simulations where PCs were entirely redundant for full GRMs, their inclusion in the real-data analysis resulted in a better model fit for heritability estimation. This suggests that in real-world cohorts, PCs do not correct for unmodeled genetic effects but instead capture variance from unmeasured environmental or social confounders that are correlated with specific axis of variation (72). This correlation may be stronger than was captured by the broad population labels used in the confounding simulations. This underscores that the fundamental problem of using PCs in mixed linear models is not one of utility, but of interpretation.

I propose that the utility of principal components in mixed linear models stems not from its ability to model phenotype covariance due to population structure itself, but from its ability to partition genetic relatedness into specific ancestry components that can be individually correlated with social and environmental confounders. However, accurately modeling complex ancestry-social environment correlations is a major challenge. One recent approach attempts to infer clusters of shared environmental exposures directly from patterns of genetic relatedness, such as shared identity-by-descent (IBD) segments (73). A more direct solution, however, would be to prioritize the rich collection of socio-demographic and environmental data, allowing these potential confounders to be modeled explicitly rather than relying on PCs or inferred ancestry clusters as proxies.

The development of scalable ancestral recombination graph (ARG) inference methods has enabled reframing existing methods but has not yet led to a new definition or measure of relatedness that accounts for the full genealogy. This work of reinterpretation is best exemplified by two recent, parallel developments. First, the GCTA-based GRM is now understood to be a statistical estimate

of its true genealogical expectation, the eGRM, a specific application of the formal duality between genomic and genealogical statistics (74,75). Second, association testing has been reframed by replacing single variants with genealogical branches, using the ARG to test for latent variation (38). In these cases, the ARG has been used to provide a genealogical interpretation or re-framing of established statistical frameworks: the GRM and the single-locus association test, respectively. In contrast to reinterpreting existing methods, the GeSi framework is derived from the ground up by modeling the correlation of genetic effects given the coalescent times, and its resulting estimator, the cosine similarity of raw genotype vectors, is an emergent property, not a predefined inference target.

3.5. Strengths and Implications

The primary strength of this work lies in the step-wise, incremental validation of GeSi properties and usage across different conditions of increasing complexity, and with a diverse set of statistical techniques. I first validated that the GeSi matrix could be calculated from mean-centered genotype data (Simulation Set #2); I then tested the validity of the geometric extension which justified calculating GeSi with α values other than zero (Simulation Set #3). Next, I used CV-BLUP prediction framework as a way to quantify the amount of information contained within a GRM, which allowed me to confirm that the matrices classified as full-GRMs do indeed contain more information than the shallow PC-Relate matrix. I also confirmed that under weak confounding conditions, the principal components remain redundant in mixed linear models using a full GRM, but are required when using the shallow GRM PC-Relate. Finally, I used real data to assess whether PCs remain unnecessary in a condition where the true sources of confounding are not fully understood.

3.6. Limitations

While this study robustly demonstrates the properties of GeSi and clarifies the role of different GRM types, its simulations have some limitations. The environmental confounder modeled was a simplification of the true, unobserved social and environmental variables that influence complex traits. However, its purpose was conceptual: to demonstrate that when a known confounder correlated with population labels is explicitly included as a covariate, PCs are redundant for full GRMs. This isolates the effect of population structure and shows that it is not, by itself, the source of confounding that requires PC-based correction.

A key practical finding of this study was that GRMs calculated from WGS data yielded the most accurate heritability estimates, outperforming those from LD-pruned data, even when using GRMs GCTA that are typically calculated from LD-pruned data. I caution that this result may be partially influenced by the simulation design, which sampled causal variants uniformly across the genome. Our derivation of GeSi (see section 2.4.1) assumed that the number of causal mutations is linear with time, an assumption that holds under uniform sampling but may be violated in the human genome, where causal variants are known to be non-uniformly distributed and enriched in specific functional regions with distinct LD patterns and subject to different levels of natural selection (76,77). In such real-world scenarios, WGS data could lead to the over- or under-representation of heritability from certain genomic regions. Consequently, while LD-pruning can introduce its own biases, some form of LD-aware weighting or pruning would likely be necessary for both GeSi and other full GRMs in analyses of real data (16,31). This practical consideration, however, does not alter our fundamental conclusions that full GRMs contain more genealogical information than shallow GRMs, and that the primary role of PCs in a mixed model is not to model the genetic

covariance but to serve as proxies for non-genetic confounders correlated with specific axes of variation.

3.7. Future Directions

The results obtained here open several avenues of future investigation. The unexpected finding that the standard GCTA method, which is an average of per-site ratios, also performs well with WGS data requires further exploration. Future simulations should explore the robustness of both GeSi and GCTA to this finding under different genetic architectures where causal variants are not sampled uniformly but are instead enriched in regions of high or low LD. On a different note, the genealogical basis of GeSi makes it a promising tool for partitioning heritability into components explained by variants across different allele age bins, providing a more detailed understanding of a trait's genetic architecture.

3.8. Conclusions

In this work, I confirmed the validity of using a GeSi matrix in mixed linear models association testing and heritability prediction. This work also redefines the role of principal components (PCs) in mixed models. I demonstrate that when a full GRM is used in the absence of confounding, PCs are redundant for modeling genetic covariance due to population structure. Thus, their demonstrated utility in real-world analyses likely does not arise from modeling genetic effects, but from serving as proxies for unmeasured, non-genetic confounders correlated with ancestry. These findings directly challenge two widespread practices in the field: the arbitrary partitioning of genetic relatedness into recent and distant components, and the conflation of population structure with confounding.

4. Aim 3 - Development of a Flexible Phenotype Simulation Tool.

4.1. Introduction

Aims 1 and 2 established a new null model for the genetic architecture of complex traits, in which the genetic covariance from population structure is modeled by a “full” Genetic Relationship Matrix (GRM). This unified model, operationalized by the Coefficient of Genealogical Similarity (GeSi), is justified by evidence that causal effect sizes are highly conserved across populations and that poor polygenic score transferability results from statistical differences in LD and allele frequency, not from different biology (20,21). GeSi’s validity relies on the assumptions that causal genetic effects accumulate linearly with time, and that the non-genetic effects are uncorrelated with the genetic effects. I argue, that this model of unified genetic covariance is not to be treated as generally correct, but rather, that it can serve as a null hypothesis.

The formalization of this null model shifts the fundamental research question from how to statistically correct for population structure to when and why this unified model of genetic covariance is insufficient. This question requires studying how two different phenomena lead to violations of the assumptions made by the GeSi model. The first is “genetic confounding”, where non-neutral evolutionary processes like selection or assortative mating cause a non-linear accumulation of causal variants across the genealogy (17,18). The second is “environmental confounding”, where non-genetic factors are correlated with genetic ancestry or polygenic score, violating the assumption that genetic and environmental effects are independent. Quantifying the impact of these phenomena is necessary to guide practical decisions about what components to include in a mixed linear model and how to interpret its results. Studying when the unified model

fails requires simulated data where different true causal models of varying complexity can be generated and are known.

Systematically quantifying the impact of genetic and environmental confounding requires a simulation tool designed to disentangle these separate causal components. To model genetic confounding, such a tool must allow a user to control the distribution of causal variants along the genome and along different depths across the genealogy, thus simulating conditions where the assumption of linear-with-time accumulation of causal effects is violated. To model environmental confounding, it must provide a mechanism to simulate a non-genetic factor and explicitly define its correlation with genetic ancestry or the polygenic score. Critically, the tool's framework must separate the generative model from the inference model, to ensure that the simulated ground truth is decoupled from the statistical assumptions of the inference methods being benchmarked.

Existing phenotype simulation tools lack modules to specify different confounding mechanisms or use assumptions that prevent decoupling the generative and inference models. For instance, Tstrait can efficiently simulate quantitative traits on large-scale Ancestral Recombination Graphs, but its environmental component is limited to random noise that is independent of the genetic background (28). Furthermore, it cannot simulate phenotypes from user-provided genotype data, instead requiring trees stored in TreeSequence format (28,56). GCTA, PLNK and LDAK include phenotype simulation modules from user-supplied genotype data, but the genetic architecture they implement is overly simple, and they cannot simulate confounders correlated with ancestry or polygenic scores (5,31,59). SIMER, implements a genetic architectures with additive, dominant, and epistatic effects; however, it only simulates environmental components independently and does not provide a mechanism to create a non-genetic factor with a defined correlation to genetic

ancestry or a polygenic score (30). Other frameworks are designed for different purposes, such as multi-trait analysis or case-control studies, and likewise do not address the problem of confounding by population structure (78,79). PhenotypeSimulator implements a framework that prevents decoupling the generative model from the inference model (80). It simulates an infinitesimal genetic effect directly from a user-provided GRM, creating a circularity where the causal mechanism complies with the assumptions underlying the GRM's model that will be used for inference in a mixed linear model, making it impossible to test the robustness to violations of the GRM's model.

Phenocause, the tool described here, addresses this methodological gap. Phenocause is an R package for simulating phenotypes under a wide variety of complex genetic and non-genetic causal architectures, with special emphasis on mechanisms of confounding. The package implements a two-step workflow that largely separates the definition of the genetic architecture from the simulation of confounding effects. First, the `sample_causal_sites` function samples variants from user-provided genotype data, allowing for precise control over the causal architecture to test assumptions like the linear-with-time accumulation of effects. Second, the `simulate_phenotype` function uses these causal variants to generate a phenotype under explicit, user-controlled models of environmental or genetic confounding. This design provides the necessary ground truth to systematically benchmark the validity and limitations of genome-wide association studies.

