Topics in the Quantitative Analysis of Complex Trait Genetic Architectures

by

Richard Border

B.A., Wesleyan University, 2010

M.A., University of Colorado Boulder, 2018

M.S., University of Colorado Boulder, 2018

A dissertation submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Psychology and Neuroscience

2019

This thesis entitled:
Topics in the Quantitative Analysis of Complex Trait Genetic Architectures
written by Richard Border
has been approved for the Department of Psychology and Neuroscience

Mathew C. Keller

Stephen Becker

Luke M. Evans

Naomi P. Friedman

Soo Hyun Rhee

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Border, Richard Schuster (Ph.D., Psychology and Neuroscience)

Topics in the Quantitative Analysis of Complex Trait Genetic Architectures

Dissertation directed by Associate Professor Matthew C. Keller

The unprecedented availability of whole-genome data in recent years has allowed researchers to ask new questions regarding the genetic underpinnings of human traits while simultaneously introducing numerous theoretical and computational challenges. The vast majority of commonly studied phenotypes have revealed themselves to be *complex traits*—outcomes influenced by numerous genetic and non-genetic factors, the former of which remain largely unknown. Large samples of high-dimensional data are required to effectively study the genetic architectures of such traits and the development of new theory and methods for effectively analyzing such data have become essential. The present dissertation investigates several topics in this emerging area in the context of three studies. In the first study, *Falsification of candidate gene hypotheses for major depression*, we apply genome-wide methods and probabilistic arguments to the critical interrogation of previous "candidate gene" hypotheses regarding the genetics of major depression. We demonstrate that modern, well-powered samples provide little evidence supporting these pre-whole genome era hypotheses and that these discrepancies cannot be explained by measurement error or non-additive effects, concluding that the majority of published findings in this area represent false positives. In the second study, *Stochastic Lanczos residual maximum likelihood algorithms*, we introduce two novel algorithms for residual maximum likelihood (REML) estimation of genomic variance components in the context of linear mixed-effects models. The principle of Krylov subspace shift-invariance is applied to exploit problem structure and speed computation beyond existing methods, which we demonstrate via theoretical argument and numerical experiments. In the final study, *Assortative mating and whole-genome heritability estimation*, we present on-going work characterizing the effects of assortative mating on commonly used heritability estimation procedures. We provide analytic and simulation-based arguments demonstrating that three widely used estimators demonstrate substantial bias when assumptions regarding independence of causal variants are violated.

in memory of Zooey Schuster

Acknowledgments

I would like to thank my committee members, Stephen Becker, Luke Evans, Naomi Friedman, Matt Keller, and Soo Rhee, for their exceptional mentoring and guidance. Additionally, I would like to thank Matt Jones of the Department of Psychology and Neuroscience and Sean O'Rourke of the Department of Mathematics, for their supervision of and contributions to the theoretical work presented in the third section, though I take full credit for any errors contained therein. Also, I would like to acknowledge Emma Johnson and Andy Smolen for their contributions to the work presented in the first section. Finally, I'd like to thank Alta, Rosemary, Buckminster, Catherine, and Kim for their love and support.

Contents

List of Tables

List of Figures

Forward

Complex traits—phenotypes reflecting numerous heritable and non-heritable influences—have been the subject of genetics research for over a century. In contrast, the availability of genome-wide data—measured genotypes at millions of loci across the genome—is a relatively recent phenomenon, and samples large enough to effectively investigate the genetic architectures of complex traits are largely limited to the previous decade. Despite this short timeline, the efforts of international consortia and research initiatives (e.g., [1, 2, 3]) have resulted in numerous large samples of individuals with measured genotypic and phenotypic data. The large scale of these data sets, comprised of the genotypes of hundreds of thousands of individuals at millions of loci, has provided researchers with unprecedented resources while simultaneously introducing numerous theoretical, numerical, and computational challenges. This dissertation focuses on a subset of these emerging topics in statistical genetics.

Chapter 1 applies whole-genome era methods to the examination of historical hypotheses regarding the genetic underpinnings of major depression[1]. Through a diverse array of analytic methods, we demonstrate that early theories, some of which continue to generate considerable research interest, were incorrect and that the discrepancies between the so-called“candidate gene” and genome-wide literatures cannot be explained by non-additive genetic effects or measurement error. Next, in Chapter 2, we introduce two novel stochastic algorithms for the estimation of genomic variance components using residual maximum likelihood (REML)[2]. By exploiting problem structure through the principle of Krylov subspace shift-invariance, our novel algorithms speed computation beyond existing methods; we demonstrate these improvements theoretically and via

[1] Originally published in the *American Journal of Psychiatry* as No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples with coauthors E.C. Johnson, L.M. Evans, A. Smolen, N. Berley, P.F. Sullivan, and M.C. Keller [4]. Reprinted with permission from the American Journal of Psychiatry, (Copyright 2019). American Psychiatric Association.

[2] Originally published in *BMC Bioinformatics* as Stochastic Lanczos estimation of genomic variance components for linear mixed-effects models with coauthor S. Becker [5]. Copyright retained by the author of this dissertation.

numerical experiment. Finally, in Chapter 3, we present ongoing work characterizing the behavior of commonly used methods for estimating heritability from whole-genome data in populations subject to assortative mating (AM). Employing theory- and simulation-based arguments, we demonstrate that three widely-used procedures generate biased estimates of the equilibrium heritability under Fisher's classical AM framework.

Chapter 1

Falsification of candidate gene hypotheses for major depression

1.1 Introduction

Major Depressive Disorder (hereafter referred to as “depression”) is moderately heritable (twin-based heritability $\approx 37\%$ [6]), but its genetic architecture is complex, and identifying specific polymorphisms underlying depression susceptibility has been challenging. With the ability to genotype particular genetic variants and optimism about the potential public health impact of identifying reliable biomarkers for depression [7], early research focused on the effects of specific candidate polymorphisms in genes hypothesized to underlie depression liability. These genes were chosen based on hypotheses regarding the biological underpinnings of depression. The 5-HTTLPR variable number tandem repeat (VNTR) polymorphism in the promoter region of the serotonin transporter gene *SLC6A4*, the most commonly studied polymorphism in relation to depression (Figure 1.1, Table S1.1[1]), serves as a prototypical example: given the theorized importance of the serotonergic system in the etiology of depression, a logical target for early association studies was a common, large (and hence relatively easy to genotype), and potentially functional repeat polymorphism in a serotonergic gene [8–10]. Early investigations, though focused on a small number of variants by necessity (low cost genome-wide arrays were not yet available), reported promising positive associations. However, replication attempts led to inconsistent results [11–13]).

To critics of candidate gene findings, replication failures suggested that initial reports were artifactual [14–16]. However, at least two alternative explanations could account for inabilities to replicate early reports and inconsistent results across studies. First, in the early 2000s, Caspi et al. posited that previous inconsistencies might reflect the effects of candidate polymorphisms that were dependent on environment exposures (gene-by-environment interaction [G × E] effects) [17]. In what would become one of the most highly-cited ($>$ 8000 citations as of July, 2018) and influential

[1] Tables and Figures prefixed with “S” are available in the online supplement to the published version of this chapter [4]

papers in psychiatric genetics, Caspi et al. reported that the impact of the *5-HTTLPR* repeat polymorphism in *SLC6A4* on depression was moderated by exposure to stressful life events, such that the positive association between stressful life events and depression was stronger in individuals carrying the "short" allele [18]. This early work led many researchers to shift their attention to G × E hypotheses, focusing on the same polymorphisms first investigated for main effects [13]. Second, in an alternative but complementary line of reasoning, other researchers suggested that polymorphisms other than those studied previously in the same candidate genes were likely to explain depression risk, given the genes' putative biological relevance [19]. All three lines of inquiry are well represented in the published literature of the past twenty-five years: thousands of investigations of depression or depression endophenotypes have examined the direct effects of 1., the most studied polymorphisms within candidate genes, 2., the moderation of their effects by environmental stressors, or 3., the effects of alternative polymorphisms within the same candidate genes. The popularity of these lines of inquiry has not diminished over time (Figure 1.1, supplement sections S1.4, S1.5), with many studies reporting statistically significant associations.

Perhaps surprisingly given the continued interest in studying these historic depression candidate genes and the large number of associations documented in the candidate gene literature, many researchers have expressed extreme skepticism about the validity of such findings [16, 20–22]. There are several reasons for this. First, genome-wide association studies (GWAS), which agnostically examine associations at millions of common single nucleotide polymorphisms (SNPs) across the genome in large samples, have consistently found that individual SNPs exert small effects on genetically-complex traits such as depression [2, 3, 23]. For example, in the most recent GWAS of depression, which utilized a sample of 135,458 cases and 344,901 controls, the strongest individual signal detected (rs12552; *odds ratio* = 1.044; p = 6.07e-19) would require a sample of approximately 34,100 individuals to be detected with 80% power at $\alpha = .05$, assuming a balanced case-control design [3]. In contrast, the median study sample size in a review of 103 candidate G × E studies published during 2000-2009 was 345, with 65% of studies reporting positive results [20]. Thus, given the small sample sizes typically employed, candidate gene research has likely been severely underpowered [24, 25]. This, in turn, may suggest that the false discovery rate for the many positive reports in the candidate gene literature is high. Consistent with this possibility, targeted, well-powered genetic association studies of depression and other psychiatric phenotypes in large samples have not supported candidate gene hypotheses [2, 26–29]. For example, a preregistered, collaborative meta-analysis of the stressful life event × 5-HTTLPR interaction in a sample of 38,802 individuals failed to support the original finding of Caspi et al. [30], though we note that this variant and several other candidate VNTRs have not been previously examined in a GWAS context [31, 32]. The absence of previous large-sample investigations of VNTR hypotheses is noteworthy as VNTRs comprise several of the earliest candidate polymorphisms to be examined in the context behavioral research; concerns about variability in VNTR genotyping procedures and analysis methods over time have further complicated the interpretation of the existing literature [33]. Additionally, a number of researchers have suggested that incorrect analytic methods and inadequate control for population stratification characterize the majority of published candidate gene studies [24, 34–36], and other researchers have questioned the clinical utility of focusing on individual polymorphisms or polymorphism-by-environment interactions [37]. Finally, there is evidence of systematic publication bias in the candidate gene literature; in the aforementioned review of all candidate G × E studies published between 2000 and 2009, 96% percent of novel findings were significant compared to only 27% of replication attempts, and replication attempts reporting null findings had larger sample sizes than those presenting positive findings [20]. In response to such skepticism, candidate gene proponents have argued that lack of replication of candidate gene associations in large sample studies may reflect poor or limited phenotyping [38–40], exclusion of non-SNP polymorphisms such as VNTRs [19, 32], the "multiple-testing burden" associated with genome-wide scans [38], and failure to account for environmental moderators [38, 39, 41].

The current study is the most comprehensive and well-powered investigation of historic candidate polymorphism and candidate gene hypotheses in depression to date. We focus on three lines of inquiry concerning how historic candidate genes may impact depression liability:

1. main effects of the most commonly studied candidate polymorphisms;

2. moderation of the effects of these polymorphisms by environmental exposures;

3. main effects of common SNPs across each of the candidate genes.

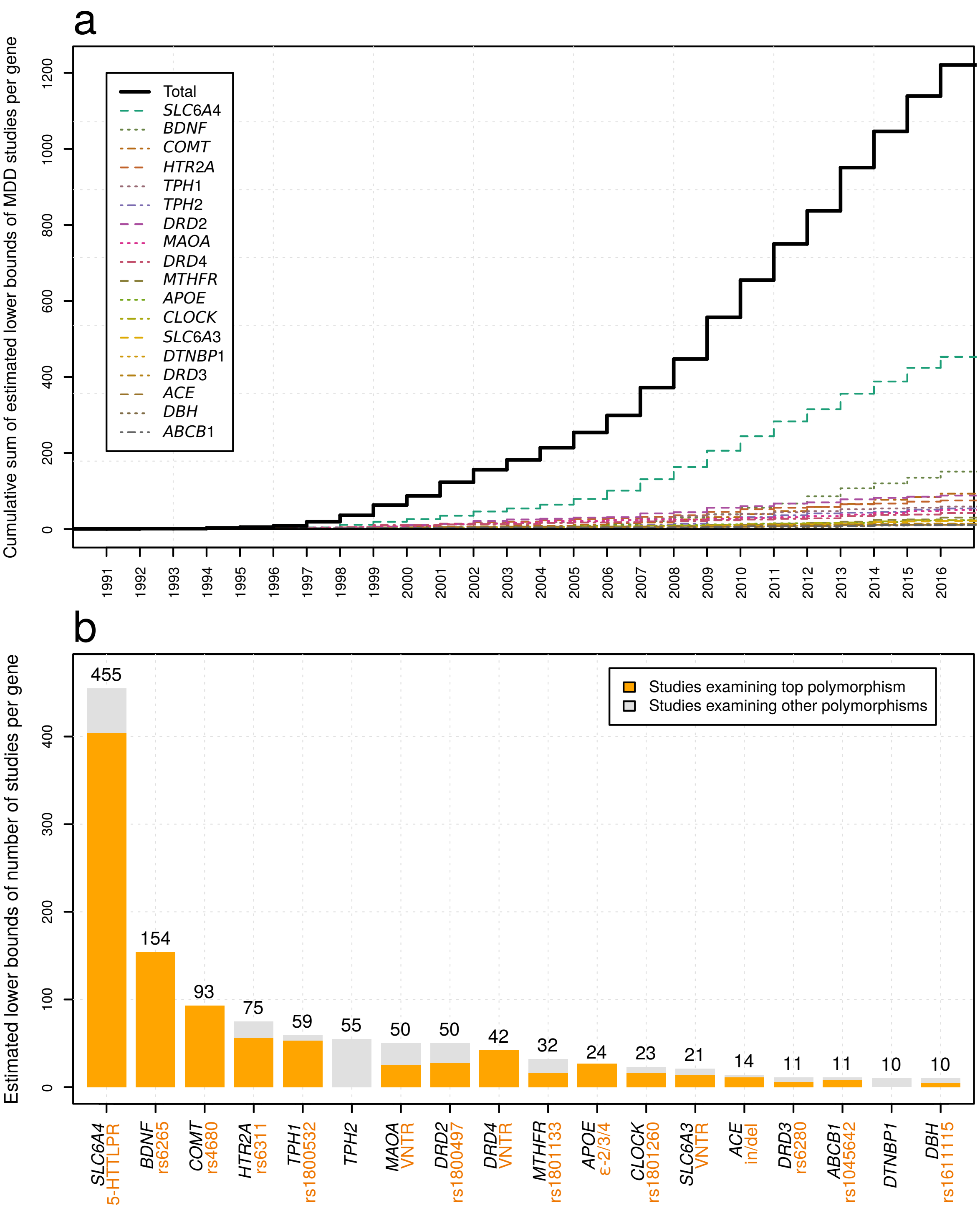
We first empirically identified 18 commonly studied candidate genes represented in at least ten peer-reviewed depression-focused journal articles between 1991 and 2016 from the body of publications indexed in the PubMed database [42]. Within these candidate genes, we identified the most commonly studied polymorphisms, as well as their canonical risk alleles, at which point our primary analysis plan was preregistered. Using multiple large samples (n ranging from 62,138 to 443,264 across subsamples; total N = 621,214 individuals), we examined multiple measures of depression (e.g., lifetime diagnostic status, symptom severity among individuals reporting mood disturbances, lifetime number of depressive episodes; Table 1.1), employing multiple statistical frameworks (e.g., main effects of polymorphisms and genes, interaction effects on both the additive and multiplicative scales) and, in G × E analyses, considering multiple indices of environmental exposure (e.g., traumatic events in childhood or adulthood,). Previous large sample studies of depression have largely focused on genetic main effects on depression diagnosis in the context of SNP data across the genome. In contrast, we examined several alternative depression phenotypes, analyzed both main effects and interactions with multiple potential moderators, included the most studied polymorphisms, including VNTRs (Figure 1), and employed a liberal significance threshold. Further, we quantified the extent to which phenotypic measurement error may have biased our results. The unifying question underlying this "multiverse" analytic approach [43] was the following: do the large datasets of the whole-genome data era support any previous depression candidate gene hypotheses?

1.2 Materials and methods

1.2.1 Identification of genes and polymorphisms

We identified eighteen candidate genes studied for their associations with depression phenotypes at least ten times from within the body of peer-reviewed biomedical literature indexed in the PubMed database [42] using the Biopython bioinformatics package [44]. We used regular expressions to find articles potentially corresponding to each gene and hand-verified the number of correctly classified articles for each gene in order to estimate hypergeometric confidence intervals for the true number of correctly classified studies. We identified single polymorphisms comprising a large proportion of study foci for 16 of the 18 candidate genes. Figure 1.1 shows the most studied candidate genes and polymorphisms within them, as well as probabilistic estimates of the minimum number of times each has been studied with respect to depression and the number of studies-per-gene-per-year (confidence intervals presented in Table S1.1).