4.2. Methods

4.2.1. Design and Implementation of the Phenocause Package

This section details the architecture and implementation choices for the phenocause R package.

4.2.1.1. Implementation and Dependencies

Phenocause was developed in the R programming language for integration with standard tools in statistical genetics. The package uses `data.table` for efficient data manipulation, reading and writing (81), `dplyr` for data-wrangling (82), and `SeqArray` for high-performance, memory-efficient access to genomic data stored in the GDS (genomic data structure) format (83). GDS files can be created from VCF file using the `seqVCF2GDS` function in the `SeqArray` package. More information about the GDS format can be found in the online tutorial at the phenocause GitHub repository:

<https://github.com/diegovelizo/phenocause>.

4.2.1.2. Modularity

The simulation framework is a two-step pipeline that separates most of the definition of the genetic architecture from the simulation of the phenotype. The first module, `sample_causal_sites()`, selects causal variants from genotype data. The second, `simulate_phenotype()`, uses these variants to generate a phenotype under a specified causal model. This modular design provides flexibility, allowing a single defined genetic architecture to be used in multiple, separate phenotype simulations.

4.2.1.3. Input and Output Formats

The package is designed to operate on genotype data stored in the Genomic Data Structure (GDS) format. This choice is driven by the computational efficiency of GDS for large-scale genomic datasets, as it allows for on-disk data chunking and filtering that minimizes memory usage. The package outputs standard R data frames that are compatible with downstream analysis tools. The

user is tasked with providing the paths to the GDS files, but all manipulation of them happens internally, thus preventing the user from needing to become familiar with the syntax to directly handle GDS files.

4.2.1.4. Wrapper-Based User Interface

To simplify usage, the `sample_causal_sites()` module uses a wrapper design pattern. Users interact with one of four mode-specific functions: `sample_causal_sites.uniform`, `sample_causal_sites.ldak`, `sample_causal_sites.custom`, and `sample_causal_sites.collider`, each with a minimal set of relevant parameters, which reduces the potential for error.

4.2.2. Module 1: Causal Sites Sampling (`sample_causal_sites`)

This module implements the first step of the pipeline, defining most of the genetic architecture-defining features by sampling causal variants from GDS files. The last genetic architecture feature, the LDAK α parameter, is defined in the second module detailed later.

4.2.2.1. Pre-sampling Filters

Before sampling, the pool of eligible variants can be filtered to ensure data quality or constrain the genetic architecture of the trait. Specifically, the function allows for the exclusion of variants based on a user-specified minor allele count (`mac_threshold`) or a variant missingness rate (`missingness_threshold`). The filters are efficiently applied to the connection to the GDS file on disk, without requiring loading the whole genotype data to memory at the same time. This is achieved by internally calling the `SeqSetFilter` function in the `SeqArray` package (83).

4.2.2.2. Uniform and Weighted Sampling Modes

The module supports four modes for defining the probability distribution of causal variants across the genome. In `uniform` mode, all variants that pass pre-sampling filters are sampled with equal probability. In `ldak_weights` mode, variants are sampled using pre-computed LD weights from the LDAK software (31). The sampling probability can be modified using the `weights_power` parameter to enrich for variants in high or low LD regions. LDAK weights range from 0 to 1, with variants in high LD regions receiving weights close to 0, and variants in low LD regions weights close to 1. Thus, as a rule of thumb, using negative `weights_power` values will enrich in high-LD variants, while using positive `weights_power` values will enrich in low-LD variants. The `custom_weights` mode allows variants to be sampled using any user-provided weights, enabling the simulation of architectures where causal status is correlated with any genomic feature.

4.2.2.3. Collider Sampling Mode

This mode is designed to generate data for studying collider bias. It first calculates the squared multiple correlation (R^2) between each variant and a user-provided matrix of principal components. It then stratifies variants by minor allele frequency (MAF) bins and samples two parallel sets of variants that are matched on MAF: a “structured” set from variants with high R^2 values with the PCA matrix, and a “non-structured” set from variants with low R^2 values. It then samples both causal and non-causal variants from each parallel set. This selection process of causal sites generates the specific data structure required to test the conditions under which adjusting for PCs induces collider bias (26), and the set of frequency- and R^2 -matched non-causal variants can be used as controls in benchmarking studies.

4.2.3. Module 2: Phenotype Simulation (`simulate_phenotype`)

This module uses the sampled causal genotypes to generate a final phenotype (Y) under an additive model composed of a mandatory base genetic causal component (g), an optional (non-genetic) confounder component (P), and a residual component e :

$$Y = g + P + e$$

Defining $g = X\beta$, and $P = \sum P_k$, this model is more specifically expressed as:

$$Y = X\beta + \sum_{k=1}^{k=K} P_k + e$$

Where $X \in \mathbb{R}^{(n \times m)}$ is the mean-centered matrix of alternate allele dosages for n subjects across m causal sites, $\beta \in \mathbb{R}^{(m \times 1)}$ is the vector of causal effects coefficients, and P_k are optional vectors of different confounding effects components, as detailed in section 4.2.3.2.

4.2.3.1. Base Genetic Model

The “total genetic effect” or “true polygenic score” (g) is the sum of dosages at causal variants, weighted by effect sizes (β). The genotype matrix is mean-centered so that the mean polygenic score in the sample is zero. The function handles missing genotype data through mean-imputation of small chunks of data at a time read from the GDS files via the `seqBlockApply` function in the `SeqArray` package (83).

The effect sizes β are first sampled from a standard normal distribution via the built-in R function `rnorm` and are then scaled using the LDAK α parameter that controls the dependency between

effect sizes and MAF. Thus, the final distribution of the effect sizes is $\beta_j \sim N(0, [2p_j(1 - p_j)]^\alpha)$, where $j = \{1 \dots p\}$ indexes the causal variants. The parameter α takes a value of -1 by default, which corresponds to the GCTA model where all variants explain the same amount of genetic variance, and rarer variants have a larger average effect than common variants.

The residual variance is scaled to meet the target heritability specified by the user via the equation $\sigma_e^2 = \frac{\sigma_a^2(1-h^2)}{h^2}$. Thus, the heritability is defined using only the additive genetic and residual variance components, as is typical in mixed linear models, while the variance explained by any fix-effect covariates is ignored.

4.2.3.2. Confounding Models

The framework implements three confounding models that can be used in any combination. If a categorical confounding effect is requested, the categorical variable must be supplied by the user, and the function `simulate_phenotype` assigns the effect sizes to each category so that the target user-supplied variance components values are met. If a quantitative confounder is requested, `simulate_phenotype` simulates the confounding variable from a normal distribution ensuring compliance with the user-specified variance components target values.

Meeting the user-requested target values requires deriving the effect size of the confounding variables, and in the case of the quantitative confounders, their residual variance. The step-by-step derivations are presented in section 4.2.3.4.

– **Categorical Confounder (cc):**

This mode assigns different effects $\beta_j^{(cc)}$ to each j category in a user-supplied categorical variable that is transformed into a one-hot encoded matrix $X^{(cc)}$, so that the categorical confounder effect component is $P^{(cc)} = X^{(cc)}\beta^{(cc)}$. The user specifies the target variance of confounding effect relative to the genetic variance $\left(\frac{\text{Var}(P^{(cc)})}{\text{Var}(g)}\right)$ via the `cc_relative_variance` parameter in the `simulate_phenotype` function.

The `simulate_phenotype` function distinguishes between nominal (input vector of class `character`) and ordinal (input vector of class `factor`) scales. For nominal variables, each category is assigned an independent random effect. For ordinal variables, the effects are simulated to be monotonically increasing across the factor levels in the order as they appear in `levels(input_vector)`, where `input_vector` is the user-provided vector containing the group assignments. This distinction can be useful to simulate categorical confounders with unsorted (e.g. zip code) or sorted (e.g. educational attainment) categories.

– **Environmental Quantitative Confounder Correlated with Polygenic Score (gc):**

This mode simulates a quantitative variable $X^{(gc)}$ that is correlated with the polygenic score g . The confounder effect caused by this variable is $P^{(gc)} = X^{(gc)}\beta^{(gc)}$.

The user specifies the target correlation between the simulated confounder and the polygenic score $\rho_{gc} := \text{Cor}(X^{(gc)}, g)$ via the `rho_gc` parameter, and the target variance of the simulated

confounding effect relative to the genetic variance $\left(\frac{\text{Var}(P^{(gc)})}{\text{Var}(g)}\right)$ via the `gc_relative_variance` parameter in the `simulate_phenotype` function.

– **Environmental Quantitative Confounder Correlated with Ancestry (ac):**

This mode simulates a quantitative variable $X^{(ac)}$ that is correlated with a user-supplied vector Q containing a measure of ancestry, such as a principal component or a continental ancestry estimate inferred via ADMIXTURE (84) or other tools. The confounder effect caused by this variable is $P^{(ac)} = X^{(ac)}\beta^{(ac)}$. The user specifies the target correlation between the simulated confounder and the ancestry component $\rho_{ac} := \text{Cor}(X^{(ac)}, Q)$ via the `rho_ac` parameter, and the target confounding effect variance relative to the genetic variance $\left(\frac{\text{Var}(P^{(gc)})}{\text{Var}(g)}\right)$ via the `ac_relative_variance` parameter in the `simulate_phenotype` function.