Figure 1.1: Popularity of candidate genes over time



a. Cumulative sums of the estimated number of depression candidate gene studies identified by our algorithm per year per gene from 1991 through 2016. Estimates reflect the number of correctly classified studies among identified studies, excluding studies not detected by our protocol, and thus comprise lower bounds for the true number of studies-per-gene. **b.** Eighteen candidate genes studied ≥ 10 times between 1991 and 2016. The estimated number of studies focused on the top polymorphism (Table S1.1) is displayed relative to the other identified studies within each gene. No top polymorphisms were identified for *DTNBP1* or *TPH2* (supplement section S1).

Table 1.1: Depression and environmental moderator phenotypes

Depression phenotypes		
Phenotype	**Description**	**Sample size**
Estimated lifetime depression diagnosis	Binary indicator of lifetime DSM-V depression diagnosis assessed in UKBB online mental health follow-up questionnaire. To meet criteria, participants had to endorse at least four of eight DSM-V depression symptoms (motor agitation/retardation symptom was not assessed), as well as duration, frequency, and impairment criteria.	*n* = 115,458 *85,513*: *controls* *29,945 cases*
Current depression severity	Sum score of all nine DSM-V depression symptom severities (using four point Likert scale to index severity of each symptom) over the two weeks leading up to assessment. Assessed in UKBB online mental health follow-up questionnaire.	*n* = 115,463 $\bar{x}$ = *2.502* s_x = *3.347*
Conditional lifetime symptom count	Sum of symptom indicators for eight of nine lifetime DSM-V depression symptoms (motor agitation/retardation symptom was not assessed) among individuals endorsing lifetime incidence of a two+ week period characterized by anhedonia and/or depressed mood (questionnaire skip patterns necessitated this precondition). Assessed in the UKBB online mental health follow-up questionnaire.	*n* = 62,138 $\bar{x}$ = *4.746* s_x = *1.745*
Lifetime episode count	Ordinal measure of incidence/recurrence of a two+ week period characterized by anhedonia and/or depressed mood indicating zero episodes, a single episode, or recurrent episodes. Assessed in UKBB online mental health follow-up questionnaire.	*n* = 115,457 *55,388: zero* *30,724: single* *26,345: recurrent*
Touchscreen probable lifetime diagnosis, ordinal classification	Ordinal measure of depression diagnostic status based on a selection items of items from the Patient Health Questionnaire, the Structured Clinical Interview for DSM-IV Axis I Disorders-Research Version, and items assessing treatment seeking behavior specific to the UKBB touchscreen interview, as described in Smith et al., 2013. Categories included no depression, single depressive episode, recurrent episodes (moderate), and recurrent episodes (severe), in that order. Assessed as part of the UKBB initial touchscreen interview.	*n* = 91,121 *66,605*: *controls* *6,209*: *1 episode* *11,634*: *≥ 2 moderate* *6,633*: *≥ 2 severe*
Touchscreen probable lifetime diagnosis	Dichotomized coding of the touchscreen probable life diagnosis ordinal classification, contrasting no depression with the three diagnosis categories.	*n* = 91,121 *66,605*: *controls* *84,516*: *cases*
Severe recurrent depression	Binary indicator of case/control status for depression excluding cases and controls with mild to moderate depression symptoms. Controls were individuals who failed to endorse incidence of a two+ week period characterized by anhedonia and/or depressed mood. Cases were individuals met criteria for estimated lifetime depression diagnosis, endorsed at least five of the eight measured DSM-V symptoms, and experienced recurrent depressive episodes. Assessed in UKBB online mental health follow-up questionnaire.	*n* = 64,432 *53,218*: *controls* *14,214*: *cases*
PGC lifetime depression diagnosis	Binary indicator of lifetime depression diagnosis as measured in the PGC2 depression GWAS. The current investigation utilized data from the full expanded cohort meta-analysis, excepting UK-based cohorts (UKBB and Generation Scotland).	*n* = 443,264 *323,063*: *controls* *120,201*: *cases*
Moderator phenotypes		
Phenotype	**Description**	**Sample size**
Childhood trauma	Binary indicator of sexual and or physical abuse during childhood. Assessed in the UKBB online mental health follow-up questionnaire.	*n* = 157,146 *118,800*: *unexposed* *38,346*: *exposed*
Adult trauma	Binary indicator of any of the following traumatic events during adulthood: physical assault, sexual assault, witness to sudden/violent death, diagnosis with life threatening illness, involvement in life threatening accident, and exposure to combat or war-zone conditions. Assessed in the UKBB online mental health follow-up questionnaire.	*n* = 157,223 *64,286*: *unexposed* *92,937*: *exposed*
Recent trauma	Binary indicator of whether any of the above events occurred in the year leading up to assessment.	*n* = 157,220 *142,008*: *unexposed* *15,212*: *exposed*
Stressor-induced depression	Binary indicator of whether period of depressed mood or anhedonia was a possible consequence of a traumatic event among individuals endorsing lifetime incidence of a two+ week period characterized by anhedonia and/or depressed mood (questionnaire skip patterns necessitated this precondition). Assessed in the UKBB online mental health follow-up questionnaire.	*n* = 88,585 *23,746*: *unrelated to stressor* *64,839*: *stressor-induced*
Townsend deprivation index (TDI)	Measure of socioeconomic adversity with higher values indicating greater adversity. Standardized to have zero mean and unit standard deviation. Assessed during the UKBB initial touchscreen interview.	*n* = 187,094

1.2.2 Samples

1.2.2.1 UK Biobank samples

A large portion of the data used in the present study was collected by the UK Biobank (UKBB), a population sample of 502,682 individuals collected at 22 centers across the United Kingdom between 2006 and 2010 [1]. Within this group, we analyzed several depression phenotypes and moderators among 177,950 unrelated (pairwise genome-wide relatedness, $\hat{\pi} < .05$) European ancestry individuals for whom relevant depression measures were collected. We analyzed two partially overlapping subsets of these individuals: 91,121 individuals for whom selected items from the initial touchscreen interview were available and 115,458 individuals who completed a series of online mental health questionnaires, 62,138 of whom endorsed a two-week period characterized by anhedonia or depressed mood at some point during their lives. DNA was extracted from whole blood and genotyped using the Affymetrix UK Biobank Axiom array or the Affymetrix UK BiLEVE Axiom array and imputed to the Haplotype Reference Consortium by the UKBB [45]. Further details on genotyping and sampling procedures are available in S2. Because VNTRs were not genotyped in the UKBB dataset, we used two independent whole-genome SNP datasets (the Family Transition Project [46] and the Genetics of Antisocial Drug Dependence [47, 48]) that also measured these repeat polymorphisms as reference panels in order to impute highly studied VNTRs within *DRD4, MAOA, SLC6A3,* and *SLC6A4* in the UKBB. The estimated out-of-sample imputed genotype match rates were ≥ 0.919 for all four VNTRs (complete details are provided in [49]).

1.2.2.2 Psychiatric Genetics Consortium sample

To investigate candidate gene polymorphism main effect hypotheses, we also used data from the most recent GWAS on depression conducted by the Major Depressive Disorder Working Group of the Psychiatric Genetics Consortium (PGC), which is described in detail in [3]. Lack of access to raw genotypes for a large number of the PGC cohorts precluded imputation of VNTRs in the PGC sample. To minimize sample overlap with UKBB, UK-based cohorts were excluded from the PGC dataset, resulting in GWAS summary statistics for a total of 443,264 individuals (120,201 cases; 323,063 controls); see S2 for further details.

1.2.3 Phenotypes

Table 1.1 describes all phenotypes examined in the present investigation, with additional information provided in S3. Correlations between depression outcomes and Cohen's κ estimates for diagnosis phenotypes are presented in Tables S3.1 and S3.2, respectively. Marker-based heritabilities of, and genetic correlations between, depression outcomes were estimated via LD score regression [50] and are presented in Tables S3.3-S3.4 and Figure S3.3 (see S4.4 for further details).

1.3 Analyses

All analyses were preregistered through the Open Science Framework [51] and are available at (https://osf.io/akgvz/). Statistical models are described in detail in S4 and departures from the preregistered analyses are documented in S5.

1.3.1 Polymorphism-wise analyses

We analyzed associations between outcomes and each of the top 16 candidate polymorphisms using a generalized linear model framework (link functions listed in Table S4.1). For two of the genes, *TPH2* and *DTNBP1*, no particular polymorphism was investigated in a preponderance of studies, and so these genes were not included in the polymorphism-wide analyses. Covariates included genotyping batch, testing center, sex, age, age^2, and the first ten European ancestry principal components. Sixteen polymorphism × environment effects were tested on both the additive and multiplicative scales for each of the 16 polymorphisms; each model tested is listed in Table S4.1. For interaction tests, we included all covariate × polymorphism and covariate × moderator terms to control for the potential confounding influences of covariates on the interaction [35]. We also tested interaction models only controlling for covariate main effects, which is incorrect but common in the candidate gene literature [34]. Across all outcomes we employed a preregistered significance threshold of $\alpha_{\text{poly}} = .05/16 = 3.13\text{e-}03$, corresponding to a Bonferroni correction across the top 16 candidate polymorphisms. This threshold is liberal because it does not account for the multiple ways each polymorphisms was analyzed or the multiple outcomes it was assessed with respect to. Further details are provided in S4.1.

1.3.2 Gene-wise and gene-set analyses

We used the NCBI Build 37 gene locations to annotate SNPs to genes, allowing SNPs within a 25kb window of the gene start and end points to be mapped to each gene. We used MAGMA software version 1.05b [52] to perform gene-wise and gene-set analyses for the top eighteen candidate genes separately in the UKBB and PGC datasets. Gene-wise tests summarize the degree of association between a phenotype and polymorphisms within a given gene; in contrast, gene-set tests examine the association between a phenotype and a set of genes rather than individual genes.

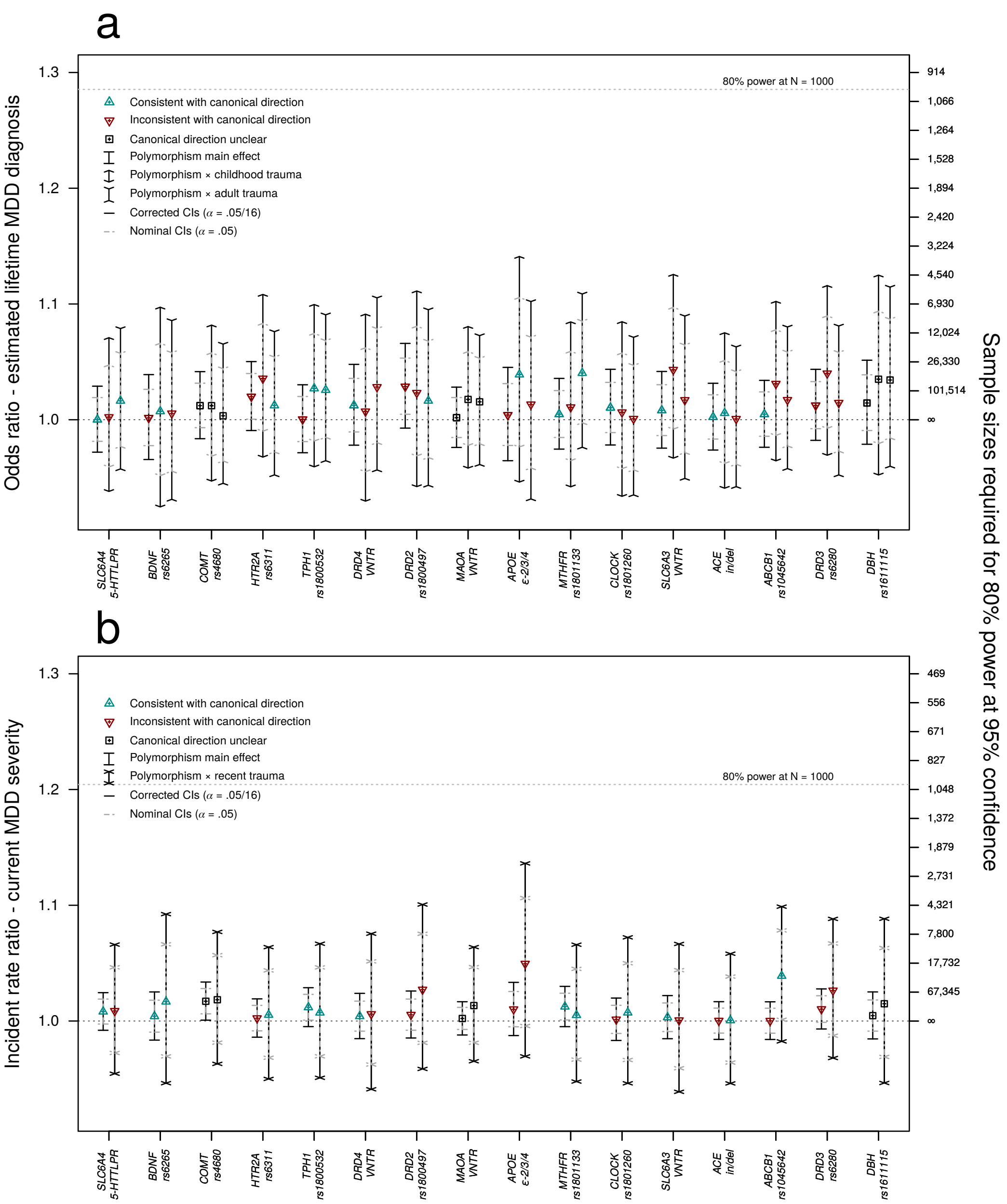
We conducted gene-wise association analyses for each gene and outcome using the MAGMA default gene-level association statistic (sum -log p-based statistics or principal components regression, for tests based on summary statistics or individual-level genotypes, respectively) and using a liberal significance threshold of $\alpha_{\text{gene}} = .05/18 = 2.78\text{e-}03$ to correct for multiple tests across the 18 candidate genes. We used summary statistics from the PGC2 depression GWAS [3] (excluding UK-based cohorts) as input for the PGC analyses, whereas individual-level genotypes were available for the UKBB. The gene-level association statistics were in turn used to perform “competitive” gene-set tests that compared enrichment of depression phenotype-associated-loci between our set of 18 candidate genes and all other genes not in the gene set, controlling for potentially confounding gene characteristics. Further analyses, which compared the 18 candidate genes to negative control sets of genes involved in type 2 diabetes, height, or synaptic processes, are described in S4.2 and reported in S11.

1.4 Results

1.4.1 Polymorphism-level analyses

Table 1.2 shows the most significant result for each of the most-studied candidate gene polymorphisms for the main effect across the eight outcomes investigated (eight main effect tests per polymorphism; first column) and the interaction effect across five moderators measured in the UKBB (32 interaction tests per polymorphism [Table S4.6]; second column). Given the number of tests conducted, there was little evidence that any effect was larger than what would be expected by chance under the null hypothesis. Only for *COMT* rs4680 on current depression severity was there was evidence of a small main effect that surpassed our liberal threshold of significance, such that

Figure 1.2: Main effects and G × E effects of 16 candidate polymorphisms on estimated lifetime depression diagnosis and current depression severity in the UK Biobank.



Effect size estimates for 16 candidate polymorphisms (in order of estimated number of tops from left to right, descending) on **a.** estimated lifetime depression diagnosis and **b.** past two-week depression symptom severity from the online mental health follow-up assessment in the UKBB sample (n = 115,257). Both polymorphism main effects and polymorphism × environmental moderator interaction effects are presented for each outcomes. Detailed descriptions of the variables, and of the association and power analysis models are provided in S3 and S4, respectively.

the rate of current depression severity scores decreased by a factor of 0.983 per copy of the G allele (*odds ratio* CI: 0.967-0.999; $p = .002$; Figure 2). Detecting an effect of this size (*genomic relative risk*

= 0.986) at α = .05 with 80% power would require a sample of approximately 214,350 individuals assuming a balanced case-control study (S4.3). Similarly, across all polymorphisms, outcomes, and exposures, on both the additive and multiplicative scales, no polymorphism-by-exposure moderation effects attained significance at α_{poly}. Failing to include all covariate × polymorphism and covariate × moderator terms as covariates, as is common in the published G × E literature [35], inflated product term test statistics on average but did not result in any additional significant effects (S10). Complete results for all outcomes are provided in S7-S10.

Despite the lack of evidence for G × E effects, all moderators exhibited large significant effects on all outcomes in the expected directions (S6). For example, experiencing childhood trauma increased odds for estimated lifetime depression diagnosis by a factor of 1.655 (z = 32.048, p = 2.33e-225) and experiencing a traumatic event in the past two years increased incidence rate of current depression severity index by a factor of 1.431 (z = 27.004, p = 1.32e-160).

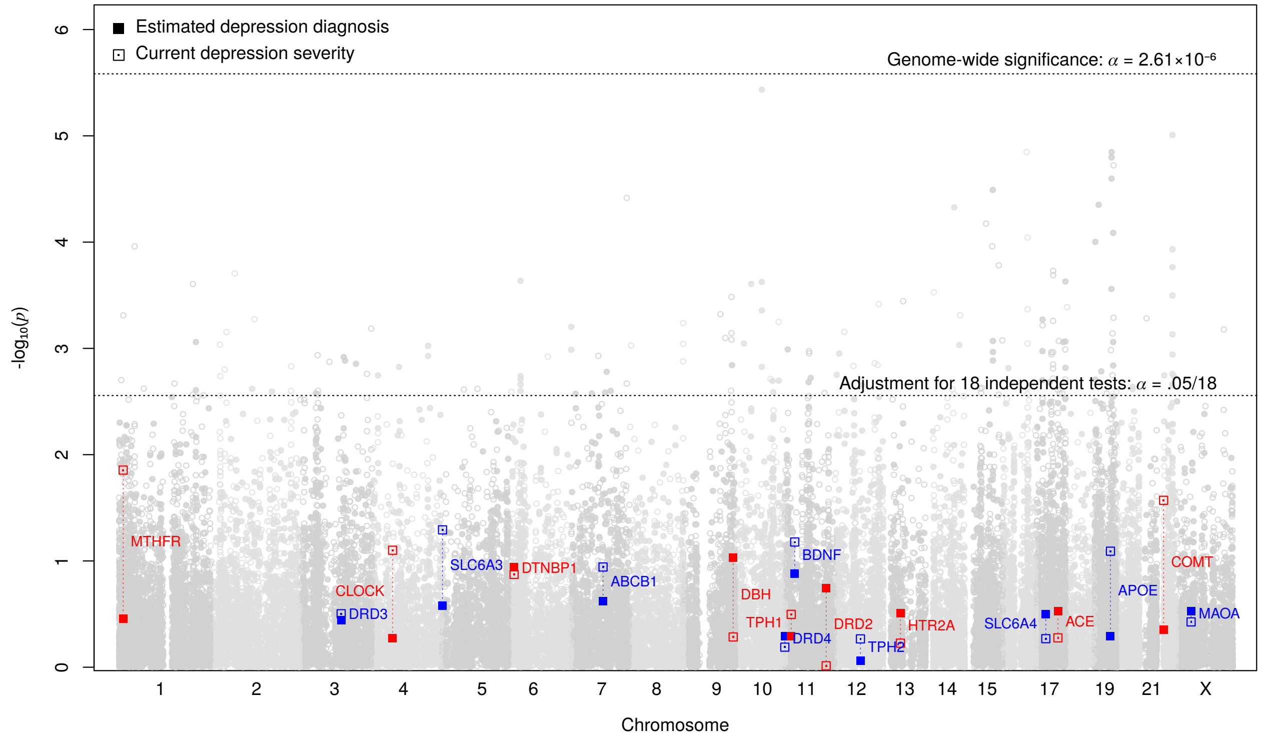
1.4.2 Gene-level analyses

Across all candidate genes and outcomes, only *DRD2* showed a significant gene-wise effect (α_{gene} =.05/18=2.78e-03) and only on PGC lifetime depression diagnosis using both the sum –log p statistic (p = 5.14e-07) as well as using the minimum p-value statistic (p = 2.74e-03; see Tables S11.1 and S11.2 for full results and section S4.2 for comparison of methods). The former estimate, based on the sum -log p statistic, was also significant at the more stringent genome-wide level (α_{GW} = .05/19,165 = 2.61e-6). *DRD2* did not exhibit a significant effect on any of the UKBB outcomes despite its high genetic correlations with the UKBB depression phenotypes (Table S3.3, Figure S3.3). Investigating the effects of the 18 genes together as a set revealed no associations with depression above what would be expected by chance under the null; the set of 18 depression candidate genes did not show stronger associations with any depression phenotype compared to all other genes at α = .05 (S11.2.1).

1.4.3 Attempted replication of top 16 loci implicated by PGC GWAS results

In order to contextualize the lack of replication of the of 16 candidate genetic polymorphisms, we sought to replicate the top 16 independent genome-wide significant loci implicated for PGC lifetime diagnosis by examining their associations with estimated lifetime diagnosis in the

Figure 1.3: Gene-wise statistics for effects of 18 candidate genes on primary depression outcomes in the UK Biobank



Gene-wise p-values across the genome, highlighting the 18 candidate polymorphisms' effects on estimated depression diagnosis (filled points) and past two-week depression symptom severity (hollow points) from the online mental health follow-up assessment in the UKBB sample (n = 115,257). Detailed descriptions of the variables, and of the association and power analysis models are provided in S3 and S4, respectively.

independent UKBB sample (see S4.5 for details). Three loci attained significance at $\alpha_{\text{poly}} = .05/16$ (rs12552, rs12658032, rs11135349; S12), which is consistent with the low power to detect small associations; median power for the 16 loci was 0.143 and the 95% CI for number of replications we'd expect given power estimates was $2 - 7$ (Figure S4.6).

1.4.4 Sensitivity of results to measurement error

One reason why candidate gene polymorphism associations detected in small samples are not replicated in large GWAS datasets is the potentially worse phenotyping and higher measurement error in predictor or outcome variables in the GWAS datasets. To investigate this possibility, we used a Monte Carlo procedure to quantify the extent to which measurement error may have impacted statistical power of our tests. As a lower bound on a candidate gene polymorphism study effect sizes, we used the minimally detectable log odds ratio for both main and interaction effects that had 50% power at $\alpha = .05$ in a balanced case/control study of 1000 individuals and where the risk allele frequency was 0.5 (e.g., for main effects, *genomic relative risk* = 1.16). Simulations demonstrated that we had ≈100% power to detect such effects under multiple severe measurement error scenarios in a sample of size typical of that in our UKBB analyses (≈ 30,000 cases and ≈ 85,000 controls; see S4.3.3). This was true even in the extreme scenario wherein half of diagnoses and half of traumatic exposures were determined via coin toss (Figure S4.5).

1.5 Discussion

The present study examined multiple types of associations between 18 highly studied candidate genes for depression and multiple depression phenotypes. The study was very well powered compared to previous candidate gene studies, with n ranging from 62,138 to 443,264 across subsamples. Despite the high statistical power, none of the most highly studied polymorphisms within these genes demonstrated substantial contributions to depression liability. Furthermore, we found no evidence to support moderation of polymorphism effects by exposure to traumatic events or socioeconomic adversity. We also found little evidence to support contributions of other common polymorphisms within these genes to depression liability excepting *DRD2*, which showed a genome-wide significant gene-wise effect on depression diagnosis in the PGC sample, though not on any outcomes in the UKBB sample. Reasons for the failure of *DRD2* to replicate in the UKBB are unclear, but could be due to sampling variability, lower statistical power in the UKBB, or false positive or negative findings. Phenotypic heterogeneity, however, is an unlikely explanation as genetic correlation estimates between depression phenotypes across samples were high (Table S3.3, Figure S3.3)—for example, PGC lifetime depression diagnosis was strongly associated with estimated lifetime depression diagnosis from the UKBB online follow-up questionnaire ($\hat{h}^2_{\text{LDSC}} = 0.085[0.004]$, $\hat{h}^2_{\text{LDSC}} = 0.057[0.007]$, respectively; $\hat{r}_g = 0.885[0.054]$, p = 2.08e-57), which was in turn strongly associated with probable lifetime diagnosis from the UKBB initial touchscreen interview ($\hat{h}^2_{\text{LDSC}} = 0.090[0.008]$, $\hat{r}_g = 0.939[0.082]$, p = 2.83e-30). Finally, as a set, depression candidate genes were no more related to depression phenotypes than non-candidate genes. Our results stand in stark contrast to the published candidate gene literature, where large, statistically significant effects are commonly reported for the specific polymorphisms in the 18 candidate genes we investigated here.

There are several features of the current investigation that set it apart from previous candidate gene replication attempts, meta-analyses of candidate gene studies, and genome-wide studies that failed to support roles for depression candidate polymorphisms. First, this is the only study to have imputed and examined the effects of several highly-studied VNTR polymorphisms in a large GWAS dataset, including 5-HTTLPR in *SLC6A4*, which was examined in 38.14% of the depression candidate gene studies we identified (see [49] for imputation details). Second, we thoroughly examined several distinct depression phenotypes (e.g., diagnosis, depressive episode recurrence, symptom count among depressed individuals) to ensure that our results did not reflect a single operationalization of depression. Some researchers have attributed the poor replicability of candidate gene findings to specificity of effects with respect to particular types of depression or stressors (e.g., prior versus subsequent depression onset with respect to stress exposure [40], recurrent versus single episode depression [53], financial versus other stress exposure [54]). As such, we examined all available depression and exposure phenotypes reflecting constructs of interest in the candidate gene literature. Results for all measures and modeling choices (e.g., multiplicative versus additive interactions), presented in detail in the supplement (S7-S11), were consistently null with respect to candidate gene hypotheses. Third, we employed exceedingly liberal significance thresholds (e.g., for polymorphism-wise analyses $\alpha_{\text{poly}} = 3.13\text{e-}03$ as opposed to the standard $\alpha_{\text{gwas}} = 5\text{e-}08$ utilized in GWAS) across all outcomes to ensure no possible effect was missed, correcting only for the number of polymorphisms we examined. As such, our results suggest that the zero or near-zero effect sizes of these candidate polymorphisms, rather than the multiple-testing burden induced by genome-wide scans, account for the previous failures of large GWAS to detect candidate polymorphisms effects. Finally, and perhaps most importantly, unlike meta-analyses that use previously published candidate gene findings, our results cannot be affected by selective publication or reporting practices that can inflate type-I errors and lead to biased representations of evidence for candidate gene hypotheses.

There are several limitations to the present investigation. First, it is possible that we failed to identify a small number of candidate gene publications and that these failures resulted in the omission of some depression candidate genes examined in ten or more publications. Nevertheless, the top nine of the eighteen identified genes accounted for 86.59% of the estimated number of studies, and it is unlikely that we omitted any depression candidate genes with popularity approaching that of, for example, *SLC6A4* or *COMT*. Second, a subset of the UKBB sample were ascertained for smoking behaviors (the BiLEVE study [55]), and controlling for genotyping batch (which differentiates the two subsamples) has the potential to induce collider bias [56]. However, only one of the sixteen candidate gene polymorphisms demonstrated allele frequency differences across these two subsamples (rs6311; =12.558, p = .002; MAF = .402 in the BiLEVE sample, MAF = .405 otherwise) and it is unlikely that ascertainment in the BiLEVE subsample unduly influenced association statistics. However, the potential influence of ascertainment in the BiLEVE subsample on interaction effect estimates, as well as other possible sources of selection-induced bias, remains unclear. Third, whereas some of phenotypes we examined closely matched standard diagnostic instruments (e.g., current depression severity was based on the widely used PHQ-9 questionnaire [57]), others were of undetermined reliability. For example, one of the nine DSM-V depression symptoms (motor agitation/retardation) was omitted from the UKBB online mental health follow-up questionnaire, and our estimated lifetime depression diagnosis phenotype required $\geq$ 4 of 8 symptoms rather than the standard $\geq$ 5 of 9 symptoms (in addition to episode duration and impairment criteria; S3.1). However, enforcing stricter case/control criteria (i.e., comparing individuals who endorsed no two-week period of either anhedonia or depressed mood throughout their lifetimes to individuals reporting recurrent episodes, endorsing $\geq$ 5 of 8 symptoms, and meeting duration and impairment criteria) failed to alter results (S7, S8, S9), despite the fact that even this diminished sample size (n = 67,304) was much larger than any previous candidate gene study we are aware of. Fourth, some of the phenotypes we examined were possibly measured with greater error than is typical in smaller candidate gene studies, an issue for which large studies are often criticized. For example, the prevalence of our measure of traumatic exposure in adulthood was uncommonly high (59.11%) and most of our retrospective measurements were likely corrupted by recall bias. However, as demonstrated in S4.3.3, even extreme measurement error cannot explain our failure to detect the relatively large effects necessary for detection in smaller samples. Further, follow-up analyses demonstrated strong effects of all environmental moderators across all outcomes (S6), suggesting that both moderators and depression phenotypes were measured with sufficient accuracy to detect known environmental effects. It is exceedingly difficult to construct a plausible measurement error model that could, for example, comfortably reconcile the large effect estimate of childhood trauma on estimated lifetime diagnosis ($odds\ ratio$ = 1.655, p = 2.96e-225) and the negligible estimate for the 5-HTTLPR × childhood trauma interaction effect ($odds\ ratio$ = 0.988,

$p = .914$) with the existence of a substantial G × E interaction effect.

The genetic underpinnings of common complex traits such as depression appear to be far more complicated than originally hoped [58, 59], and large collaborative efforts have not supported the existence of common genetic variants with large effects on depression liability [3]. In the context of our understanding of psychiatric genetics in the 1990s and early 2000s, the most studied candidate genes and the polymorphisms within them were defensible targets for association studies. However, our results demonstrate that historic depression candidate gene polymorphisms do not have detectable effects on depression phenotypes. Further, the candidate genes themselves (with the possible exception of *DRD2*) were no more associated with depression phenotypes than genes chosen at random. The present study had $\geq 99.99\%$ power at $\alpha_{\text{gwas}} = 5\text{e-}08$ to detect a main effect of the magnitude commonly reported in candidate gene studies, even allowing for extreme measurement error in both outcome and moderator phenotypes (S4.3). Thus, it is extremely unlikely that we failed to detect any true associations between depression phenotypes and these candidate genes. The implication of our study, therefore, is that previous positive main effect or interaction effect findings for these 18 candidate genes with respect to depression were false positives. Our results mirror those of well-powered investigations of candidate gene hypotheses for other complex traits including those of schizophrenia [21] and white matter microstructure [60]. The potential for self-correction is an essential strength of the scientific enterprise; it is with this mechanism in mind that we present these findings. In agreement with the recent recommendations of the National Institute of Mental Health Council Workgroup on Genomics [61], we conclude that it is time for depression research to abandon historic candidate gene and candidate gene-by-environment interaction hypotheses.

Table 1.2: Minimum p-value effect across 8 main effect models and 32 interaction effect models per polymorphism

	Polymorphism	***MAF***	**Outcome: Additive effect**	***β***	**min *p***	**Outcome: Interaction effect**	**Moderator**	**Scale**	***β***	**min *p***
1	*SLC6A4* – 5-HTTLPR†⁑	0.499	Current depression severity	0.008	.138	Lifetime episode count	TDI	primary	0.019	.041
2	*BDNF* – rs6265	0.188	Severe recurrent depression	0.018	.325	Estimated lifetime depression diagnosis	TDI	alternate	0.007	.008
3	*COMT* – rs4680	0.483	Current depression severity	-0.017	.002*	Conditional lifetime symptom count	Stressor-induced depression‡	alternate	0.048	.040
4	*HTR2A* – rs6311	0.402	Estimated lifetime depression diagnosis	0.020	.045	Estimated lifetime depression diagnosis	Childhood trauma	alternate	0.008	.072
5	*TPH1* – rs1800532	0.391	Current depression severity	-0.012	.036	Conditional lifetime symptom count	Childhood trauma	primary	-0.045	.049
6	*DRD4* – VNTR†	0.223	Touchscreen probable lifetime diagnosis (ord.)	0.022	.079	Severe recurrent depression	TDI	primary	0.011	.094
7	*DRD2* – rs1800497	0.201	PGC lifetime diagnosis	-0.019	.006	Conditional lifetime symptom count	Stressor-induced depression‡	alternate	-0.044	.134
8	*MAOA* – VNTR†	⁂	Severe recurrent depression	0.023	.073	Conditional lifetime symptom count	TDI	primary	-0.024	.014
9	*APOE* – rs429358/rs7412†	0.148	Lifetime episode count	0.019	.091	Current depression severity	Recent trauma	alternate	-0.182	.009
10	*MTHFR* – rs1801133	0.334	Current depression severity	-0.012	.034	Estimated lifetime depression diagnosis	Adult trauma	alternate	-0.007	.054
11	*CLOCK* – rs1801260	0.268	Touchscreen probable lifetime diagnosis	0.030	.013	Severe recurrent depression	TDI	primary	0.014	.012
12	*SLC6A3* – VNTR†	0.255	Touchscreen probable lifetime diagnosis	0.019	.114	Estimated lifetime depression diagnosis	Childhood trauma	alternate	-0.008	.099
13	*ACE* – in/del	0.474	Touchscreen probable lifetime diagnosis	0.016	.143	Lifetime episode count	TDI	primary	0.015	.107
14	*ABCB1* – rs1045642	0.456	PGC lifetime diagnosis	-0.006	.164	Current depression severity	Recent trauma	alternate	-0.108	.027
15	*DRD3* – rs6280	0.336	Current depression severity	-0.010	.078	Current depression severity	Recent trauma	alternate	-0.111	.031
16	*DBH* – rs1611115	0.205	Estimated lifetime depression diagnosis	-0.014	.236	Severe recurrent depression	Adult trauma	alternate	-0.005	.087

Minimum p value for each polymorphism across outcomes/moderators for additive and interaction effects (on additive and multiplicative scales), respectively. Interaction tests were not conducted in the PGC sample because moderators were unavailable for the PGC sample. Only one effect was significant after a liberal correction for the number of polymorphisms (but not for outcomes or moderators; $\alpha_{\text{poly}} = .05/16 = 3.125\times10^{-3}$). Details of each model are provided in supplement section S4, with all interaction models listed in Table S4.6. Complete results are presented in sections S7-S9. *Significant at $\alpha_{\text{poly}} = .05/16$. †VNTRs and the triallelic *APOE* polymorphism were unavailable for the PGC samples, and thus these variants were examined only across the seven UKBB outcomes. ‡Variant × stressor-induced depression estimates reflect differences in the magnitude of variant/outcome associations between individuals reporting that their depression was induced by a stressful event and those reporting otherwise. ⁑allele frequency reflects the low activity VNTR/rs25531 haplotype. ⁂*MAOA* is located on the X chromosome; frequencies were 0.336, 0.341 for females, males, respectively. *MAF* indicates the minor allele frequency in the subset of UKBB sample for whom estimated lifetime depression diagnosis was available.