4.2.3.3. Liability Threshold Model for Binary Traits

If the user requests a binary quantitative trait (e.g. case/control status), the quantitative variable Y is treated as a liability score, and the final phenotype is assigned by applying a threshold to the liability score defined by the `prevalence` parameter. The liability score distribution is controlled via the `liab_dist`, which accepts the values `auto` (default), `gaussian` or `empirical`. The `gaussian` mode assumes that the liability score follows a standard normal distribution, so that the phenotype assignment is done by the inverse cumulative distribution function via the built-in R function `qnorm`. The `empirical` mode assigns the binary phenotypes using the percentiles of the liability distribution. The `auto` setting defaults to the empirical model when a categorical

confounder is present to account for potential non-normality in the liability distribution, and to gaussian otherwise.

4.2.3.4. Derivation of the Confounder Parameters

The mathematical derivation of the categorical confounder was presented in the methods section of Aim 2. Here I present the mathematical derivation of the quantitative confounder.

– Quantitative Confounder Correlated with Polygenic Score

Using the same definitions provided before, let:

$$Y := g + P^{(gc)} + e \quad \text{Equation (22)}$$

$$P^{(gc)} := X^{(gc)}\beta^{(gc)} \quad \text{Equation (23)}$$

The goal is to simulate the variable $X^{(gc)}$, which must be simulated from g and comply with the user-provided target values. Define the generative equation:

$$X^{(gc)} := b^{(gc)}g + e^{(gc)} \quad \text{Equation (24)}$$

Where $b^{(gc)}$ is a scalar defining the relationship between the polygenic score and the confounder, and $e^{(gc)} \sim N(0, \sigma_{gc}^2)$ is random noise. The goal is to find $b^{(gc)}$ and σ_{gc}^2 . Remember the user-provided parameters:

$$w_{gc} := \frac{\text{Var}(P^{(gc)})}{\text{Var}(g)} \Rightarrow \text{Var}(P^{(gc)}) = w_{gc}\text{Var}(g)$$

$$\rho_{gc} := \text{Cor}(X^{(gc)}, g)$$

Replace Equation (24) into Equation (23):

$$P^{(gc)} = (b^{(gc)}g + e^{(gc)})\beta^{(gc)} \quad \text{Equation (25)}$$

Take the variance on both sides of Equation (23):

$$\text{Var}(P^{(gc)}) = \text{Var}(X^{(gc)}\beta^{(gc)}) = (\beta^{(gc)})^2 \text{Var}(X^{(gc)}) \quad \text{Equation (26)}$$

Replace definition of w_{gc} into equation (26):

$$w_{gc}\text{Var}(g) = (\beta^{(gc)})^2 \text{Var}(X^{(gc)})$$

$$SD(X^{(gc)}) = \frac{\sqrt{w_{gc}}SD(g)}{|\beta^{(gc)}|} \quad \text{Equation (27)}$$

Where SD is used as a shortcut for square root of the variance. Replace Equation (24) into the user-defined correlation:

$$\rho_{gc} := \text{Cor}(X^{(gc)}, g) = \text{Cor}((b^{(gc)}g + e^{(gc)}), g) = \frac{b^{(gc)}\text{Var}(g)}{SD(X^{(gc)})SD(g)} = \frac{b^{(gc)}SD(g)}{SD(X^{(gc)})}$$

$$\rho_{gc} := \frac{b^{(gc)}SD(g)}{SD(X^{(gc)})}$$

Replace in Equation (27):

$$b^{(gc)} = \frac{\rho_{gc}\sqrt{w_{gc}}}{|\beta_{gc}|} \quad \text{Equation (28)}$$

To obtain σ_{gc}^2 , take the variance on both sides of the equation (23) and replace in the definition of w_{gc} :

$$w_{gc}Var(g) = (\beta^{(gc)})^2 Var(X^{(gc)})$$

$$w_{gc}Var(g) = (\beta^{(gc)})^2 (b^{(gc)})^2 Var(g) + (\beta^{(gc)})^2 Var(e^{(gc)})$$

After some algebra, obtain:

$$Var(e^{(gc)}) := \sigma_{gc}^2 = Var(g) \left(\frac{w_{gc}}{(\beta^{(gc)})^2} - (b^{(gc)})^2 \right) \quad \text{Equation (29)}$$

The genetic variance can be calculated from the base genetic model. Thus, the only missing parameter to calculate b_{gc} and σ_{gc}^2 in Equations (28) and (29) is $\beta^{(gc)}$, which can take any arbitrary real value and must be supplied by the user via the `gc_coefficient`. The specific value of $\beta^{(gc)}$ has no effect on the causal components of the trait, thus the user can supply any random value from, for instance, a standard normal distribution, or can keep it fixed at any arbitrary constant to ensure reproducibility.

– Quantitative Confounder Correlated with an Ancestry Component

Proceed in a similar manner as before. Let:

$$Y = g + P^{(ac)} + e \quad \text{Equation (30)}$$

$$P^{(ac)} = X^{(ac)} \beta^{(gc)} \quad \text{Equation (31)}$$

The goal is to simulate the variable $X^{(ac)}$, which must be simulated from the vector Q containing the values of an ancestry component, and comply with the user-provided target values. Define the generative equation:

$$X^{(ac)} := b^{(ac)}Q + e^{(ac)} \quad \text{Equation (32)}$$

Where $b^{(ac)}$ is a scalar defining the relationship between the ancestry component Q and the confounder, and $e^{(ac)} \sim N(0, \sigma_{ac}^2)$ is random noise. The goal is to find $b^{(ac)}$ and σ_{ac}^2 . Remember the user-provided parameters:

$$w_{ac} := \frac{\text{Var}(P^{(ac)})}{\text{Var}(g)} \Rightarrow \text{Var}(P^{(ac)}) = w_{ac}\text{Var}(g)$$

$$\rho_{ac} := \text{Cor}(X^{(ac)}, Q)$$

Replace Equation (32) into Equation (31):

$$P^{(ac)} = (b^{(ac)}Q + e^{(ac)})\beta^{(ac)} \quad \text{Equation (33)}$$

Take the variance on both sides of Equation (31):

$$\text{Var}(P^{(ac)}) = \text{Var}(X^{(ac)}\beta^{(ac)}) = (\beta^{(ac)})^2 \text{Var}(X^{(ac)}) \quad \text{Equation (34)}$$

Replace definition of w_{ac} into equation (34):

$$w_{ac}\text{Var}(g) = (\beta^{(ac)})^2 \text{Var}(X^{(ac)})$$

$$SD(X^{(ac)}) = \frac{\sqrt{w_{ac}}SD(g)}{|\beta^{(ac)}|} \quad \text{Equation (35)}$$

Replace Equation (32) into the user-defined correlation:

$$\rho_{ac} := \text{Cor}(X^{(ac)}, Q) = \text{Cor}\left((b^{(ac)}Q + e^{(ac)}), Q\right) = \frac{b^{(ac)}\text{Var}(Q)}{SD(X^{(ac)})SD(Q)} = \frac{b^{(ac)}SD(Q)}{SD(X^{(ac)})}$$

$$b^{(ac)} := \frac{\rho_{ac}SD(X^{(ac)})}{SD(Q)}$$

Replace in Equation (35):

$$b^{(ac)} := \frac{\rho_{ac}SD(X^{(ac)})}{SD(Q)} = \frac{\rho_{ac}SD(P^{(ac)}/\beta^{(ac)})}{SD(Q)}$$

$$b_{ac} = \frac{\rho_{ac}\sqrt{w_{ac}}SD(g)}{|\beta^{(ac)}|SD(Q)} \quad \text{Equation (36)}$$

To obtain σ_{ac}^2 , take the variance on both sides of the equation (32) and replace in the definition of

w_{ac} :

$$\text{Var}(X^{(ac)}) = \text{Var}(b^{(ac)}Q) + \text{Var}(e^{(ac)})$$

$$\text{Var}(e^{(ac)}) = \text{Var}(X^{(ac)}) - (b^{(ac)})^2\text{Var}(Q)$$

Replace in Equation (35):

$$\text{Var}(e^{(ac)}) := \sigma_{ac}^2 = \frac{w_{ac}\text{Var}(g)}{(\beta^{(ac)})^2} - (b^{(ac)})^2\text{Var}(Q) \quad \text{Equation (37)}$$

The variance of the ancestry component Q can be directly calculated from the user-supplied vector.

Likewise, the genetic variance can be calculated from the base genetic model. Thus, the only

missing parameter to calculate b_{ac} and σ_{ac}^2 in Equations (36) and (37) is $\beta^{(ac)}$, which can take any arbitrary real value and must be supplied by the user via the `ac_coefficient` parameter. The specific value of $\beta^{(ac)}$ has no effect on the causal components of the trait, thus the user can supply any random value from, for instance, a standard normal distribution, or can keep it fixed at any arbitrary constant to ensure reproducibility.

4.2.4. Accompanying Data and Distribution

The package includes example data that can be used to follow the tutorial shared via GitHub. Phenocause is distributed as a free and open-source R package under an MIT license, with the source code, documentation and a tutorial publicly available on GitHub at <https://github.com/diegovelizo/phenocause>.