Chapter 2

Stochastic Lanczos residual maximum likelihood algorithms

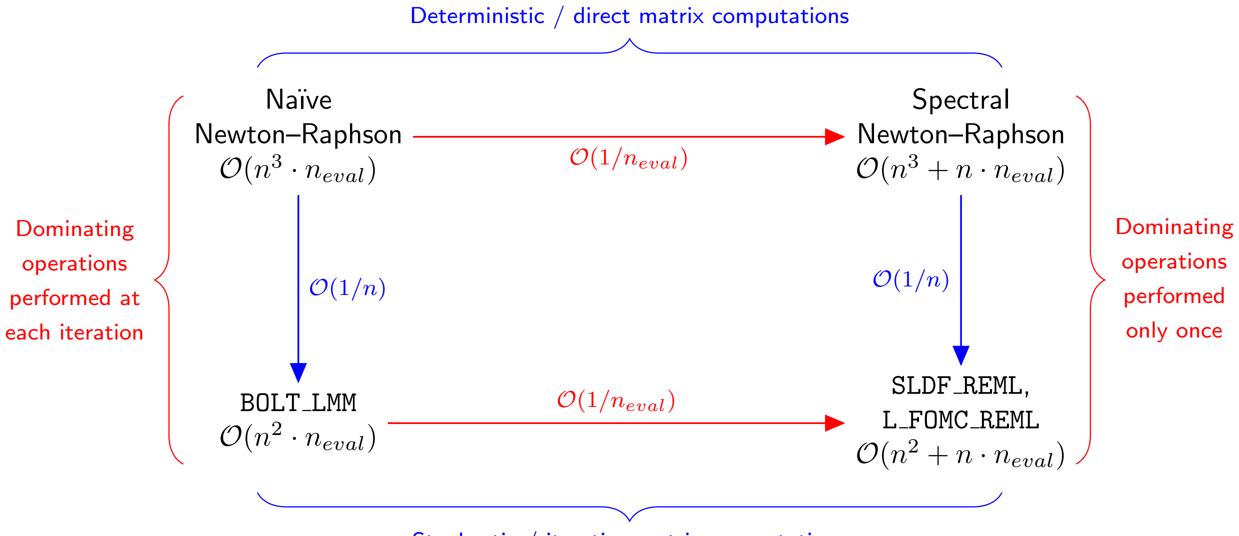
2.1 Background

Linear mixed-effects modeling (LMM) is a leading methodology employed in genome-wide association studies (GWAS) of complex traits in humans, offering the dual benefits of controlling for population stratification while permitting the inclusion of data from related individuals [62]. However, the implementation of LMM comes at the cost of increased computational burden relative to ordinary least-squares regression, particularly in performing residual maximum likelihood (REML) estimation of genomic variance components. Conventional REML algorithms require multiple $\mathcal{O}(n^3)$ or $\mathcal{O}(mn^2)$ matrix operations, where m and n are the numbers of markers and individuals, respectively, rendering them infeasible for large biobank scale data sets. Further, common numerical methods for REML estimation rely on sparse matrix methods suitable for traditional LMM applications (e.g., pedigree data or experiments with repeated measures [63]) that are inapplicable to genomics variance components models since these models involve dense relatedness matrices. As a result, the problem of increasing the computational efficiency of REML estimation of genomic variance components has generated considerable research activity [64–69].

In the case of the standard two variance component model (2.2.1), the estimation of which is the focus of the current research, previous efforts toward increasing computational efficiency fit into two primary categories: 1., reducing the number of cubic time complexity matrix operations needed to achieve convergence; and 2., substituting stochastic and iterative matrix operations for deterministic, direct methods to obtain procedures with quadratic time complexity. The first approach is embodied by the methods implemented in the FaST-LMM and GEMMA packages [64, 66, 67], which take advantage of the fact that the genetic relatedness matrix (GRM) and identity matrix comprising the covariance structure are simultaneously diagonalizable. As a result, after performing a single spectral decomposition of the GRM and a small number of matrix-vector multiplications, the REML criterion (2.2.3) and its gradient and Hessian can be repeatedly evaluated using only vector operations. The second approach is exemplified by the popular BOLT-LMM software [68, 69], which avoids all cubic operations by solving linear systems via the method of conjugate gradients (CG) and employing stochastic trace estimators in place of deterministic computations.

In the current research, we propose two algorithms, stochastic Lanczos derivative-free residual maximum likelihood (`SLDF_REML`; algorithm 3) and Lanczos first-order Monte Carlo residual maximum likelihood (`L_FOMC_REML`; algorithm 4), that combine features of both approaches (figure 2.1). Here, we translate the simultaneous diagonalizability of the heritable and non-heritable components of the covariance structure to stochastic and iterative methods via the principle of Krylov subspace shift-invariance. As a result, we only need to compute the costliest portions of the objective function once (via stochastic/iterative methods), computing all subsequent iterations of the REML optimization problem only using vector operations. We develop the theory underlying these methods and demonstrate their performance relative to previous methods via numerical experiment.

Figure 2.1: Time complexity analogies with respect to existing and proposed methods.


Deterministic / direct matrix computations
Naïve Newton–Raphson $\mathcal{O}(n^3 \cdot n_{eval})$
$\mathcal{O}(1/n_{eval})$
Spectral Newton–Raphson $\mathcal{O}(n^3 + n \cdot n_{eval})$
Dominating operations performed at each iteration
$\mathcal{O}(1/n)$
$\mathcal{O}(1/n)$
Dominating operations performed only once
BOLT_LMM $\mathcal{O}(n^2 \cdot n_{eval})$
$\mathcal{O}(1/n_{eval})$
SLDF_REML, L_FOMC_REML $\mathcal{O}(n^2 + n \cdot n_{eval})$
Stochastic / iterative matrix computations

Heuristically, the novel algorithms (bottom right) are to the stochastic, iterative algorithm implemented in the BOLT-LMM software [68, 69] (bottom left) as the direct methods exploiting the shifted structure of the the two component genomic variance component model 2.2.1 (e.g., FaST-LMM and GEMMA [64, 66] (top right) are to standard direct methods (top left). For simplicity, we assume here that the number of markers is equal to the number of observations and omit low-order terms related to the spectral conditioning of the covariance structure and the number of random vectors generated by the stochastic methods; further details are provided in Table 2.1. n_{eval} denotes the number of objective function evaluations needed to achieve convergence.

2.2 Method

We consider the two component genomic variance components model commonly employed in LMM association studies [62], which is of the form

$$\begin{gathered} y = X\beta + \frac{1}{\sqrt{m}} Z u + e, \\ u \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_g^2), \quad e \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2), \end{gathered} \tag{2.2.1}$$

where y is a measured phenotype, the $c \ll n$ columns of $X \in \mathbb{R}^{n \times c}$ are covariates (including an intercept term) with corresponding fixed effects β, and $Z \in \mathbb{R}^{n \times m}$ is a matrix of n individuals' standardized genotypes at m loci. Without loss of generality, we assume that X has full column rank; in the case of numerical rank deficiency we can simply replace X by the optimal full rank approximation generated by its economy singular value decomposition or rank revealing QR decomposition. The latent genetic effects $u \in \mathbb{R}^m$ and residuals $e \in \mathbb{R}^n$ are random variables with distributions parametrized by the heritable and non-heritable variance components, σ_g^2 and σ_e^2, respectively. The REML criterion corresponds to the marginal likelihood of $\sigma_g^2, \sigma_e^2 | K^T y$, where K^T projects to an $(n - c)$-dimensional subspace orthogonal to the covariate vectors such that the null space of K^T is exactly the column space of X [70]. In other words $K^T : \mathbb{R}^n \to \mathcal{S} \subset \mathbb{R}^{n-c}$ such that $\mathbb{R}^n = \mathcal{S} \oplus \text{col } X$. The transformed random variable $K^T y$ has the marginal distribution $K^T y \sim \mathcal{MVN}(0, \sigma_g^2 \frac{1}{m} K^T Z Z^T K + \sigma_e^2 K K^T)$, which we reparametrize as $K^T y \sim \mathcal{MVN}(0, \sigma_g^2 K^T H_\tau K)$, where

$$H_\tau = \frac{1}{m} Z Z^T + \tau I_n, \quad \tau = \sigma_e^2 / \sigma_g^2. \tag{2.2.2}$$

Here, $\frac{1}{m} Z Z^T$, which indicates the average covariance between individuals' standardized genotypes, is often referred to as the *genomic relatedness matrix* (GRM). The *REML criterion*, or marginal log likelihood, can be expressed as a function of τ:

$$\begin{aligned} \ell(\tau | K^T y) \propto & -(n - c) \ln(\hat{\sigma}_g^2(\tau)) - \hat{\sigma}_e^2(\tau)^{-1} y^T P_\tau y \\ & - \ln(\det(K^T H_\tau K)), \end{aligned} \tag{2.2.3}$$

where $P_\tau = K(K^T H_\tau K)^{-1} K^T$, and, as implied by the REML first-order (stationarity) conditions, $\hat{\sigma}_e^2(\tau)$ is the expected residual variance component given τ and $\hat{\sigma}_g^2(\tau) = \hat{\sigma}_e^2(\tau)/\tau$ [70, 71]. In practice, K is never explicitly formed.

Naïve procedures for maximizing the REML criterion require evaluating (2.2.3) or its derivatives at each iteration of the optimization procedure. Previous methods either reduce the number of necessary cubic time complexity operations to one by exploiting problem structure, or substitute quadratic time complexity iterative and stochastic matrix operations for direct computations (Figure 2.1). Here, we unify these approaches via the principle of Krylov subspace shift invariance to achieve methods that only require a single iteration of quadratic time complexity operations.

In what follows, we first present a brief survey of the Lanczos process, its applications to families of shifted linear systems, and its use in constructing Gaussian quadratures for spectral matrix functions. We assume familiarity with the method of conjugate gradients, an iterative procedure for approximating solutions to symmetric positive definite linear systems, and Gaussian quadrature, a method for approximating the integral of a given function by a well chosen weighted sum of its values; if not, see [72] and [73], respectively. We present these methods toward the goal of efficiently evaluating the quadratic form and log-determinant terms appearing in the REML criterion (2.2.3). We then present the details of the `SLDF_REML` and `L_FOMC_REML` algorithms, both of which exploit problem structure via Lanczos process-based methods in order to speed computation. Finally, we derive expressions for the computational complexity of the present algorithms, which we confirm via numerical experiment.

2.2.1 Preliminaries

The notation in this section is self-contained. Our presentation borrows from the literature extensively; further details on the (block) Lanczos procedure [72, 74], conjugate gradients for shifted linear systems [75, 76], stochastic trace estimation [77, 78], and stochastic Lanczos quadrature [79–81] are suggested in the bibliography.

2.2.1.1 Krylov subspaces

Consider a symmetric positive-definite matrix A and nonzero vector b. Define the m^{th} Krylov subspace by the span of the first $m - 1$ monomials in A applied to b; that is, $\mathcal{K}_m(A, b) =$

span$\{A^k b : k = 0, \ldots, m-1\}$. Krylov subspaces are *shift invariant*—i.e., for real numbers σ, we have $\mathscr{K}_m(A,b) = \mathscr{K}_m(A+\sigma I, b)$.

2.2.1.2 The Lanczos procedure

The Lanczos procedure generates the decomposition $AU_m = U_m T_m$, where the columns $u_1, \ldots, u_m$ of U_m form an orthonormal basis for $\mathscr{K}_m(A,b)$ and the *Jacobi matrices* $T_m \in \mathbb{R}^{m \times m}$ are symmetric tridiagonal. Choosing $u_1 = b/\|b\|$, successive columns are uniquely determined by the sequence of Lanczos polynomials $\{p_k\}_{k=1}^{m-1}$ such that each $u_k = p_{k-1}(A)u_1$ and each p_k is the characteristic polynomial of Jacobi matrix T_k consisting of the first k rows and columns of T_m. The Lanczos procedure is equivalent to the well-known method of conjugate gradients (CG) for solving the linear system $Ax = b$ in that the m^{th} step CG approximate solution $x^{(m)}$ is obtained from the above decomposition using only vector operations (see algorithm 1). The number of steps m prior to termination corresponds to the number of CG iterations need to bound the norm of the residual below a specified tolerance: $\|Ax^{(m)} - b\| < \epsilon$. The rate of convergence depends on the spectral properties of A and can be controlled in terms of the spectral condition number $\kappa(A)$. In the present application, the fact that all complex traits of interest generally have a non-trivial non-heritable component results in well-conditioned systems [68, 82].

2.2.1.3 Solving families of shifted linear systems

Having applied the Lanczos process to the *seed system* $Ax = b$, shift-invariance can be exploited to obtain the m^{th} step CG approximate solution $x_\sigma^{(m)}$ to the *shifted linear system* $A_\sigma x_\sigma = (A + \sigma I)x_\sigma = b$, only using vector operations [75]. It can be shown that any positive shift by $\sigma \geq 0$ improves the rate of convergence such that $\|A_\sigma x_\sigma^{(m)} - b\| = \frac{\delta_m}{\delta_m + \sigma}\|Ax^{(m)} - b\|$, where $\delta_m > 0$ is the m^{th} diagonal element of the Lanczos Jacobi matrix corresponding to $\mathscr{K}_m(A,b)$.

2.2.1.4 Lanczos polynomials and Gaussian quadrature

Additionally, the Lanczos polynomials comprise a sequence of orthogonal polynomials with respect to the *spectral measure*

$$\mu_{A,v}(t) = \sum_{j=1}^{\ell:\lambda_\ell \leq t} (Q^T v)_j^2,$$

Algorithm 1: Lanczos conjugate gradients solver for shifted systems (`L_Solve`)

```
input  : shift $\sigma \geq 0$, right hand sides $B = \{b_j \in \mathbb{R}^n\}_{j=1}^c$ and their Lanczos decompositions
         $U_j \in \mathbb{R}^{n\times m}$, $T_j \in \mathbb{R}^{m\times m}$ where col $U_j = \mathcal{K}_m(A, b_j)$.
output: Approximate solution $X_\sigma^{(m)} \approx (A + \sigma I)^{-1}B$.
begin
    for $j = 1, \dots, c$ do                                          // iterate over RHSs
        // initialize coefficients:
        $\delta_{1:m} \leftarrow \{(T_j)_{i,i}\}_{i=1}^m + \sigma\vec{1}_m$;                    // recycle Jacobi
        $\beta_{2:m} \leftarrow \{(T_j)_{i,i-1}\}_{i=2}^m$;                          //   coefficients
        $\omega_0 \leftarrow 0$ $\gamma_0 \leftarrow 1$ $\rho_1 \leftarrow \|b_j\|$ // initialize vectors:
        $x_j \leftarrow \vec{0}_n$;                                      // CG approx solutions
        $r_j \leftarrow b_j$;                                              // CG residuals
        $p_j \leftarrow b_j$;                                            // search directions
        // main loop:
        for $k = 1, \dots, m$ do
            // update coefficients
            $\gamma_k \leftarrow (\delta_k - \omega_{k-1}/\gamma_{k-1})^{-1}$ $\omega_k \leftarrow (\beta_{k+1}\gamma_k)^2$ $\rho_{k+1} \leftarrow -\beta_{k+1}\gamma_k\rho_k$ // update CG soln,
               residual, search dir
            $x_j \leftarrow x_j + \gamma_k p_j$ $r_j \leftarrow \rho_{k+1}(\{U_j\}_{k+1})$;                        // recycle basis
            $p_j \leftarrow r_j + \omega_k p_j$
        end
    end
    return $X_\sigma^{(m)} = [x_1|\cdots|x_c]$
end
```

where $A = Q\Lambda Q^T$ is the spectral decomposition [79, 80]. Quadratic forms $v^T f(A)v$ involving *spectral functions* $f(A) = Qf(\Lambda)Q^T$, e.g., for the matrix logarithm, $v^T(\log A)v = \sum_{i=1}^n [\ln(\lambda_i)(Q^T v)_i^2]$, can be written as Riemann–Stieltjes integrals of the form

$$v^T Q f(\Lambda) Q^T v = \int_{\lambda_1}^{\lambda_n} f(t) d\mu_{A,v}(t). \tag{2.2.4}$$

The Lanczos decomposition $AU_m = U_m T_m$ generates the weights and nodes for an m-point Gaussian quadrature approximating the above integral. Denoting the spectral decomposition of the j^{th} Jacobi matrix $T_j = W_j D_j W_j^T$ for $j = 1, \dots, m$, we approximate (2.2.4) as

$$\int_{\lambda_1}^{\lambda_n} f(t) d\mu_{A,v}(t) \approx \sum_{\ell=1}^m \omega_{j,\ell} f(\theta_{j,\ell}),$$

where $\theta_{j,\ell} = \{D_j\}_{\ell,\ell}$ and $\omega_{j,\ell} = \{e_1^T W_j\}_\ell$. As m here corresponds to the number of CG iterations needed to ensure that $\|Ax^{(m)} - v\|$ is smaller than a specified tolerance, the tridiagonal Jacobi matrices are small and calculating their spectral decompositions is computationally trivial.