4.2.4.1. Example Genetic Data Distributed with Phenocause

The raw DNA sequences and genealogies were generated using a Msprime (27). Specifically, a demographic model of Latin American and reference continental populations was used (20,23) to generate data resembling the populations in the 1000 Genome Project (85) and Mexicans with indigenous American ancestry in the Mexican Biobank (86). A total of 5,000 individuals were simulated from eight populations: Han Chinese in Beijing, China (CHB); Colombian in Medellin, Colombia (CLM); Iberian in Spain (IBS); Indigenous American Mexicans in the Mexican Biobank (MXB); Mexican Ancestry in Los Angeles, California (MXL); Peruvian in Lima, Peru (PEL); Puerto Rican in Puerto Rico (PUR); and Yoruba in Ibadan, Nigeria (YRI).

To keep the data size manageable, only the first one-third of the genetic length (in cM) of the human chromosomes 20 to 22 was simulated, using the human recombination maps for

chromosomes for genome build GRCh38. The Discrete-Time Wright-Fisher (DTWF) model was used. This model was chosen over the standard Hudson model because it more accurately handles large sample sizes and recent admixture events, preventing known genealogical distortions caused by the Hudson model (57,87). A constant mutation rate of 1.25×10^{-8} per base pair per generation was used, with the Jukes-Cantor (JC69) nucleotide substitution model (58) .

The resulting tree sequence data was exported to VCF format using the `write_vcf` function in Tskit (36). Bcftools (88) was used to remove multiallelic sites and variants with $MAF \leq 1\%$. Sites in high LD were removed using the `--indep-pairwise` function in Plink v1.9 (59), with a window size of 500 sites, steps of 50 sites, and r^2 threshold of 0.2. Finally, the pruned VCF for each chromosome was converted to the GDS format using the `seqVCF2GDS` function in the SeqArray package (83).

4.2.4.2. Metadata and LD Weights

The package includes two data objects that accompany the simulated genetic data.

- `phenocause.metadata`: A data frame containing sample-level information for the 5,000 simulated individuals, including a unique sample ID, the simulated population label, and the first 20 principal components calculated from the associated genetic data using GCTA.
- `phenocause.ldak_weights`: A list of data frames containing pre-computed Linkage Disequilibrium (LD) weights calculated with the `--calc_weights` function in the LDAK software (31). Each data frame corresponds to a chromosome in the example dataset and contains the variant's ID and its corresponding LD weight. These weights can be used in the

`sample_causal_sites.ldak` function in phenocause to simulate genetic architectures where the probability of a variant being causal is related to its local LD.

4.2.5. Usage Demonstration and Validation

To demonstrate the functionality of phenocause and validate its implementation, four simulation experiments were designed and run using the example data and metadata distributed with the package. In all the examples, the heritability was fixed at 0.4 and the missingness threshold was set at 0.02 unless otherwise specified. More examples are shown in the online tutorial distributed with the package. The full code to run these examples is deposited on https://github.com/diegovelizo/phenocause/extra_examples/.

4.2.5.1. Example 1: Sampling Causal Sites with Varying Levels of Linkage Disequilibrium

This example demonstrates the weighted and uniform sampling modes of the `sample_causal_sites` function. Three sets of 200 causal sites were sampled from the example genotype data in a single replicate. The first set was sampled with uniform probability. The second and third sets were sampled using the provided LDK weights to simulate LD depletion (`weights_power=4`) and LD enrichment (`weights_power=-0.25`), respectively. Low-frequency sites were filtered out with the parameter `mac_threshold=100`, and sites with high missingness rate were removed by setting `missingness_threshold=0.02`. The results are presented as an overlaid density plot showing the distribution of LD scores for the sites selected under each of the three sampling modes.

4.2.5.2. Example 2: Confounder Correlated with the Polygenic Score

This example demonstrates the simulation of a quantitative trait with a confounder correlated with the polygenic score. Thirty replicates were run. In each replicate, 200 causal sites were first sampled with uniform probability. Phenotypes were then simulated using the `simulate_phenotype` function with the LDAK's `alpha=-1`. A quantitative confounder was included with a defined correlation to the true polygenic score (g) via the parameter `gc_rho=0.3`, and a variance explained by the confounder relative to the genetic component set to `gc_relative_variance=0.2`. The results are presented as distributions of the observed heritability, the relative variance explained by the confounder, and the correlation between the polygenic score and the confounder, each compared to the corresponding pre-specified target parameter value. In addition, a scatter plot was used to compare the observed vs the expected squared effect sizes given the minor allele frequency. Specifically, the expected squared effect size of variant j was calculated as $E[\beta_j^2] = [2p_j(1 - p_j)]^\alpha$.

4.2.5.3. Example 3: Binary Trait with a Categorical Confounder in Ordinal Scale

This example demonstrates the simulation of a binary trait with an ordinal categorical confounder and non-uniform distribution of causal sites along the genome. Thirty replicates were run. In each replicate, causal sites were sampled with LDAK weights to enrich for variants in high-LD regions (`weights_power=-0.25`), and their effect sizes were simulated with `alpha=-0.50`. Phenotypes were simulated as a binary outcome with `prevalence=0.15`, using the empirical liability distribution to define the liability threshold for cases via the argument `liab_dist="empirical"`. The categorical confounder was the vector of population labels, ordered by their mean value along PC1 to create an ordinal scale. The variance explained by the categorical confounder relative to

the genetic effects was set via the argument `cc_relative_variance=0.2`. The results are presented by plotting the distribution of prevalence and polygenic scores stratified by population label across the 30 replicates. In addition, the PC1 is also shown stratified by population label.

4.2.5.4. Example 4: A Factorial Experiment to Isolate Collider Bias

This experiment uses a 2x2x2 factorial design to attempt to isolate the statistical signal of PC-induced collider bias from that of the baseline covariance created by population structure. Specifically, this experiment showcases how one would try to replicate and generalize the findings principal components estimated from genetic data without LD pruning can induce collider bias due to certain PCs capturing local genomic features as opposed to genome-wide ancestry (26). According to Grinde et al., collider bias would arise if the following conditions are met:

- There is a causal site correlated with a principal component.
- There is a non-causal site correlated with the same principal component.
- The principal component captures local genomic features as opposed to genome-wide ancestry.

In the work by Grinde et al., the third condition is met by calculating the principal components from genetic data that were not LD-pruned. However, for this example I use the genetic data distributed with the `phenocause` package, which was simulated under a neutral model, thus, there is no evolutionary force that could induce specific regions of the genome to be overly differentiated, with respect to the genome-wide average, across populations. Therefore, this example is more a demonstration of how to tackle complex problems using `phenocause` rather than a strict attempt to replicate Grinde et al. findings.

The three factors analyzed in this study were:

- The causal architecture of the phenotype: Determined by either PC-correlated (“structured”) causal variants or non-PC-correlated (“non-structured”) causal variants.
- Data used to calculate the principal components used to defined whether a genetic variant is structured or non-structured sites: Either LD-pruned or non-LD-pruned.
- Data used to calculate the principal components included as covariates in the GWAS model: Either LD-pruned or non-LD-pruned.

In order to detect the collider bias described by Grinde et al., we would need to observe a higher inflation in the scenario where a) the causal variants are correlated with principal components calculated from non-LD pruned data, and b) the GWAS model included PCs calculated from non-LD-pruned data. If swapping either set of PCs to another set calculated from LD-pruned data reduces the statistical inflation of the structured non-causal variants, it would be evidence of collider bias.

The `sample_causal_sites.collider` function was used to generate the four variant sets. Using the first ten PCs as the structural component (potential colliders), 100 “structured” causal sites and 100 “non-structured” causal sites were sampled. Two corresponding MAF-matched sets of non-causal variants were also sampled to serve as *test sets* for measuring spurious associations.

One hundred replicates were run. In each, two phenotypes ($h^2=0.5$) were simulated: one from the structured causal set and one from the non-structured set. Each phenotype was then analyzed in a multiple linear regression model including as covariates ten PCs calculated either from LD-pruned or non LD-pruned genotype data. The Type I nominal error rate (significance level = 0.05) and the

genomic inflation factor (λ_{GC}) were calculated separately for the non-structured and structured non-causal sites as measures of inflation under the null hypothesis of no association.

4.3. Results

With the purpose of validating and showcasing the flexibility of the causal genetic and non-genetic models that can be simulated by phenocause, four examples are presented next. It should be noted, however, that this is a non-exhaustive demonstration of the capabilities of phenocause. More examples are included in the online tutorial shared via phenocause's GitHub repository at <https://github.com/diegovelizo/phenocause>.

4.3.1. Example 1: Sampling Causal Sites with Varying levels of Linkage Disequilibrium

I first validate and showcase how selecting different sampling modes can lead to different LD profiles among the causal variants. In this example, I used the `sample_causal_sites` function to sample causal sites with uniform probability along the genome, or enriched in regions of low or high LD via the LDAK weights distributed with the package in combination with the `weights_power` parameter. **Figure 22** shows the distribution of LD scores when the sampling is done with uniform probability (yellow box). The distribution of LD scores among causal sites sampled in this mode reflects the genome-wide LD score distribution. In contrast, both the high- and low-LD modes lead to shifts in the LD weights distribution, thus confirming that the sampler successfully samples causal sites with an LD score-dependent probability

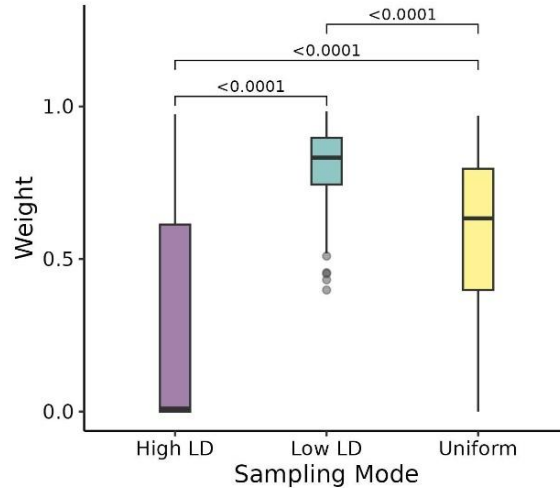


Figure 22. Example 1: Comparison of the LD-dependent and the uniform sampling of causal sites. The boxplots display the distribution of the LD score among causal sites sampled under three different modes in Example 1, as described in the main text. The numbers above the square brackets are FDR-adjusted p-values of two-sided Wilcoxon rank sum tests comparing the distribution of the mean LD score within each modality across 30 replicates.