2.2.1.5 Stochastic Lanczos quadrature

Stochastic Lanczos quadrature (SLQ) combines the above quadrature formulation with Hutchinson-type stochastic trace estimators [79]. Such estimators approximate the trace of a matrix $H \in \mathbb{R}^{n\times n}$ by a weighted sum of quadratic forms $\operatorname{tr}(H) \approx \frac{n}{n_{\text{rand}}} \sum_{k=1}^{n_{\text{rand}}} v_k^T H v_k$ for normalized, suitably distributed i.i.d. random *probing vectors* $\{v_j\}_{j=1}^{n_{\text{rand}}}$ [77]. The SLQ approximate trace of a spectral function of a matrix, $\operatorname{tr}(f(A))$, is then

$$\begin{aligned}
\operatorname{tr}(f(A)) &\approx \frac{n}{n_{\text{rand}}} \sum_{k=1}^{n_{\text{rand}}} v_k^T Q f(A) Q^T v_k \\
&= \frac{n}{n_{\text{rand}}} \sum_{k=1}^{n_{\text{rand}}} \int_{\lambda_1}^{\lambda_n} f(t) d\mu_{A,v_k}(t) \\
&\approx \frac{n}{n_{\text{rand}}} \sum_{k=1}^{n_{\text{rand}}} \sum_{\ell=1}^{m_\kappa} \omega_{k,\ell} f(\theta_{k,\ell}). \qquad (2.2.5)
\end{aligned}$$

Whereas the number of probing vectors n_{rand} is chosen *a priori*, the number quadrature nodes m_κ corresponds to the number of conjugate gradient iterations needed to ensure $\|A_\sigma x_{j\sigma}^{(m_\kappa)} - v_j\|$ is less than a specified tolerance for each $j = 1, \dots, n_{\text{rand}}$.

2.2.1.6 SLQ and shift invariance

For a fixed probing vector v_i, we can exploit the shift invariance of $\mathcal{K}_m(A, v_i)$ to efficiently update Gaussian quadrature generated by the corresponding Lanczos decomposition $AU_m = U_m T_m$. Again denoting the spectral decomposition of the Jacobi matrix $T_i = W_i D_i W_i^T$, the Lanczos decomposition of the shifted system is simply $A_\sigma U_m = U_m W_m (D_m + \sigma I_m) W_m^T$. Thus, given the approximation (2.2.5) for $\operatorname{tr}(f(A))$, we can efficiently compute an approximation of $\operatorname{tr}(f(A_\sigma))$ for any $\sigma > 0$. In algorithm 2 we implement a method for estimating $\operatorname{tr}(\log(A_\sigma))$ in $\mathcal{O}(()n_{\text{rand}})$ operations given the spectral decompositions of the Jacobi matrices corresponding to $\mathcal{K}_m(A, v_j)$ for probing vectors $\{v_j\}_{j=1}^{n_{\text{rand}}}$.c

Algorithm 2: Stochastic Lanczos quadrature approximate log determinant of shifted systems (`SLQ_LDet`)

```
input : shift $\sigma \geq 0$, eigenvectors and eigenvalues $W_{V_j} \in \mathbb{R}^{m \times m}, D_{V_j} \in \mathbb{R}^m$ of Jacobi matrices
        corresponding to $\mathcal{K}(A, v_j)$ for each probing vector, $j = 1, \dots, n_{\text{rand}}$
output: approximate log determinant soln $\approx \log(\det(A + \sigma I))$

begin
    soln = 0
    for $j = 1, \dots, n_{\text{rand}}$ do
        for $i = 1, \dots, m$ do
            soln $\leftarrow$ soln + $(W_{V_j})^2_{i,1} \ln((D_{V_j})_i + \sigma)$
        end
    end
    return $(n/n_{\text{rand}})$soln
end
```

2.2.1.7 Block methods

For multiple right hand sides $B = [b_1 | \cdots | b_c]$, the Lanczos procedure can be generalized to the *block Krylov subspace* $\mathcal{K}_m(A, B) = \bigotimes_{j=1}^{c} \mathcal{K}_m(A, b_j)$, resulting in a collection of Lanczos decompositions $AU_j = U_j T_j$ such that $\{U_j\}_1 = b_j / \|b_j\|$ for $j = 1, \dots, c$. This process is equivalent to block CG methods in that the Jacobi matrices can again be used to generate an approximate solution $X^{(m)}$ to the matrix equation $AX^{(m)} = B$. We provide an implementation of the block Lanczos procedure in `L_Seed` [82], employing a conservative convergence criterion defined in terms of the magnitude of the $(1, 2)$ operator norm of the residual $\|AB - X^{(m)}\|_{1 \to 2} = \max_j \|Ab_j - x_j^{(m)}\|_2$. Compared to performing c separate Lanczos procedures with respect to $\{\mathcal{K}_m(A, b_j)\}_{j=1}^{c}$, block Lanczos with respect to $\mathcal{K}_m(A, B)$, with $B = [b_1 | \cdots | b_c]$, produces the same result (for a fixed number of steps). However, block Lanczos employs BLAS-3 operations and is thus more performant, especially when implemented on top of parallelized linear algebra subroutines.

2.2.2 A derivative-free REML algorithm

We propose the stochastic Lanczos derivative-free residual maximum likelihood algorithm (`SLDF_REML`; algorithm 3), a method for efficiently and repeatedly evaluating the REML criterion, which is then subject to a zeroth-order optimization scheme. To achieve this goal, we first identify the parameter space of interest with a family of shifted linear systems. We then develop a scheme for evaluating the quadratic form $y^T P_\tau y$ and log determinant $\ln(\det(K^T H_\tau K))$ terms in the REML criterion (2.2.3) that use the previously discussed Lanczos methods to exploit this shifted structure. Specifically, after obtaining a collection of Lanczos decompositions, we can repeatedly solve the linear systems involved in the quadratic form term via Lanczos conjugate gradients and approximate the log determinant term via stochastic Lanczos quadrature.

Algorithm 3: Stochastic Lanczos derivative-free residual maximum likelihood (SLDF-REML)

input : standardized genotype matrix $Z \in \mathbb{R}^{n\times m}$ or genomic relatedness matrix $ZZ^T \in \mathbb{R}^{n\times n}$, phenotype vector $y \in \mathbb{R}^n$ covariate matrix $X \in \mathbb{R}^{n\times c}$ with $c \ll n$, range of values to consider for standardized genomic variance component $\Theta = [h^2_{\text{min}}, h^2_{\text{max}}]$, number of probing vectors for trace estimator n_{rand}, scalar optimization routine over search interval optimize($f : \Theta \to \mathbb{R}, a, b$)

output: estimated variance components $\hat{\sigma}^2_g, \hat{\sigma}^2_e$

define : qr: economy QR decomposition, Rademacher: generates Rademacher random samples, L_Seed: block Lanczos procedure as implemented in [82], eigh_tridiagonal: spectral decomposition of Hermitian tridiagonal matrix

```
begin
    $\tau_0 \leftarrow (1 - h^2_{\text{max}})/h^2_{\text{max}}$ ;                                  // minimum value of $\tau$
    $\tau_{\text{max}} \leftarrow (1 - h^2_{\text{min}})/h^2_{\text{min}}$ ;                           // maximal value of $\tau$
    $Q, R \leftarrow$ qr($X$);                                              // economy QR decomp. of $X$
    $H_0 : u \mapsto \frac{1}{m} ZZ^T u + \tau_0 u$;                              // LHS of seed system
    $S : u \mapsto u - QQ^T u$;                                          // projection to (col $X$)$^\perp$
    for $j = 1, \dots, n_{rand}$ do                                         // draw random probes
        $V_j \leftarrow$ Rademacher($n$) $V_j \leftarrow V_j/\|V_j\|$
    end
    // Lanczosdecompositions of seed systems:
    $U_y, T_y \leftarrow$ L_Seed($SH_0S, Sy$);                                       // proj. pheno.
    for $j = 1, \dots, c$ do
        $U_{X_j}, T_{X_j} \leftarrow$ L_Seed($H_0, X_j$);                                        // covariates
    end
    for $j = 1, \dots, n_{rand}$ do
        $U_{V_j}, T_{V_j} \leftarrow$ L_Seed($H_0, V_j$);                                            // probes
        // decompose Jacobi matrices for SLQ:
        $W_{V_j}, D_{V_j} =$ eigh_tridiagonal ($T_{V_j}$)
    end
    // construct REML criterion function:
    def REML_criterion ($h^2 \leq h^2_{\text{max}}$):
        global $\hat{\sigma}^2_g, \hat{\sigma}^2_e$ $\tau = (1 - h^2)/h^2$ $\sigma \leftarrow \tau - \tau_0$ $\gamma \leftarrow (1 + \tau)^{-1}$
        ldet $\leftarrow X^T$(L_Solve($\sigma, \{U_{X_j}, T_{X_j}\}_{j=1}^{c}$)) ldet $\leftarrow$ldet +SLQ_LDet($\sigma, \{W_{V_j}, D_{V_j}\}_{j=1}^{n_{rand}}$)
         qform $\leftarrow y^T S$(L_Solve($\sigma, U_y, T_y$)) $\hat{\sigma}^2_e \leftarrow$ qform$/(n - c)$ $\hat{\sigma}^2_g \leftarrow \hat{\sigma}^2_e/\tau$ return
         $(n - c)\ln(\hat{\sigma}^2_g) -$ ldet $-$ qform$/\hat{\sigma}^2_e$
    // apply zeroth-order optimization routine:
    optimize(REML_criterion,$h^2_{\text{min}}, h^2_{\text{max}}$) return $\hat{\sigma}^2_g, \hat{\sigma}^2_e$
end
```

2.2.2.1 The parameter space as shifted linear systems

Given a range of possible values of the *standardized genetic variance component*, or *heritability*,

$$h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2), \quad h^2 \in [h^2_{\min}, h^2_{\max}], \tag{2.2.6}$$

we set $\tau_0 = (1 - h^2_{\max})/h^2_{\max}$ and define $H_0 = H_{\tau_0}$, noting that for all $\tau \in \Theta = \{(1-h^2)/h^2 : h^2 \in [h^2_{\min}, h^2_{\max}]\}$, the spectral condition number of H_τ will be less than that of H_0 as the identity component of H_τ will only increase. Further, we have now identified elements of our parameter space $\tau \in \Theta$ with the family of shifted linear systems

$$\mathcal{H}_{\tau_0} = \{H_\sigma = H_\tau = H_0 + \sigma I_n : \sigma = \tau - \tau_0\}.$$

For any vector v for which we have computed the Lanczos decomposition $H_0 U = UT$ with the first column of U equal to $v/\|v\|$, we can use algorithm 1 to obtain the CG approximate solution $x_\sigma \approx H_\sigma^{-1} v$ for all $\sigma \geq 0$ in $O(n)$ operations.

2.2.2.2 The quadratic form

Directly evaluating the quadratic form

$$y^T P_\tau y = y^T K (K^T H_\tau K)^{-1} K^T y \tag{2.2.7}$$

is computationally demanding and is typically avoided in direct estimation methods [70, 71]. Writing the complete QR decomposition of the covariate matrix $X = [Q_X | Q_{X^\perp}] R$ allows us to define $K^T = Q^T_{X^\perp}$, noting that substituting $Q_{X^\perp} Q^T_{X^\perp}$ for K^T preserves the value of (2.2.7). $Q_{X^\perp} Q^T_{X^\perp}$ is equivalent to the orthogonal projection operator $S : v \mapsto v - Q_X Q_X^T v$, which admits an efficient implicit construction and is computed in $O(nc^2)$ operations via the *economy* QR decomposition $X = Q_X R_X$. Then, reexpressing (2.2.7) as $y^T S (S H_\tau S)^\dagger S y$, we can use the Lanczos process to construct an orthonormal basis and corresponding Jacobi matrix for the Krylov subspace $\mathcal{K}(SH_0S, Sy)$. We can then obtain the CG approximation of $y^T S(SH_\sigma S)^{-1} Sy$ using vector operations as, for any shift σ, we have $y^T S (S H_\sigma S)^\dagger S y = y^T S (S H_0 S + \sigma I_n)^{-1} S y$ (see Lemma 1 in Additional File 1 for proof).

2.2.2.3 The log determinant

We use an equivalent formulation [70, 83] of the term $\ln(\det(K^T H_\tau K))$, rewriting it as

$$\ln(\det(H_\tau)) + \ln(\det(X^T H_\tau^{-1} X)) - \ln(\det(X^T X)).$$

The $\det(X^T X)$ term is constant with respect to τ and can be disregarded. For $c \ll n$, $\det(X^T H_\tau^{-1} X)$ is computationally trivial via direct methods given $H_\tau^{-1} X$, which we can compute for all parameter values of interest in $O(n)$ operations having first applied the block Lanczos process with respect to $\mathcal{K}(H_0, X)$. Computing the block Lanczos decomposition corresponding to $\mathcal{K}(H_0, X)$, which is only performed once, unfortunately scales with the number of covariates c, a disadvantage not shared by our second algorithm (algorithm 4). The remaining term, $\ln(\det(H_\tau))$, is approximated by applying SLQ (algorithm 2) to a special case of (2.2.5): We rewrite the log determinant as the trace of the matrix logarithm

$$\begin{aligned}\ln(\det(H_\tau)) &= \operatorname{tr}\left(\log(H_\tau)\right)\\ &= \operatorname{tr} Q[\ln(\lambda_1 + \sigma)| \cdots | \ln(\lambda_n + \sigma)]Q^T,\end{aligned}$$

where we have spectrally decomposed $H_0 = Q\Lambda Q^T$ for some $\tau_0 \le \tau$ with $\sigma = \tau - \tau_0$. We draw n_{rand} *i.i.d.* normalized Rademacher random vectors $v_1, \dots, v_{n_{\text{rand}}}$, where each element of each vector v_i takes values of either $1/\|v_i\|$ or $-1/\|v_i\|$ with equal probability. The SLQ approximate of the log determinant for the seed system is

$$\ln(\det(H_\sigma)) \approx \frac{n}{n_{\text{rand}}} \sum_{i=1}^{n_{\text{rand}}} \sum_{\ell=1}^{m_i} \omega_{i,\ell} \ln(\theta_{i,\ell} + \sigma),$$

where the weights $w_{i,\ell}$ and nodes $\theta_{i,\ell}$ are respectively derived by using the Lanczos process to construct orthonormal bases for $\mathcal{K}(H_0, v_i)$ (in practice, we apply block Lanczos to $\mathcal{K}(H_0, (v_1, \dots, v_{n_{\text{rand}}}))$ [79, 80].

2.2.2.4 The `SLDF_REML` algorithm

Stochastic Lanczos derivative-free residual maximum likelihood (`SLDF_REML`; algorithm 3), conceptually similar to the derivative-free algorithm of Graser and colleagues [71], applies the previously introduced Lanczos methods to approximate the above reparametrization of the REML criterion. Shift-invariance is then exploited such that, with the exception of the initial Lanczos decompositions, the REML log likelihood can be repeatedly evaluated using only vector operations. `SLDF_REML` takes a phenotype vector $y \in \mathbb{R}^n$, a covariate matrix $X \in \mathbb{R}^{n \times c}$, either the genetic relatedness matrix $ZZ^T \in \mathbb{R}^{n \times n}$ or the standardized genotype matrix $Z \in \mathbb{R}^{n \times m}$ (in which case the action of the GRM as a linear operator is coded implicitly as $v \mapsto Z(Z^T v)$), and a range of possible standardized genomic variance component values $\Theta = [h^2_{\text{min}}, h^2_{\text{max}}]$ as arguments and generates a function `REML_criterion`: $\Theta \to \mathbb{R}$ that efficiently computes the log-likelihood of $\tau | K^T y$. This function is then subject to scalar optimization via Brent's method, which is feasible given the low cost of evaluation and low dimension of Θ. Hyperparameters include the number of probing vectors to be used for the SLQ approximation of the log determinant n_{rand}, as well as tolerances corresponding to the REML criterion, parameter estimates, and the Lanczos residual norms. Convergence to a given tolerance on a sensible scale is ensured by optimizing with respect to the heritability $h^2 \in \Theta \subseteq [0, 1]$ and evaluating the REML criterion at $\tau = (1 - h^2)/h^2$. The REML criterion can be repeatedly evaluated in $\mathcal{O}(()n)$ operations, making high accuracy computationally feasible.

2.2.3 A first-order Monte Carlo REML algorithm

We additionally propose the Lanczos first-order Monte Carlo residual maximum likelihood algorithm (`L_FOMC_REML`; algorithm 4), which also takes advantage of the shifted structure of the standard genomic variance components model to speed computation. We first present the related first-order algorithm implemented in the efficient and widely-used BOLT-LMM software [68, 69], which we refer to as `BOLT_LMM` and of which the proposed `L_FOMC_REML` algorithm is a straightforward extension.

2.2.3.1 `BOLT_LMM` (First-order Monte Carlo REML)

The `BOLT_LMM` algorithm is based on the observation that at stationary points of the REML criterion (2.2.3), the first-order REML conditions (i.e., $\nabla \ell = 0$) imply that

$$\mathbb{E}[\tilde{u}^T \tilde{u} | y] = \tilde{u}^T \tilde{u}, \quad \mathbb{E}[\tilde{e}^T \tilde{e} | y] = \tilde{e}^T \tilde{e}, \tag{2.2.8}$$

where $\tilde{u}$ and $\tilde{e}$ are the best linear unbiased predictions (BLUPs) of the latent genetic effects and residuals, respectively [84]. The BLUPs are functions of τ given by

$$\begin{aligned} \tilde{u}(\tau) &= m^{-1/2} Z^T S \acute{H}_\tau^{-1} S y, \\ \tilde{e}(\tau) &= \tau \acute{H}_\tau^{-1} S y, \end{aligned} \tag{2.2.9}$$

where we have defined $\acute{H}_\tau = \frac{1}{m} S Z Z^T S + \tau I_n$. The expectations (2.2.8) are approximated via the following stochastic procedure: Monte Carlo samples of the latent variables, $\breve{u}_k \overset{i.i.d.}{\sim} \mathcal{MVN}(0, I_m)$, $\breve{e}_k \overset{i.i.d.}{\sim} \mathcal{MVN}(0, S)$ are used to generate samples of the projected phenotype vector

$$\breve{y}_k = m^{-1/2} S Z \breve{u}_k + \breve{e}_k, \quad k = 1, \ldots n_{\text{rand}}.$$

BLUPs are then computed as in (2.2.9), yielding the approximations

$$\begin{aligned} \underset{\text{MC}}{\mathbb{E}}[\tilde{u}^T \tilde{u} | y] &= \frac{n_{\text{rand}}^{-1}}{\sqrt{m}} \sum_{k=1}^{n_{\text{rand}}} \left\| Z^T S \acute{H}_\tau^{-1} S \breve{y}_k \right\|^2, \\ \underset{\text{MC}}{\mathbb{E}}[\tilde{e}^T \tilde{e} | y] &= n_{\text{rand}}^{-1} \sum_{k=1}^{n_{\text{rand}}} \left\| \tau \acute{H}_\tau^{-1} S \breve{y}_k \right\|^2. \end{aligned}$$

Using the above expressions, Loh et al. [68, 69] apply a zeroth-order root-finding algorithm to the quantity

$$f_r(\tau) = \ln \left[\frac{\tilde{u}^T \tilde{u}}{\tilde{e}^T \tilde{e}} \right] - \ln \left[\frac{\mathbb{E}_{\text{MC}}[\tilde{u}^T \tilde{u} | y]}{\mathbb{E}_{\text{MC}}[\tilde{e}^T \tilde{e} | y]} \right],$$

noting that $f_r = 0$ is a necessary condition (and, in practice, a sufficient condition) for (2.2.8). Using CG to approximate solutions to the linear systems involved in BLUP computations results in an efficient REML estimation procedure involving $O(n \cdot m \cdot n_{\text{rand}})$ operations for well-conditioned covariance structures (i.e., for nontrivial non-heritable variance component values). As noted in [69], implicit preconditioning of H_0 can be achieved by including the first few right singular vectors of the genotype matrix (or eigenvectors of the GRM) as columns of the covariate matrix X.

Algorithm 4: Lanczos first-order Monte Carlo residual maximum likelihood (`L_FOMC_REML`)

```
input  : standardized genotype matrix $Z \in \mathbb{R}^{n \times m}$, phenotype vector $y \in \mathbb{R}^n$, covariate matrix
         $X \in \mathbb{R}^{n \times c}$ with $c \ll n$, range of values to consider for standardized genomic variance
         component $\Theta = [h^2_{\min}, h^2_{\max}]$, number of Monte Carlo samples $n_{\mathrm{rand}}$, zeroth order
         scalar root finding routine root($f : \Theta \to \mathbb{R}, h^2_{\min}, h^2_{\max}$)
output: estimated value of heritable variance component $\hat{\sigma}^2_g$
define : qr: economy QR decomposition, Gaussian: generates standard normal random
         samples, L_Seed: block Lanczos procedure as implemented in [82]

begin
    $\tau_0 \leftarrow (1 - \sigma^2_{g\max})/\sigma^2_{g\ \max}$ ;                    // minimum value of $\tau$
    $\tau_{\max} \leftarrow (1 - \sigma^2_{g\ \min})/\sigma^2_{g\ \min}$ ;                 // maximal value of $\tau$
    $Q, R \leftarrow$ qr($X$);                                   // economy QR decomp. of $X$
    $S : u \mapsto u - QQ^T u$;                                 // projection to (col $X$)$^\perp$
    $\acute{H}_0 : u \mapsto \frac{1}{m} SZZ^T Su + \tau_0 u$;                        // LHS of seed system
    for $k = 1, \dots, n_{\mathrm{rand}}$ do                              // sample latent variables
        $\check{u}_k \leftarrow$ Gaussian($m$) $\check{e}_k \leftarrow$ Gaussian($n$) $\check{e}_k \leftarrow S\check{e}_k$;          // latent residual
        $\check{g}_k \leftarrow m^{-1/2} SZ\check{u}_k$;                             // latent genetic value
    end
    $U_y, T_y \leftarrow$ L_Seed($\acute{H}_0, Sy$);                 // Lanczos decompositions of seed systems
    for $k = 1, \dots, n_{\mathrm{rand}}$ do                                // can use block Lanczos
        $U_{\check{e}_k}, T_{\check{e}_k} \leftarrow$ L_Seed($H_0, \check{e}_k$) $U_{\check{g}_k}, T_{\check{g}_k} \leftarrow$ L_Seed($H_0, \check{g}_k$)
    end
    def f_reml ($h^2 \leq h^2_{\max}$):
        global $\hat{\sigma}^2_e$;                             // construct objective for root finding
        $\tau = (1 - h^2)/h^2$ $\sigma \leftarrow \tau - \tau_0$ soln $\leftarrow$ L_Solve($\sigma, U_y, T_y$) ;           // compute BLUPs
        $\tilde{u} \leftarrow m^{-1/2} Z^T S$(soln) $\tilde{e} \leftarrow \sqrt{\tau}$(soln) for $k = 1, \dots, n_{\mathrm{rand}}$ do            // MC samples
            soln_u[$k$] $\leftarrow$ L_Solve($\sigma, U_{\check{u}_k}, T_{\check{u}_k}$) soln_e[$k$] $\leftarrow$ L_Solve($\sigma, U_{\check{e}_k}, T_{\check{e}_k}$)
            $\check{\tilde{u}}_k \leftarrow m^{-1/2} Z^T S$(soln_u[$k$] $+ \sqrt{\tau}$soln_e[$k$]) $\check{\tilde{e}}_k \leftarrow \sqrt{\tau}$(soln_u[$k$]$+\sqrt{\tau}$soln_e[$k$])
        end
        E[$\tilde{u}^T \tilde{u}$] $\leftarrow n^{-1}_{\mathrm{rand}} \sum_{k=1}^{n_{\mathrm{rand}}} \|\check{\tilde{u}}_k\|^2$ E[$\tilde{e}^T \tilde{e}$] $\leftarrow n^{-1}_{\mathrm{rand}} \sum_{k=1}^{n_{\mathrm{rand}}} \|\check{\tilde{e}}_k\|^2$ $\hat{\sigma}^2_e \leftarrow \tilde{e}^T \tilde{e}/(n - c)$ return
         $\ln(\tilde{u}^T \tilde{u}/\tilde{e}^T \tilde{e}) - \ln(\tilde{u}^T \tilde{u}/$E[$\tilde{e}^T \tilde{e}$])
    $\hat{h}^2 \leftarrow$ root(f_reml,$h^2_{\min}, h^2_{\max}$);        // apply zeroth-order root finding routine
    $\hat{\sigma}^2_g \leftarrow \hat{\sigma}^2_e(\hat{h}^2/(1 - \hat{h}^2))$ return $\hat{\sigma}^2_g, \hat{\sigma}^2_e$
end
```

2.2.3.2 The L_FOMC_REML algorithm

The BOLT_LMM algorithm described above involves solving $n_{\text{rand}} + 1$ linear systems

$$\acute{H}_{\tau_\ell}^{-1} S\breve{y},\ \acute{H}_{\tau_\ell}^{-1} S\breve{y}_1,\ \dots,\ \acute{H}_{\tau_\ell}^{-1} S\breve{y}_{n_{\text{rand}}},$$

at each iteration of the optimization scheme in order to compute BLUPs of the latent variables for the observed phenotype vector and each of the Monte Carlo samples. However, each iteration involves spectral shifts of the left hand side of the form

$$\acute{H}_{\tau_{\ell+1}}^{-1} = (\acute{H}_{\tau_\ell} + \sigma I_n)^{-1}, \quad \sigma = (\tau_{\ell+1} - \tau_\ell).$$

As in the SLDF_REML algorithm, the underlying block Krylov subspace is invariant to these shifts (i.e., $\mathcal{K}_m(\acute{H}_\tau, Y) = \mathcal{K}_m(\acute{H}_\tau + \sigma I, Y)$, where $Y = [y|\breve{y}_1|\cdots|\breve{y}_{n_{\text{rand}}}]$). Thus, having performed the Lanczos process for an initial parameter value τ_0, we can use L_Solve (algorithm 1) to obtain the block CG approximate solution $X_\sigma^{(m)} \approx \acute{H}_{\tau+\sigma}^{-1} Y$ in $O(n \cdot n_{\text{rand}})$ operations. We are thus able to avoid solving linear systems in all subsequent iterations, though the relatively small number of matrix-vector products involved in computing BLUPs for the latent genetic effects at each step are unavoidable. The requirement of the genotype matrix for computing (2.2.9) prevents both L_FOMC_REML and BOLT_LMM from efficiently exploiting precomputed GRMs.

2.2.4 Comparison of methods

We compare theoretical and empirical properties of our proposed algorithms, SLDF_REML and L_FOMC_REML, to those of BOLT_LMM.

2.2.4.1 Computational complexity

In contrast to BOLT_LMM, the Lanczos-decomposition based algorithms we have proposed only need to perform the computationally demanding operations necessary to evaluate the REML criterion once. As such, we differentiate between *overhead* computations, which occur once and do not depend on the number of iterations needed to achieve convergence, and *per-iteration* computations, which are repeated until convergence of the optimization process (table 2.1,figure 2.4).

The overhead computations of `SLDF_REML` are dominated by the need to construct bases for the $n_{\text{rand}}+c+1$ subspaces $\mathcal{K}(H_0, [\check{v}_1, \dots, \check{v}_{n_{\text{rand}}}, x_1, \dots, x_c, y])$, and are thus $\mathcal{O}(n^2(n_{\text{rand}}+c)n_\kappa)$ when a precomputed GRM is available and $\mathcal{O}(2m \cdot n(n_{\text{rand}}+c)n_\kappa)$ otherwise. Here, n_κ denotes the number of Lanczos iterations needed to achieve convergence at a pre-specified tolerance and increases with h^2_{max}. Subsequent iterations are dominated by the cost of solving $c+1$ shifted linear systems via `L_Solve` and are thus $\mathcal{O}(n \cdot c \cdot n_\kappa)$. The overhead computations in `L_FOMC_REML` are dominated by the Lanczos decompositions corresponding to the $2n_{\text{rand}}+1$ seed systems, where the GRM is implicitly represented in terms of the standardized genotype matrix, and is thus $\mathcal{O}(4m \cdot n \cdot n_{\text{rand}} \cdot n_\kappa)$. Operations of equivalent complexity are needed at *every* iteration of `BOLT_LMM`.

2.2.4.2 Numerical experiments

We compared wall clock times for genomic variance component estimation for height in nested random subsets of 16,000, 32,000, 64,000, 128,000, and 256,000 unrelated ($\hat{\pi} < .05$) European ancestry individuals from the widely used UK Biobank data set [1]. All analyses included 24 covariates consisting of age, sex, and testing center and used hard-called genotypes from 330,723 array SNPs remaining after enforcing a 1% minor allele frequency cutoff. We compared `SLDF_REML`, with and without a precomputed GRM, to `L_FOMC_REML` which requires the genotype matrix. For the novel algorithms, absolute tolerances for the Lanczos iterations and the REML optimization procedure were set to 5e-5 and 1e-5, respectively. Additionally, we compared our interpreted Python 3.6 code to BOLT-LMM versions 2.1 and 2.3.3 (C++ code compiled against the Intel MKL and Boost libraries) [68, 69, 86, 87]. We ran each algorithm twenty times per condition, trimming away the two most extreme timings in each condition. Mirroring the default settings of the BOLT-LMM software packages, we set $n_{\text{rand}} = 15$ across both of our proposed methods.

Novel algorithms were implemented in the Python v3.6.5 computing environment [82], using NumPy v1.14.3 and SciPy v1.1.0 compiled against the Intel Math Kernel Library v2018.0.2 [87–89]. Optimization was performed using SciPy's implementation of Brent's method, with convergence determined via absolute tolerance of the standardized genomic variance component $\hat{h}^2$. Timing results (table 2.2, figure 2.3, figure 2.5). do not include time required to read genotypes into memory, or, when applicable, to compute GRMs, and reflect total running time on an Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz with 32 physical cores and 1 terabyte of RAM. Timing experiments excluded methods with cubic time complexity, including GCTA, FaST-LMM, and GEMMA. Accuracy was assessed by comparing heritability estimates generated by the stochastic algorithms to those estimated via the direct, deterministic average-information Newton–Raphson algorithm as implemented in the GCTA software package v1.92.0b2 [65] (figure 2.2, figure 2.5).

2.3 Results

Across 20 replications per condition for random subsamples of n=16,000 to 256,000 unrelated European-ancestry individuals, both `SLDF_REML` and `L_FOMC_REML` produced heritability estimates for height consistent with those generated by the GCTA software package (figure 2.2, figure 2.5). For large samples, the novel algorithms achieved greater accuracy than either version of BOLT-LMM (e.g., for n=250,000, mean-squared error was 1.74×10^{-6} for BOLT-LMM v2.3.2 versus 1.24×10^{-7} for `L_FOMC_REML`). Particularly, the time required per additional iteration after initial overhead computations was low for the novel algorithms (e.g., $\bar{t}$=20.07 minutes for BOLT-LMM v2.3.2 versus 2.06 minutes for `SLDF_REML`; table 2.2), enabling increased precision at minor cost. With respect to total timings, `SLDF_REML` dramatically outperformed all other methods when the precomputed GRM was available (figure 2.3, figure 2.3), which we expect whenever the number of markers exceeds the sample size. Examining methods taking genotype matrices as inputs, `SLDF_REML` and `L_FOMC_REML` performed similarly, whereas BOLT-LMM v2.3.2 converged more quickly than either in smaller samples (figure 2.3), though the differences for n=256,000 were relatively minor (e.g., $\bar{t}$=91.09 minutes for BOLT-LMM v2.3.2 versus 102.21 minutes for `SLDF_REML`; table 2.2). The older version of BOLT-LMM, v2.1, performed significantly more slowly than any of the other implementations examined (e.g., average wall clock time was 177.95 minutes at n=256,000), demonstrating the importance of implementation optimization.