4.3.2. Example 2: Confounder Correlated with the Polygenic Score

In this example I sought to showcase the simulation of a phenotype influenced by a non-genetic confounder correlated with the polygenic score, as well as to demonstrate that the distribution of the simulated parameter values across 30 replicates matches the target parameter values. I first validate that the distribution of the simulated effect sizes of the causal variants matches the distribution of their expected effect sizes given their minor allele frequencies given by the equation $E[\beta_j^2] = [2p_j(1 - p_j)]^\alpha$. As shown in **Figure 23a**, the identity line (observed values = expected value) matches almost perfectly the regression line of the observed values on the expected values, thus confirming that the `simulate_phenotype` function generates the requested distribution of causal effect sizes. In **Figure 23b**, I demonstrate that the distribution of the heritability, confounder-polygenic score correlation and relative variance of the confounder effects matches the requested parameter values. Thus, this example demonstrates that the function

`simulate_phenotype` generates phenotypes with the target genetic and non-genetic causal architecture.

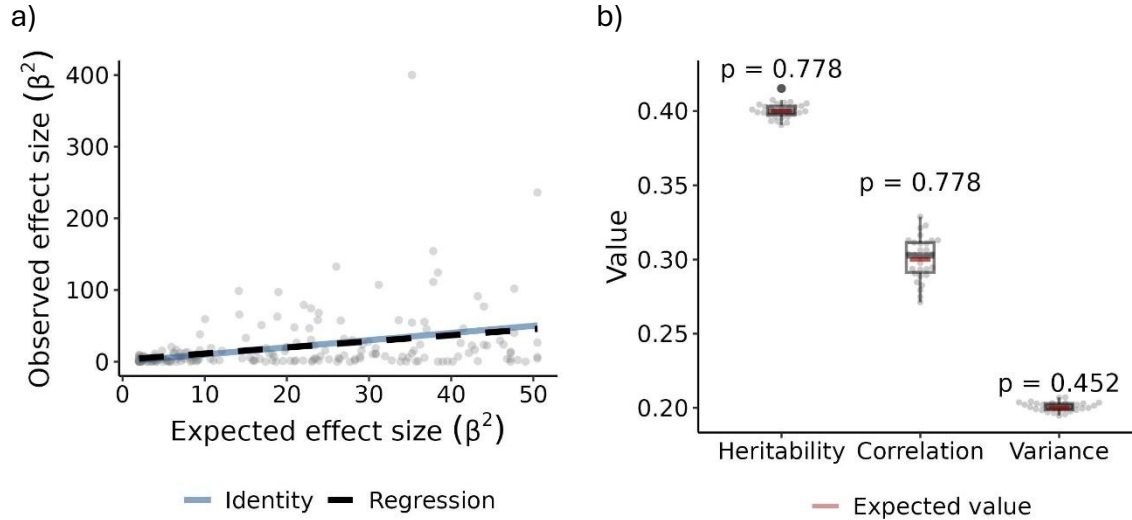


Figure 23. Example 2: Simulation of a confounder effect correlated with the polygenic score. **a)** Scatter plot of the observed causal effect sizes against the expected effect sizes given the minor allele frequencies. The identity line (intercept =0, slope=1) is shown in solid blue, and the regression line in dashed black. **b)** Box and beeswarm plot of the simulated values of for heritability, confounder-polygenic score correlation, and relative variance of the confounder across 30 replicates. The numbers above the boxplot are FDR-adjusted p-values of two-sided t-tests comparing the mean values (shown in red) against the corresponding target parameter values.

4.3.3. Example 3: Binary Trait with a Categorical Confounder in Ordinal Scale

In this example I showcase how phenocause can be used to simulate the relationship between different causal components of a binary polygenic trait. Specifically, I simulated a binary trait with global prevalence of 0.15. The trait is partially determined by an environmental confounder differentially distributed across populations. The confounder increases monotonically, but non-linearly, with the mean value of the first principal component. The confounder was simulated to account for half as much variance in the liability scale as the polygenic score. The LD enrichment functionality was previously validated in example 1 and is thus not repeated in this example.

Figure **Figure 24b** shows that the prevalence of the binary trait is associated with the population labels, and that this association mirrors the association of PC1 with the population labels **Figure 24a**. In contrast, the polygenic score had similar mean values across all populations (p -values > 0.05 , see **Figure 24c**). Thus, this example showcases how phenocause can be used to finely tune the genetic and non-genetic causal components of a polygenic trait.

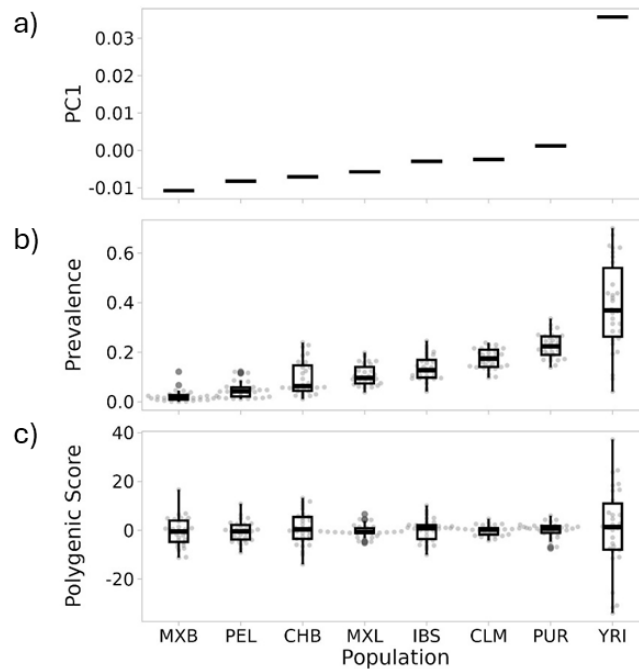


Figure 24. Example 3: Simulation of a binary trait with a categorical confounder in ordinal Scale. **a)** Mean value of the first principal component across populations. The PC1 was used to determine the order in which the confounder effect would increase monotonically. Note that because the PCs are calculated from LD-pruned genome-wide data, they are kept fixed across replicates. **b)** Box and beeswarm plots of population-specific mean prevalence across 30 replicates, demonstrating that it increases monotonically with the mean value of PC1. **c)** Box and beeswarm plots of the mean polygenic score across 30 replicates, showing that it had similar mean values across all populations (all pairwise two-sided t-tests FDR-adjusted p -values were above 0.05).

4.3.4. Example 4: A Factorial Experiment to Isolate Collider Bias

The phenocause package was used to run a 2x2x2 factorial experiment designed to attempt to isolate PC-induced collider bias from the baseline inflation by population structure (26). The

experiment used `phenocause::simulate_phenotype` in `collider` sampling mode to generate phenotypes determined by either PC-correlated (“structured”) or uncorrelated (“non-structured”) causal variants. These phenotypes were then analyzed using GWAS models adjusted with PCs calculated from LD-pruned or non-LD-pruned data. The inflation was measured on the non-causal variant sets, which were also either structured or non-structured.

The structured non-causal variants had a higher inflation rate (**Figure 25a**) and type-I error rate (**Figure 25b**) than the non-structured non-causal variants regardless of the set of data used to calculate the PCs used to define the causal sites (X axis in **Figure 25**) or to adjust the GWAS models (blue vs beige in **Figure 25**). In the analogous control experiment where the causal variants are non-structured, both the median genomic inflation factor (**Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components. a**) and the median type-I error rate (**Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components. b**) of the structured non-causal variants set fall to values lower than those of the matched non-structured non-causal variants. Thus, the simulation via `phenocause::simulate_phenotype` successfully created the statistical inflation expected by genetic confounding.

Notably, **Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components. a** and **Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components. b** reveals a baseline inflation due to population structure even though the causal variants were specifically sampled not to be correlated with the first ten PCs. This baseline was maintained on the left panels of **Figure 25a** and **Figure 25b**, in which the non-causal variants are non-structured. However, the right panels of these figures have an increased inflation, suggesting that there are two different sources of inflation in this experiment. This second source of inflation is compatible with genetic confounding due to specific ancestry components, namely, the first ten PCs. This source of additional confounding can be represented via the following Directed Acyclic Graph (DAG):

Phenotype ← Causal variant ← Ancestry component → Marker variant

In this DAG, the path from the non-causal marker variant to the phenotype exists only if both the marker and the causal site are correlated with the same specific ancestry components beyond the background levels due to population structure.