As the computations needed to compute the Lanczos decompositions in `L_FOMC_REML` and BOLT-LMM v2.3.2 are equivalent in time and memory complexity, we expect that an optimized compiled-language implementation of `L_FOMC_REML` would reduce the overhead computation time by a significant linear factor (≈3 for n=256,000, comparing the sum of the overhead time and single objective function evaluation time for BOLT-LMM v2.3.2 to its total running time; figure 2.3). Consistent with theory, the wall clock times per objective function evaluation for the novel algorithms were trivial given the Lanczos decompositions (e.g., for n=256,000, $\bar{t}$ = 2.06 versus 20.07

minutes for `L_FOMC_REML` and BOLT-LMM v2.3.2, respectively; table 2.2, figure 2.4).

2.4 Discussion

We have proposed stochastic algorithms for estimating the two variance component model (2.2.1), both of which theoretically offer substantial time savings relative to existing methods. Our methods capitalize on the principle of Krylov subspace shift invariance to reduce the number of steps involving n^2) or $\mathcal{O}(mn)$ computations to one, whereas existing methods perform equivalent computations at each iteration of the REML optimization procedure. For large samples, when taking genotype matrices as inputs, our interpreted-language implementations of `L_FOMC_REML` and `SLDF_REML` [82] produced more accurate variance component estimates than the highly-optimized, compiled BOLT-LMM implementations, while taking similar amounts of time. Thus, we expect comparably-optimized implementations of the novel algorithms to compute high accuracy REML estimates in close to the time required by BOLT-LMM v2.3.2 for a *single* objective function evaluation. Further, in contrast to the `BOLT_LMM` algorithm, which requires the genotype matrix, `SLDF_REML` can exploit precomputed GRMs to reduce operation count by an $\mathcal{O}(2m/n)$ factor (table 2.1), which yields dramatic time savings when the number of markers greatly exceeds the number of individuals (figure 2.3). While GRM precomputation is itself $\mathcal{O}(mn^2)$, it can be effectively and asynchronously parallellized across multiple compute nodes, substantially mitigating computational burden (though we note that serial input/output constraints can interfere with efficient parallelization). However, as the `L_FOMC_REML` algorithm involves the computation of BLUPs of SNP effects, `L_FOMC_REML` is preferable to `SLDF_REML` when BLUP estimates are desired for prediction (as in [90]).

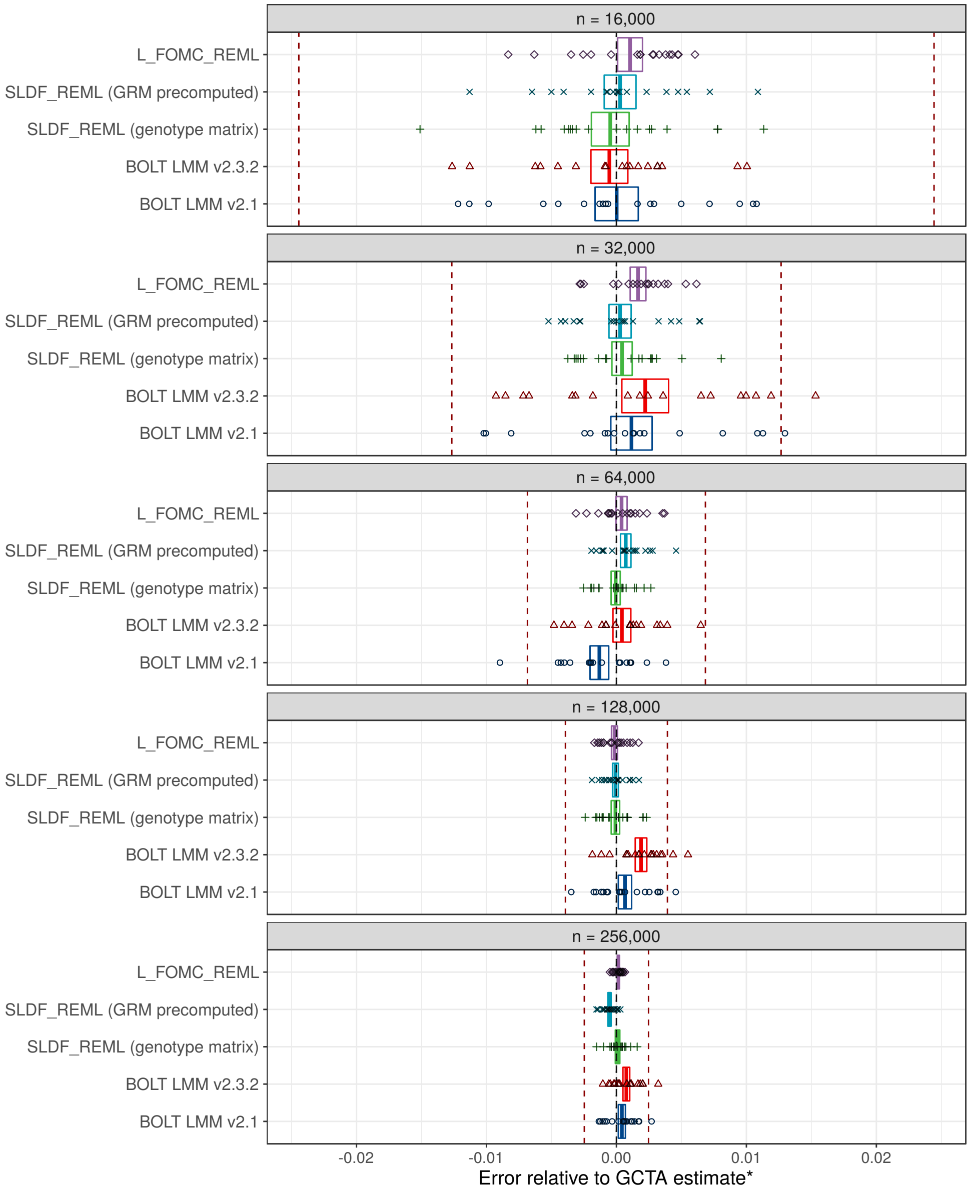
There are several limitations to the proposed approaches. First, `SLDF_REML`, which benefits from the ability to take GRMs as input, depends linearly on the number of included covariates, which might grow prohibitive in samples spanning numerous genotyping batches and ascertainment locations. However, as in `BOLT_LMM`, `L_FOMC_REML` requires $\mathcal{O}(mn)$ matrix multiplications for BLUP computation at each step of the REML optimization procedure, whereas `SLDF_REML` requires only vector operations. Thus, though the options provided by the two novel algorithms increase researchers' flexibility overall, the choice of whether to employ `SLDF_REML` versus `L_FOMC_REML` is problem-specific and necessitates greater researcher attention to resource allocation. For example, even when a precomputed GRM is available, it might be preferable to use L_FOMC_REML if BLUPs of latent SNP effects are desired. On the other hand, if a researcher intends to sequentially analyze a large number of phenotypes in a relatively small sample of individuals, it might prove most efficient to compute a GRM, despite the involved computational burden, in order to speed subsequent computations by supplying the GRM to the SLDF_REML algorithm. Second, neither algorithm mitigates the substantial $\mathcal{O}(mn)$ or $\mathcal{O}(n^2)$ memory complexity common to all algorithms for REML estimation of genomic variance components, requiring that researchers have access to high-memory compute nodes to work with large samples (though we note that neither of the novel algorithms substantial increases this burden either). Finally, for the same reasons that the spectral decomposition-based direct methods implemented in the FaST-LMM and GEMMA packages [64, 66, 67] are restricted to the simple two component model (2.2.1) (i.e., whereas the GRM and identity matrix are simultaneously diagonalizable, the same doesn't hold for arbitrary collections of three or more symmetric positive semidefinite matrices), the shift-invariance property exploited by the proposed methods does not extend to multiple genomic variance components. Given that the two component model is insufficient for precise heritability estimation for many complex traits [91], our novel algorithms apply to the particular, though common, tasks of variance component and BLUP estimation for LMM in association studies.

Despite these limitations, the proposed algorithms have clear advantages over existing methods in terms of flexibility, accuracy, and speed of computation. We provide both pseudocode and heavily annotated Python 3 implementations [82] to facilitate their incorporation into existing software packages. Further, though our algorithms are restricted to the two variance component model, they can be used to generate the inputs necessary for estimation of more complex models, such as the mixture model estimated via variational approximation implemented in [68], and thus have applications to non-infinitesimal models. Finally, we suggest that the methods presented in our theoretical development, in particular stochastic trace estimation and stochastic Lanczos quadrature, are likely to find uses in REML estimation of other models of interest to researchers in genomics. In particular, we suggest the development of models that exploit Krylov subspace shift-invariance to speed up variance/covariance component estimation for the case of multivariate phenotypes as a target for future research. Such models necessarily involve the computation or approximation of Hessian matrices, thereby introducing additional complexity in comparison to the univariate case considered above. However, the extension of fast cubic complexity methods based on the spectral decomposition of the covariance matrix [64, 66] to the multivariate case [67] suggests the potential for multivariate analogues of the algorithms presented here.

2.5 Conclusions

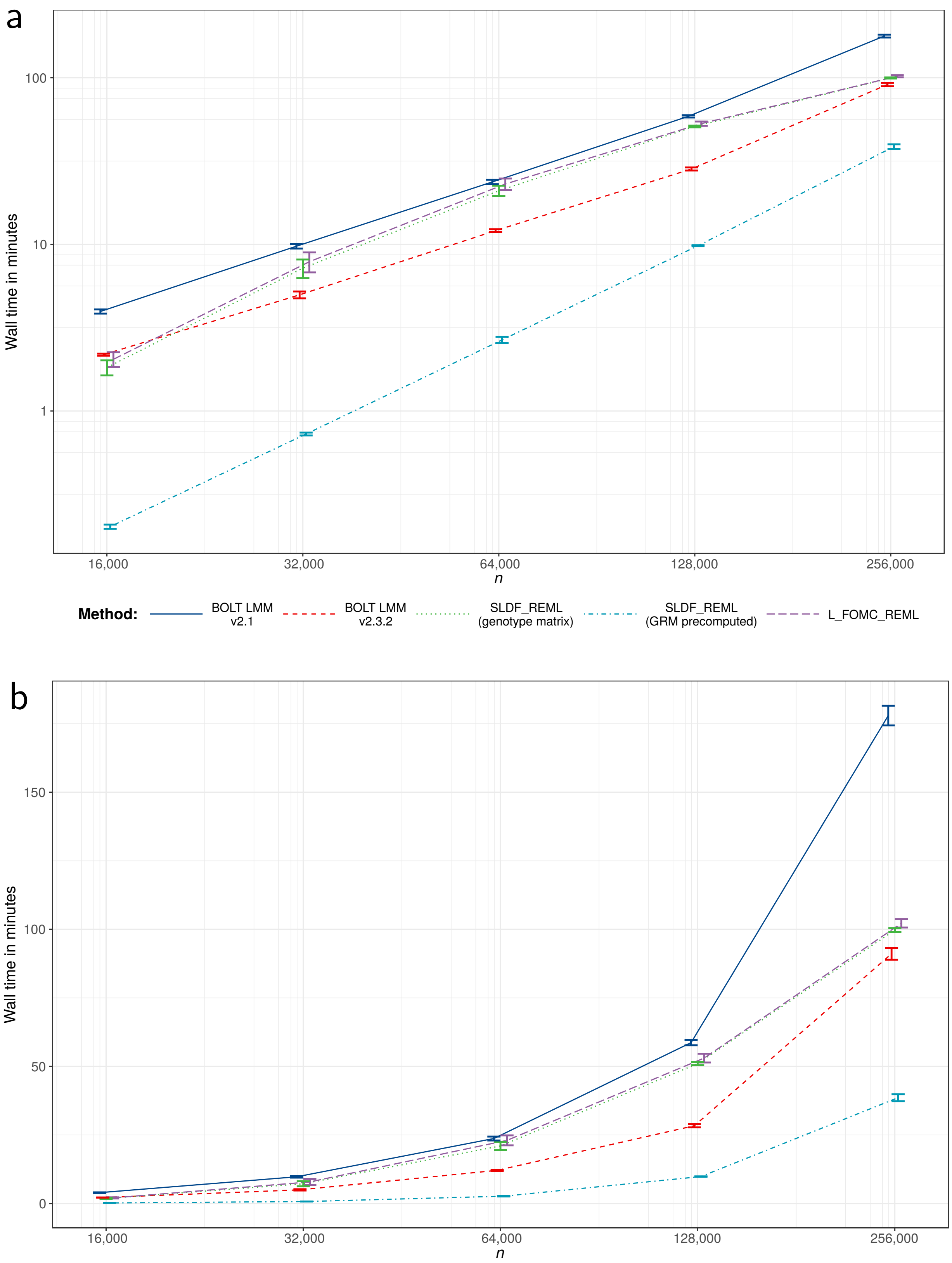
The proposed algorithms, `SLDF_REML` and `L_FOMC_REML`, unify previous approaches to estimating the two variance component model (2.2.1) by exploiting the simultaneous diagonalizability of the covariance structure components while avoiding matrix operations with cubic time complexity. As a result, the most expensive operations only need to be performed once, as with the spectral decomposition performed in the FaST-LMM and GEMMA software packages [64, 66, 67], but these operations consist only of matrix-vector products, as in the BOLT-LMM software package [68, 69]. All but one iteration of the REML optimization procedure requires only vector operations, yielding increased speed and numerical precision relative to existing methods. Furthermore, the unique strengths of the two methods lead to a flexible approach depending on researcher goals: `SLDF_REML` is capable of operating on precomputed GRMs when available, whereas `L_FOMC_REML` can generate BLUPs of latent SNP effects without added computational burden. We recommend these algorithms for incorporation into GWAS LMM implementations.

Figure 2.2: Accuracy results



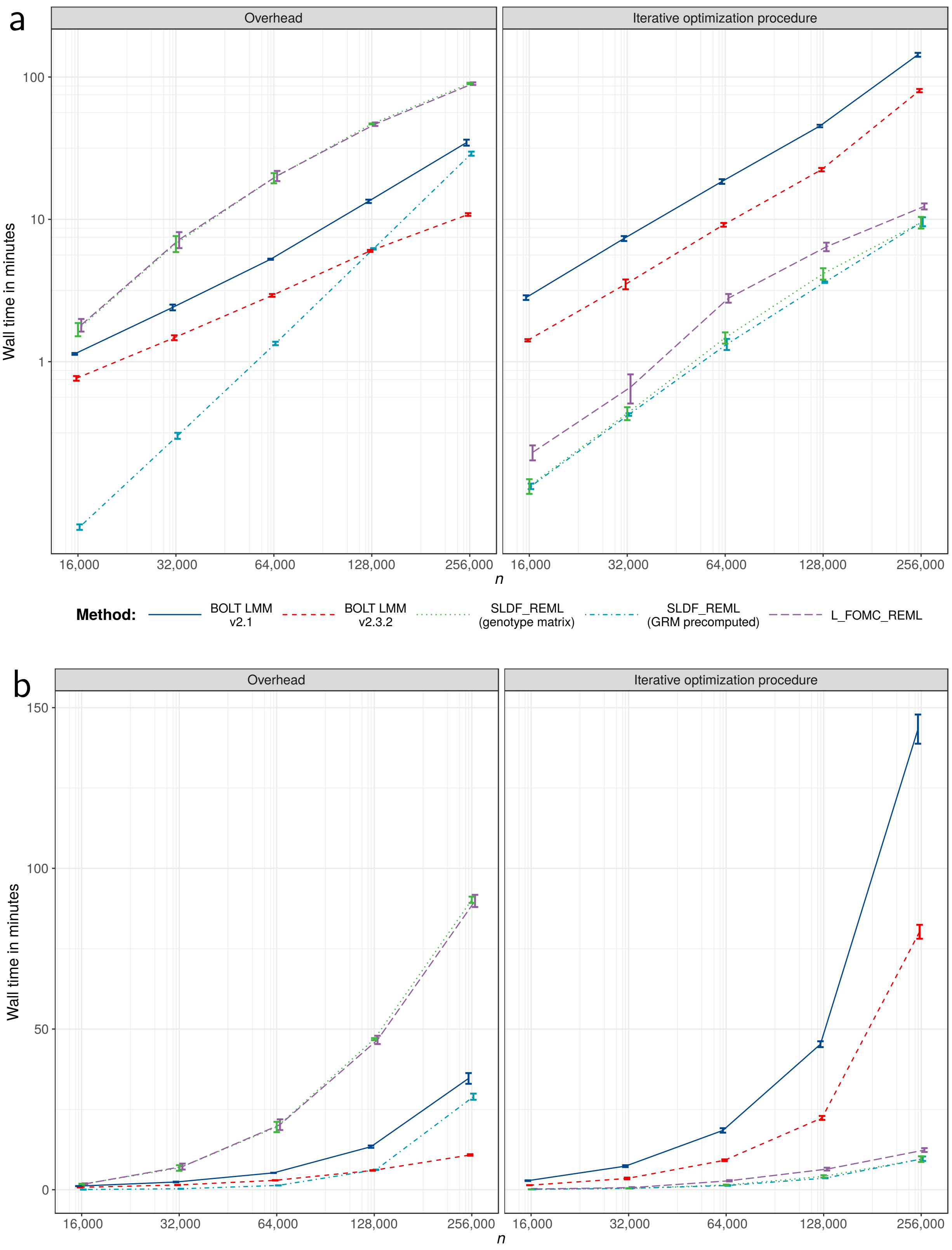
Comparison of heritability estimates for height generated by BOLT-LMM versions 2.1 and 2.3.2, `SLDF_REML`, and `L_FOMC_REML` versus those generated by the deterministic algorithm implemented in the GCTA software package [65], for varying sub-samples of 16,000 to 256,000 unrelated European-ancestry UK Biobank participants. Data are comprised of twenty independent replications per condition. Red dashed lines indicate standard errors of GCTA estimate. Points represent individual observations whereas boxes indicate the 95% confidence intervals for the trimmed mean estimate after a Bonferroni correction for 25 comparisons. The bias evidenced by the BOLT-LMM estimators is likely due to the combination of performing a small number of secant iterations with fixed start values and loose tolerances for determining convergence. *For n=256,000, memory requirements prohibited the use of GCTA, so we instead averaged ten estimates generated by the high-accuracy stochastic estimator implemented in BOLT-REML [85] (standard errors were 6.32e-5 and 2.45e-7 for the mean REML heritability estimate and its standard error, respectively).