Next, I analyzed whether there is evidence of statistical inflation beyond that of genetic confounding due to collider bias. Collider bias would have manifested as an increased inflation and type I error rate of structured non-causal sites when both sets of PCs (the ones used to select the causal sites and to adjust the GWAS models) were calculated from non-LD-pruned data. When the causal sites were structured, there was no difference in either the genomic inflation (**Figure 25a**) or the type-I error rate (**Figure 25b**) after stratifying by whether the non-causal variants were non-structured or structured (all-vs-all comparisons within each facet of figures **Figure 25a** and **Figure 25b**). Thus, this experiment revealed no evidence of additional inflation caused by collider bias induced by using PCs calculated from non-LD-pruned data.

Thus, this example showcases how phenocause can be used to simulate complex scenarios and investigate the performance of different inference models under complex causal architectures. This experiment revealed at least two different sources of inflation that would commonly be attributed to population structure without further investigation of the causal mechanism. However, In this test, where the genetic data was simulated under a neutral model, no evidence of collider bias when the PCs are calculated from LD pruned data was found.

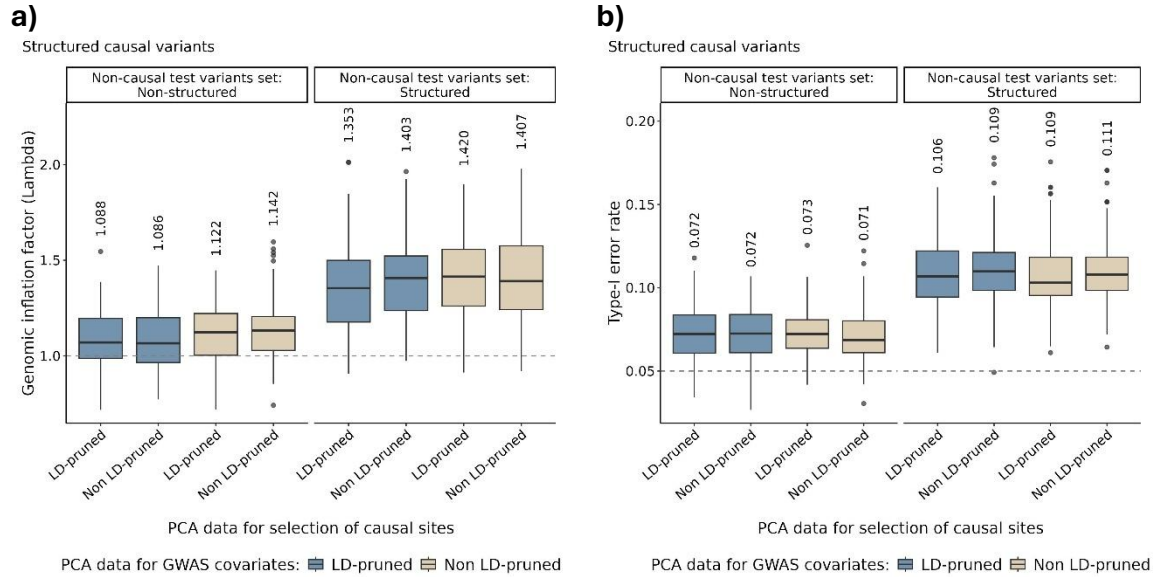


Figure 25. Test statistic inflation when the causal variants are correlated with principal components. The boxplots show the distribution of the **a)** genomic inflation factor (λ_{gc}) and **b)** the type I error rate across 100 replicates. There were eight experimental conditions defined by whether the tested non-causal variants were structured or non-structured (left vs right facet within each figure), whether the causal variants were selected using PCs calculated from LD-pruned or non-LD-pruned data (horizontal axis), or whether the PCs included as GWAS covariates were calculated from LD-pruned or non-LD-pruned data (color-coded). The number above each boxplot is the median value across all 100 replicates. A two-sided Wilcoxon Rank Sum Test was used to compare all-vs-all groups within each facet, but all FDR-adjusted p-values were greater than 0.05 and are therefore not shown.

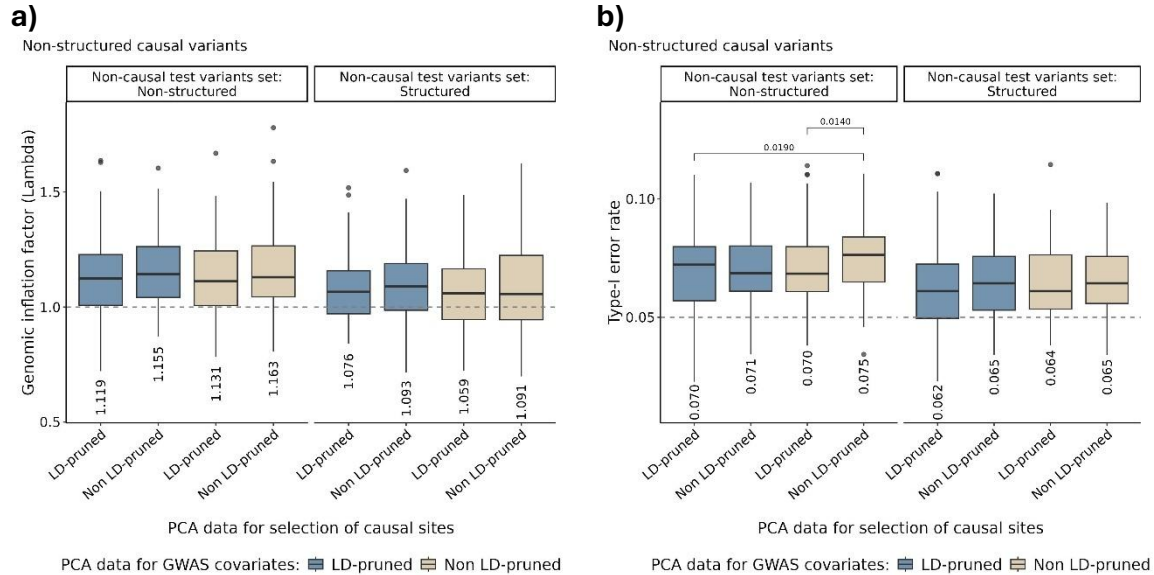


Figure 26. Test statistic inflation when the causal variants are uncorrelated with principal components. The design is the same as in **Figure 25**, except that the causal variants were not correlated with the top ten principal components. The number below each boxplot is the median value across all 100 replicates. A two-sided Wilcoxon Rank Sum Test was used to compare all-vs-all groups within each facet, only FDR-adjusted p-values greater than 0.05 are shown.

4.4. Discussion

The phenocause package was developed to address a methodological gap by providing a framework to simulate phenotypes under complex causal architectures. Previous simulation tools were insufficient for systematically testing the causal assumptions of modern GWAS models. Coalescent-based simulators like tstrait can generate phenotypes from a known genealogy but cannot implement the non-genetic confounding necessary to test for violations of an inference model's independence assumption (28). Conversely, tools such as PhenotypeSimulator create analytical circularity by simulating a genetic effect directly from a user-provided GRM, a design which prevents a rigorous test of that GRM's own underlying assumptions (80). Phenocause was built to overcome these specific limitations.

The design of phenocause avoids analytical circularity by simulating a phenotype directly onto a user-provided genotype matrix, decoupling the true causal model from the assumptions of the inference method. This is achieved through a two-step process: first, the `sample_causal_sites` function defines most aspects of the genetic architecture by sampling causal variants from the input data; second, the `simulate_phenotype` function generates a phenotype based on those variants. Because the phenotype is constructed from scratch based on a specified causal model, the resulting ground-truth data is independent of the inference GWAS model that will be tested on it. This design ensures that any evaluation of an inference model's performance is valid and non-tautological.

The factorial experiment in Example 4 shows that simulating structured causal variants increases test statistic inflation above the baseline caused by population structure alone. The experiment first established a baseline inflation due to population structure, which exists even when causal variants are selected to be uncorrelated with the top ten PCs (**Figure 26**. Test statistic inflation when the causal variants are uncorrelated with principal components.). This baseline is consistent with the continuous model of genetic covariance developed in Aim 1: because a GRM is informationally equivalent to the full set of PCs (5), ensuring zero correlation with the top ten PCs does not remove the remaining covariance encoded by the thousands of remaining PCs that represent the rest of the genealogy. On top of this baseline, simulating a phenotype with causal variants correlated with the top PCs created a significant increase in inflation at structured non-causal markers (**Figure 25**). This additional inflation is the empirical signal of genetic confounding, defined in Aim 1 as the violation of the linear-with-time accumulation of causal effects.

Adjusting the GWAS models with PCs calculated from non-LD-pruned data did not create additional inflation, providing no evidence of collider bias. The experiment was designed to test if using non-LD-pruned PCs would create a third layer of inflation, as predicted by the collider bias

mechanism previously described (26). However, the results showed no differences in inflation whether LD-pruned or non-LD-pruned PCs were used for adjustment. This result can be attributable to the input genetic data. The collider bias phenomenon described by Grinde et al. arises from PCs that capture local genomic features, likely due to non-neutral evolution. Because this experiment used neutrally simulated genetic data, the necessary conditions to create a collider PC were not met, which explains the absence of the artifact and demonstrates its context-dependency. Because of the modularity of phenocause, this explanation could be tested by using other sources of genetic data, possibly from real genomes or simulated under selection. Thus, in spite of the negative results, Example 4 showcases phenocause’s versatility to simulate and investigate more complex causal mechanisms.

This simulation framework provides the necessary tool for a hypothesis-driven investigation of the specific causal architectures under which different GWAS inference models fail or work. By enabling the explicit and separate simulation of environmental confounding, genetic confounding, and the conditions that may lead to statistical artifacts like collider bias, phenocause enables moving beyond correcting for population structure. It allows researchers to systematically investigate the conditions under which mixed linear model specifications can fail, and to develop more robust methods grounded in a more rigorous, mechanistic understanding of confounding (17,18).

4.5. Strengths and implications

The primary contribution of this aim is the development of phenocause, a simulation framework that addresses a critical gap identified in the literature. Its strength lies in two key design features that are absent in other tools. First, its approach of sampling explicit causal variants decouples the

simulated ground truth from the statistical assumptions of the inference models being benchmarked, ensuring that the genetic covariance is an emergent property of a clearly detailed genetic architecture, not an assumption made by the inference model. Second, phenocause implements explicit, user-controlled confounding modules to model environmental factors with a defined correlation to genetic ancestry or the polygenic score. This is a novel feature not present in other simulators.

The main implication of this framework is that it enables rigorous, hypothesis-driven testing of GWAS methods under complex causal models. The factorial experiment in this chapter, which successfully isolated the signal of genetic confounding, was made possible through the tool's unique sampling modes and its ability to simulate phenotypes based on structured causal variants. Thus, phenocause provides a practical framework to move beyond statistical correction and toward the investigation of causal genetic and non-genetic architectures.

4.6. Limitations and future directions

While phenocause provides a flexible framework for studying confounding, its current implementation has two primary limitations regarding the scope of its causal models. First, the genetic component is restricted to an additive model; non-additive effects such as dominance and epistasis are not simulated (3). Second, the framework does not model indirect genetic effects (for example, dynastic or parental effects), which are another recognized source of confounding in both population and family-based studies (18).

The application of phenocause within this dissertation is also necessarily limited in scope. The factorial experiment presented serves as a proof-of-concept to demonstrate the tool's utility for

disentangling statistical inflation mechanisms. However, it does not represent a systematic exploration of all the mechanisms that could potentially violate either the GRM-only or GRM+PCA model. A comprehensive investigation of how phenomena like collider bias and different modes of environmental confounding interact is an area of necessary future work and that phenocause now facilitates.

4.7. Conclusions

This aim successfully developed phenocause, a flexible simulation framework designed to disentangle the causal mechanisms underlying test statistic inflation in GWAS. By implementing a bottom-up simulation of genetic effects and explicit modules for modeling confounding, phenocause addresses the critical limitations of prior software, such as analytical circularity and an inability to simulate non-genetic confounders correlated with ancestry or polygenic score. The usefulness of this framework was demonstrated through a factorial experiment that successfully isolated the statistical signal of genetic confounding from that of baseline inflation due to population structure, providing empirical validation of the model of continuous genetic covariance developed in Aim 1. Finally, phenocause provides the field with a necessary tool to move beyond the simple paradigm of statistical correction. It enables a more rigorous, hypothesis-driven investigation of the causal architecture of complex traits and the conditions under which standard GWAS models succeed or fail.

5. Overall Conclusions of this Study

5.1. Main Findings in this Study

This work first introduced the Coefficient of Genealogical Similarity (GeSi), a measure of relatedness derived from coalescent theory. Under a unified null genetic model, GeSi represents the correlation of total additive genetic effects and can be accurately estimated directly from genotype data without inferring the genealogy. This theoretical work led to a new classification of genetic relationship matrices (GRMs) into “full” GRMs, which capture the entire continuum of shared ancestry, and “shallow” GRMs, which capture only recent relatedness.

Subsequent empirical tests demonstrated that full GRMs, unlike shallow GRMs, do not require principal components (PCs) to model the genetic covariance arising from population structure. In simulations without confounding from non-genetic factors, full GRMs provided more accurate heritability estimates and superior prediction of total genetic value. The analysis of human height, where PC adjustment did improve model fit, is interpreted under this framework as evidence for the presence of unmeasured non-genetic factors that are correlated with ancestry.

Finally, this work developed phenocause, a novel simulation framework designed to disentangle complex causal mechanisms. Using phenocause, a factorial experiment isolated two different sources of statistical inflation commonly interpreted as a single component of population structure. The experiment confirms the hypothesis presented in Aim 1 that the population structure and confounding lead to statistical inflation through different mechanisms.

5.2. Strengths and Implications of this Study

The primary strength of this work is its formal critique of the standard analytical framework. It makes explicit the implicit causal assumptions of the partitioned GRM+PC model, which have historically guided study design and analysis without formal justification. The main implication for all researchers is a more rigorous interpretation of their results. A standard GWAS finding remains statistically valid, but the role of PCs is reframed: their significance in a model is no longer a simple “correction for structure” but is evidence that falsifies a unified null model, pointing to distortions such as genetic or non-genetic confounding that could require further investigation.

For statistical geneticists, the implications are more direct. This dissertation provides a tool to formally investigate these complexities. The GeSi statistic operationalizes the unified null model, providing a theoretical baseline. The phenocause software provides the first ready-to-use tool that implements a mechanistically detailed causal framework to simulate and test the conditions under which this baseline model is insufficient. This work therefore equips the methods-development community to systematically define the boundaries of analytical models, a task previously hindered by a lack of conceptual and practical tools.

5.3. Limitations

A key limitation of this work is that statistical evidence for a model violation does not formally resolve the underlying causal mechanism. The analysis of human height, for example, found that including PCs improved model fit, providing evidence against the unified null model of genetic covariance. While this framework interprets such a result as evidence of either genetic or environmental confounding, leading to the violation of GeSi’s assumptions, it cannot formally rule

out alternative explanations. The same statistical signal could arise from a misspecification of the GRM itself, where PCs capture residual genetic effects not properly modeled by the GRM. This issue of statistical identifiability is a fundamental challenge.

Second, the finding that GRMs built from whole-genome sequence (WGS) data provided the most accurate heritability estimates in simulated data may be an artifact of the simulation design. The simulations sampled causal variants uniformly, an assumption which may not reflect the architecture of real traits where causal variants are non-uniformly distributed (31). Consequently, the performance of WGS-based GRMs in real-world scenarios may require LD- or functional annotation-aware weighting to avoid bias (67).

Third, this work focuses exclusively on quantitative traits and does not formally model the effects of ascertainment bias, a major source of confounding in case-control studies. The non-random sampling of cases and controls can create a spurious correlation between ancestry and disease status that is independent of the underlying genetic architecture (8,12,13). While this represents a form of confounding by a non-genetic factor (the study design itself), it was not explicitly simulated, and what is the best strategy to deal with it remains an open question. Of particular relevance is the question of how to choose which PCs to include in a case-control GWAS considering that the goal should not be to remove all variation caused by population structure, but only the differential ancestry distribution created by the sample selection process.

Finally, the phenocause software is limited to an additive genetic model. The current implementation does not simulate non-additive effects, such as dominance or epistasis, or indirect genetic effects from relatives, which are known to influence complex traits (18).

5.4. Future Directions

The findings of this work motivate two main lines of future research. First, the theoretical framework can be extended by developing formal statistical tests to distinguish between violations of the unified model's assumptions in real data, i.e. to distinguish genetic from non-genetic confounding. The genealogical basis of GeSi could also be leveraged to partition heritability by the age of causal variants, providing deeper insight into a trait's evolutionary history. Second, the phenocause software enables the systematic investigation of the standard GRM+PCA limitations. Future studies should use this tool to investigate the interaction between collider bias and environmental confounding, and to test whether GRMs, in addition to PCs, can induce such artifacts. The tool also facilitates the benchmarking of full and shallow GRMs under more realistic, non-uniform genetic architectures and can be extended to include non-additive and indirect genetic effects.

Finally, the phenocause package, with its implementation of the liability threshold model for binary traits, is suited to formally investigate the impact of ascertainment bias. Future studies could use phenocause to simulate a binary trait and then sample cases and controls with different ancestry proportions to create known levels of ascertainment bias. This would enable a direct test of whether a full GRM is sufficient to account for the resulting inflation or if PC adjustment remains necessary. Such an experiment would be a critical step in extending the unified causal framework to case-control studies.

5.5. Summary

This dissertation introduced a unified causal framework for the study of genetic relatedness and population structure. It began by deriving the Coefficient of Genealogical Similarity (GeSi) from coalescent theory, providing a measure of relatedness that captures the full continuum of shared ancestry. This theoretical work led to a new classification of genetic relationship matrices into “full” and “shallow” GRMs. Empirical tests demonstrated that full GRMs, unlike shallow GRMs, are sufficient to model the genetic covariance from population structure without requiring principal components (PCs). This work reframes the role of PCs in mixed models. Their utility is not to correct for genetic structure itself, but to serve as proxies for unmeasured non-genetic factors correlated with ancestry, or as a necessary component for shallow GRMs that are not genealogically complete. The role of PCs in traits affected by genetic confounding remains unclear, and requires further investigation, since such conditions can lead to collider bias. Finally, this research developed phenocause, a novel tool designed to simulate traits with complex genetic and non-genetic causal architectures. An experiment using this tool confirmed that population structure can cause statistical inflation through at least two different mechanisms even in absence of environmental confounders. Together, these contributions provide a more rigorous conceptual and practical foundation for the analysis of complex traits.

6. References

1. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sinauer; 1998. 980 p.
2. Henderson CR. *Applications of Linear Models in Animal Breeding*. University of Guelph; 1984. 462 p.
3. VanRaden PM. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci*. 2008 Nov 1;91(11):4414–23.
4. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet*. 2016 Jan 7;98(1):127–48.
5. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat Sci*. 2009 Nov;24(4):451–71.
6. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010 Jul;11(7):459–63.
7. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*. 2006 Feb;38(2):203–8.
8. Neale B, Ferreira M, Medland S, Posthuma D. *Statistical Genetics: Gene Mapping Through Linkage and Association*. Garland Science; 2007. 607 p.
9. Falconer DS. *Introduction to Quantitative Genetics*. 4th edition. Harlow: Longman; 1995. 480 p.
10. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet*. 2011 Jan 7;88(1):76–82.
11. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. 2019 Dec 15;35(24):5346–8.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006 Aug;38(8):904–9.
13. Devlin B, Roeder K. Genomic Control for Association Studies. *Biometrics*. 1999 Dec 1;55(4):997–1004.
14. Hernan MA, Robins JM. *Causal Inference: What If*. CRC Press; 2024. 312 p.

15. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*. 2008 Mar 1;178(3):1709–23.
16. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*. 2014 Feb;46(2):100–6.
17. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. *Nat Rev Genet*. 2013 Jan;14(1):1–2.
18. Veller C, Coop GM. Interpreting population- and family-based genome-wide association studies in the presence of confounding. *PLOS Biol*. 2024 Apr 11;22(4):e3002511.
19. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015 Mar;47(3):291–5.
20. Hu S, Ferreira LAF, Shi S, Hellenthal G, Marchini J, Lawson DJ, et al. Fine-scale population structure and widespread conservation of genetic effect sizes between human groups across traits. *Nat Genet*. 2025 Feb;57(2):379–89.
21. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun*. 2020 Jul 31;11(1):3865.
22. Serre D, Pääbo S. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Res*. 2004 Sep 1;14(9):1679–85.
23. Liu H, Prugnolle F, Manica A, Balloux F. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *Am J Hum Genet*. 2006 Aug 1;79(2):230–7.
24. Meirmans PG. Seven common mistakes in population genetics and how to avoid them. *Mol Ecol*. 2015;24(13):3223–31.
25. Gouveia MH, Meeks KAC, Borda V, Leal TP, Kehdy FSG, Mogire R, et al. Subcontinental genetic variation in the All of Us Research Program: Implications for biomedical research. *Am J Hum Genet*. 2025 Jun 5;112(6):1286–301.
26. Grinde KE, Browning BL, Reiner AP, Thornton TA, Browning SR. Adjusting for principal components can induce collider bias in genome-wide association studies. *PLoS Genet*. 2024 Dec;20(12):e1011242.
27. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022 Mar 1;220(3):iyab229.
28. Tagami D, Bisschop G, Kelleher J. tstrait: a quantitative trait simulator for ancestral recombination graphs. *Bioinformatics*. 2024 Jun 3;40(6):btae334.

29. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Mol Biol Evol.* 2019 Mar 1;36(3):632–7.
30. Yin D, Zhang X, Yin L, Zhang H, Tang Z, Xu J, et al. SIMER: an accurate and intelligent tool for simulating customizable population data across species in complex scenarios. *J Big Data.* 2025 Dec;12(1):1–23.
31. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. *Am J Hum Genet.* 2012 Dec 7;91(6):1011–21.
32. Kingman JFC. The coalescent. *Stoch Process Their Appl.* 1982 Sep 1;13(3):235–48.
33. Wakely J. *Coalescent Theory: An Introduction.* Macmillan Learning; 2016. 352 p.
34. Griffiths RC, Marjoram P. An ancestral recombination graph. *Inst Math Its Appl.* 1997;87:257.
35. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. *Nat Genet.* 2019 Sep;51(9):1330–8.
36. Kelleher J, Thornton KR, Ashander J, Ralph PL. Efficient pedigree recording for fast population genetics simulation. *PLOS Comput Biol.* 2018 Nov 1;14(11):e1006581.
37. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet.* 2019 Sep;51(9):1321–9.
38. Zhang BC, Biddanda A, Gunnarsson ÁF, Cooper F, Palamara PF. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat Genet.* 2023 May;55(5):768–76.
39. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genet.* 2014 May 15;10(5):e1004342.
40. Palamara PF, Terhorst J, Song YS, Price AL. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat Genet.* 2018 Sep;50(9):1311–7.
41. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet.* 2009 Jun;10(6):381–91.
42. Campos G de los, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLOS Genet.* 2013 Jul 11;9(7):e1003608.
43. Ober C, Abney M, McPeck MS. The Genetic Dissection of Complex Traits in a Founder Population. *Am J Hum Genet.* 2001 Nov 1;69(5):1068–79.

44. Medina-Muñoz SG, Vecchyo DOD, Cruz-Hervert LP, Ferreyra-Reyes L, García-García L, Moreno-Estrada A, et al. Demographic modeling of admixed Latin American populations from whole genomes. *Am J Hum Genet.* 2023 Oct 5;110(10):1804–16.
45. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021 Feb;590(7845):290–9.
46. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018 Sep 1;34(17):i884–90.
47. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; 2013 [cited 2025 Jun 25]. Available from: <http://arxiv.org/abs/1303.3997>
48. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021 Feb 1;10(2):giab008.
49. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature.* 2022 Jul;607(7920):732–40.
50. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell.* 2022 Sep 1;185(18):3426-3440.e19.
51. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 2006 Sep 1;16(9):1182–90.
52. GATK [Internet]. 2025 [cited 2025 Jun 25]. GATK Resource bundle. Available from: <https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle>
53. Picard toolkit [Internet]. Broad Institute; 2025 [cited 2025 Jun 25]. Available from: <https://github.com/broadinstitute/picard>
54. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput Biol.* 2016 May 4;12(5):e1004842.
55. Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genet.* 2020 May 5;16(5):e1008619.
56. Jukes TH, Cantor CR. Evolution of Protein Molecules. In: *Mammalian protein metabolism.* New York: Academic Press; 1969. p. 21–132.
57. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015 Dec;4(1):7.

58. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010 Nov 15;26(22):2867–73.
59. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012 Dec 1;28(24):3326–8.
60. Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, et al. Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics*. 2017 Sep 1;207(1):75–82.
61. McVean G. A Genealogical Interpretation of Principal Components Analysis. *PLOS Genet*. 2009 Oct 16;5(10):e1000686.
62. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012 Mar;44(3):243–6.
63. Sofer T, Zheng X, Gogarten SM, Laurie CA, Grinde K, Shaffer JR, et al. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol*. 2019;43(3):263–75.
64. Henderson CR. Selection index and expected genetic advance. *Statistical Genet Plant Breed* [Internet]. 1963 [cited 2024 Sep 30]; Available from: <https://cir.nii.ac.jp/crid/1573105975550488448>
65. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using summary statistics. *Nat Genet*. 2020 Apr;52(4):458–62.
66. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLOS Genet*. 2006 Dec 22;2(12):e190.
67. Bryc K, Bryc W, Silverstein JW. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theor Popul Biol*. 2013 Nov 1;89:34–43.
68. François O, Gain C. A spectral theory for Wright’s inbreeding coefficients and related quantities. *PLOS Genet*. 2021 Jul 19;17(7):e1009665.
69. Benning JW, Carlson J, Smith OS, Shaw RG, Harpak A. Confounding Fuels Misinterpretation in Human Genetics [Internet]. *bioRxiv*; 2024 [cited 2025 Jun 12]. p. 2023.11.01.565061. Available from: <https://www.biorxiv.org/content/10.1101/2023.11.01.565061v3>
70. Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. Perry GH, Turchin MC, Martin AR, editors. *eLife*. 2020 Nov 17;9:e61548.
71. Marsico F, Buonaiuto S, Amos–Abanyie EK, Chinthala LK, Mohammed A, Center RG, et al. Identity-by-descent captures Shared Environmental Factors at Biobank Scale [Internet]. *bioRxiv*; 2025 [cited 2025 Jun 12]. p. 2025.05.03.652048. Available from: <https://www.biorxiv.org/content/10.1101/2025.05.03.652048v1>

72. Fan C, Mancuso N, Chiang CWK. A genealogical estimate of genetic relationships. *Am J Hum Genet.* 2022 May 5;109(5):812–24.
73. Ralph P, Thornton K, Kelleher J. Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics.* 2020 Jul 1;215(3):779–97.
74. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet.* 2015 Nov;47(11):1228–35.
75. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat Genet.* 2017 Oct;49(10):1421–7.
76. Porter HF, O’Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci Rep.* 2017 Mar 13;7(1):38837.
77. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics.* 2011 Aug 15;27(16):2304–5.
78. Meyer HV, Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics.* 2018 Sep 1;34(17):2951–6.
79. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, et al. data.table: Extension of `data.frame` [Internet]. 2025. Available from: <https://r-datatable.com>
80. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation [Internet]. 2025. Available from: <https://dplyr.tidyverse.org>
81. Zheng X, Gogarten SM, Lawrence M, Stilp A, Conomos MP, Weir BS, et al. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics.* 2017 Aug 1;33(15):2251–7.
82. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009 Sep 1;19(9):1655–64.