Figure 2.3: Timing results



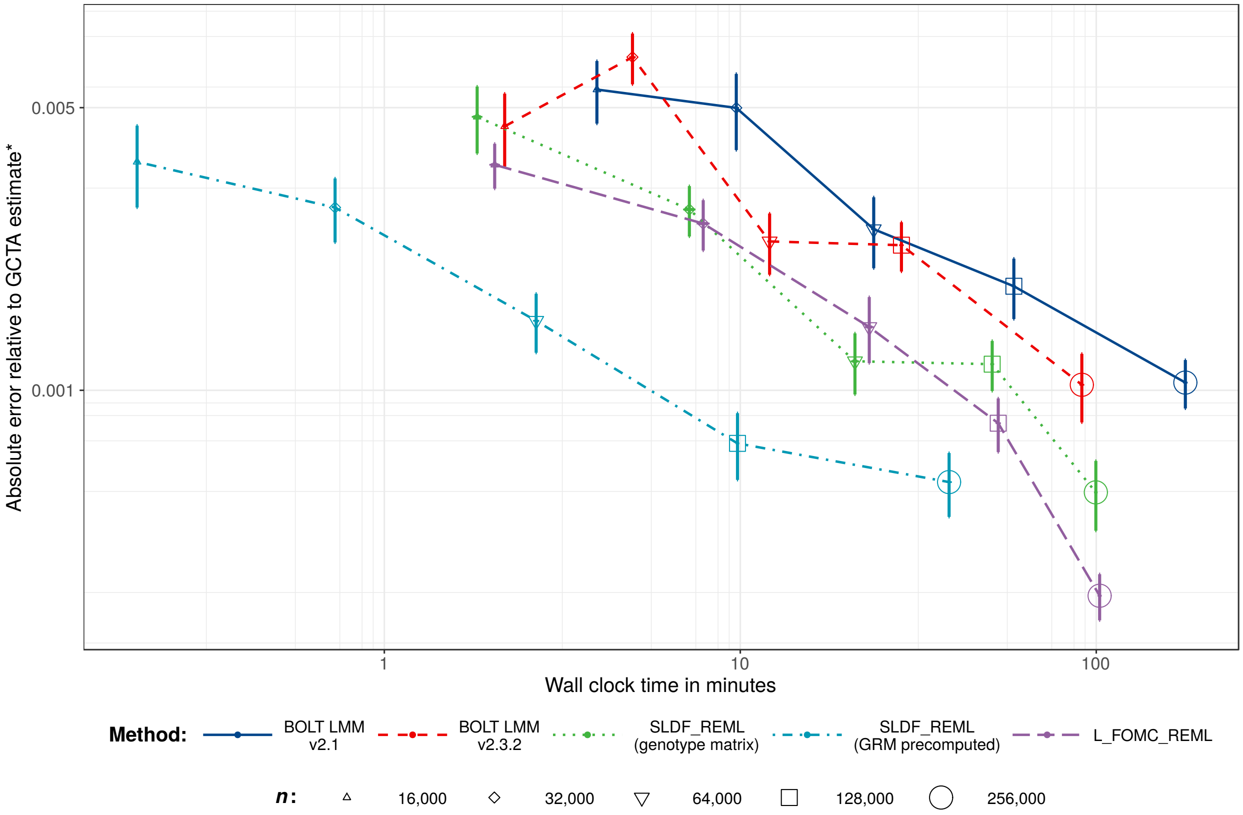
Trimmed mean wall clock time across twenty replications for per condition on the $\log_{10}$ scale (a) and natural scale (b). Error bars reflect per condition standard errors and lines connect per condition trimmed means.

Figure 2.4: Overhead versus iterative optimization procedure timing results



Trimmed mean wall clock time for overhead computations and iterative REML optimization procedures across twenty replications per condition on the $\log_{10}$ scale (a) and natural scale (b). Error bars reflect per condition standard errors and lines connect per condition means.

Figure 2.5: Numerical experiments: accuracy versus time



Average absolute error on the $\log_{10}$ scale with respect to the GCTA estimate* versus trimmed mean wall clock time across twenty replications per condition. Error bars reflect per condition standard errors and lines connect per condition trimmed means. *For n=256,000, memory requirements prohibited the use of GCTA, so we instead averaged ten estimates generated by the high-accuracy stochastic estimator implemented in BOLT-REML v2.3.2 [85] (standard errors were 6.32e-5 and 2.45e-7 for the mean heritability and its standard error, respectively).

Table 2.1: Time complexity of stochastic algorithms

Method		**Overhead**	**Objective function evaluation**
SLDF_REML	with precomputed GRM	$\mathcal{O}(n^2 \cdot (n_{\text{rand}} + c) \cdot n_{\kappa})$	$\mathcal{O}(n \cdot c \cdot n_{\kappa})$
	with genotype matrix	$\mathcal{O}(2m \cdot n \cdot (n_{\text{rand}} + c) \cdot n_{\kappa})$	$\mathcal{O}(n \cdot c \cdot n_{\kappa})$
L_FOMC_REML		$\mathcal{O}(4m \cdot n \cdot n_{\text{rand}} \cdot n_{\kappa})$	$\mathcal{O}(m \cdot n \cdot n_{\text{rand}})$
BOLT_LMM		$\mathcal{O}(n \cdot c^2 + m \cdot c)$	$\mathcal{O}(4m \cdot n \cdot n_{\text{rand}} \cdot n_{\kappa})$

n denotes the number of individuals, m the number of markers, and c the number of covariates. n_{rand} indicates the number of random probing vectors and is fixed at 15 in all numerical experiments. n_{κ} reflects the number of conjugate gradient iterations required to achieve convergence at a specified tolerance and can be bounded in terms of the spectral condition number of H_0. As noted in [69], implicit preconditioning of H_0 can be achieved by including the first few right singular vectors of the genotype matrix (or eigenvectors of the GRM) as covariates.

Table 2.2: Empirical timings

Method		Overhead	Per eval.	Eval. count	Total
BOLT-LMM v2.1		34.63	35.83	4	177.95
BOLT-LMM v2.3.2		10.82	20.07	4	91.09
L_FOMC_REML		89.87	2.06	6	102.21
SLDF_REML	with genotype matrix	90.22	1.06	9	99.73
	with precomputed GRM	28.95	1.07	9	38.60

Overhead and per objective function evaluation timings of stochastic algorithms for samples of 256,000 individuals. Data reflect trimmed mean wall clock time in minutes over 20 iterations per condition.

Chapter 3

Assortative mating and whole-genome heritability estimation

3.1 Introduction

3.1.1 Overview

Assortative mating (AM) refers to the phenomenon of “like choosing like”—mates selecting one another based on phenotypic similarity. As a result, mates’ phenotypes become positively correlated, which, for heritable traits, leads to a variety of consequences for the joint distribution of causal variants and the trait of interest. There is abundant empirical evidence that a variety of behavioral and non-behavioral phenotypes are subject to AM–e.g., height [92], political attitudes [93], and a variety of psychiatric traits [94]. Further, recent advances in methods for analyzing whole genome data have led to a resurgence of interest in characterizing AM via molecular genetic data [95, 96]. However, despite this growing interest, the effects of AM on a variety of commonly used methods for estimating heritability from genome-wide data remain unknown. Here, we present ongoing work that aims this close this gap by characterizing the behavior of three widely-used estimators—Haseman-Elston regression [97, 98], residual maximum likelihood [99, 100, 65], and linkage disequilibrium score regression [50, 101]—when applied to traits subject to AM.

Though a variety of AM models have been considered in the literature, we restrict our analysis to the earliest and one of the most popular models of AM (hereafter to referred to as the *phenotypic assortment model)* first introduced by Fisher a century ago [102] and further elaborated in the latter half of the twentieth century [103, 104, 105, 106, 107, 108]. The justification for this restriction is two-fold: first, Fisher’s model has formed the foundation for recent work in the quantification of AM using whole-genome data [95, 96, 109, 110], and second, the model is simple enough to be theoretically tractable yet general enough to permit extension to other models of interest (e.g., vertical transmission). In the remaining introductory sections, we introduce the phenotypic assortment model and the three aforementioned heritability estimators. In section §3.2, we present our primary theoretical and computational results thus far, which are justified in section §3.3 and section §3.4, respectively. Finally, in section §3.5 we briefly discuss limitations and future directions.

3.1.2 The phenotypic assortment model

3.1.2.1 A bird's eye view

Here we provide an overview of the phenotypic assortment model as elaborated by [103, 104, 105, 106, 107, 108], drawing heavily upon the exposition of [106]. We consider a heritable phenotype subject to assortment such that mates' phenotypes are correlated at $r \in (0, 1)$, with heritable component comprised solely of the additive effects of m unlinked single nucleotide polymorphism (SNP) variants. The population, which is assumed to be infinitely large, is initially subject to randomly mating at generation $t = 0$ and subject to a time-invariant spousal correlation throughout subsequent generations $t \geq 1$. Further, the non-heritable components of parent offspring phenotypes are assumed to be uncorrelated—i.e., there is no environmental transmission. Additionally, we make the following technical assumptions:

Primary phenotypic assortment Mates' genotypic values are conditionally independent given their phenotypic values.

Independence of variance components The heritable and non-heritable components of the phenotype are independent and follow Gaussian distributions with variances σ_g^2 and σ_e^2, respectively.

Exchangeable loci Haploid allele substitution effect is inversely proportional to minor allele frequency (MAF) such that a causal haploid locus with MAF $q = 1 - p$ taking values $\{-p, 1 - p\}$ has an allele substitution effect $\mathbb{E}[\eta] = \sigma_g / \sqrt{mpq}$.

Linearity of allelic effects The regression of haploid allele substitution effects onto the phenotype is linear.

Multivariate normality Parent-parent-offspring phenotypes are trivariate Gaussian.

Over successive generations, the correlation between mates' phenotypes induces linkage among all causal variants irrespective of their physical location. The additional covariance among previously independent causal variants increases the total genetic variance, which approaches a stable equilibrium as the within-individual and the cross-mate correlations between causal haploid variants approach one another. For large m (i.e., for a polygenic trait), the equilibrium genetic variance is approximately

$$\sigma^2_{g,\infty} \approx \frac{2(1-h_0^2)}{1-2h_0^2+\sqrt{1-4rh_0^2+4rh_0^4}}\sigma^2_{g,0}, \tag{3.1.1}$$

where $\sigma^2_{g,0}$ is the random mating genetic variance, $h_0^2 = \sigma^2_{g,0}/(\sigma^2_{g,0}+\sigma^2_e)$ is the random mating heritability, and r is the (time-invariant) phenotypic correlation between mates. Likewise, the equilibrium heritability is approximately

$$h^2_\infty \approx 2\left(1+\sqrt{1-4r(1-h_0^2)h_0^2}\right)^{-1} h_0^2. \tag{3.1.2}$$

Figure 3.1 illustrates equilibrium heritability for varying generation zero heritability and spousal correlation and figure 3.10 illustrates convergence to the equilibrium heritability across generations. Finally, the equilibrium correlation among individual haploid loci, a quantity we will refer to extensively, can be expressed as

$$\mu_\infty \approx \left(\frac{1-\sqrt{1-4h_0^2r(1-h_0^2)}}{(2m-1)\left(\sqrt{1-4h_0^2r(1-h_0^2)}+1-2h_0^2\right)}\right), \tag{3.1.3}$$

for polygenic traits under the exchangeable loci assumption. Note that in the above, m is the number of *diploid* causal loci. Derivations of the above quantities are presented in the following section.

3.1.2.2 Technical details

Consider a phenotype

$$y = \sum_{k=1}^{m} z_k\alpha_k + e$$

where at generation $t=0$, $e \sim \mathcal{N}(0,\sigma^2_e)$, $Var(Z\alpha)_0 = \sigma^2_{g,0}$, and each diploid locus $z_{k,0} = a^1_{k,0} + a^2_{k,0}$ is composed of two haploid loci $a^1_{k,0}, a^2_{k,0} \overset{i.i.d.}{\sim}$ bernoulli(q_k) independent of e. Denote the correlations

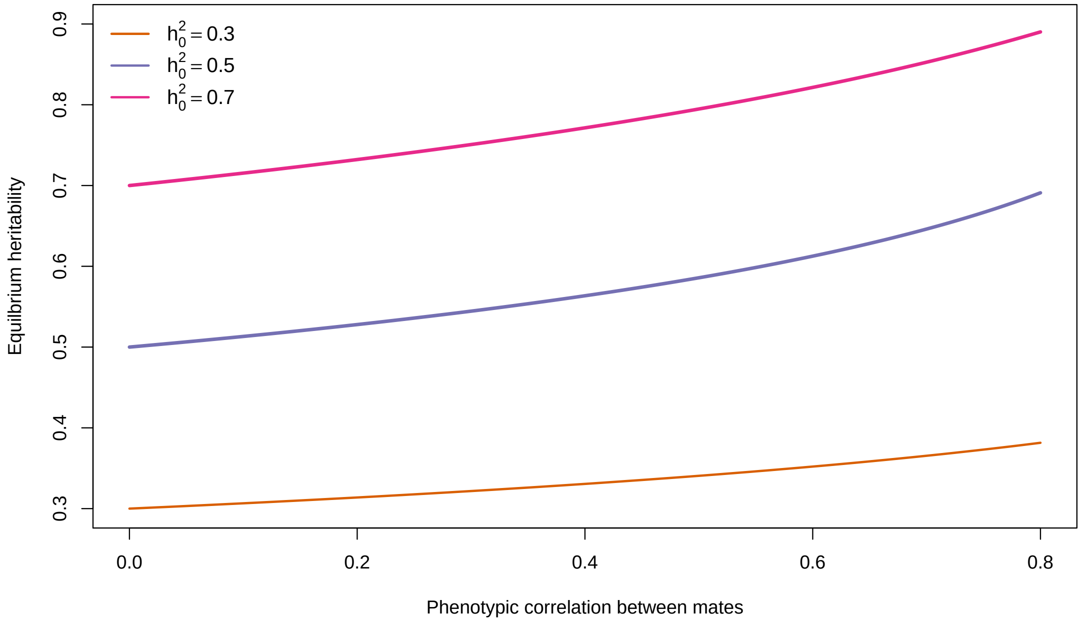


Figure 3.1: Equilibrium heritability as a function of generation zero heritability and spousal correlation

among haploid loci by

$$\begin{aligned}
\kappa_{kl,t} &= Corr(a^1_{k,t}, a^1_{l,t}) = Corr(a^2_{k,t}, a^2_{l,t}),\\
\ell_{kl,t} &= Corr(a^1_{k,t}, a^2_{l,t}) = Corr(a^2_{k,t}, a^1_{l,t}),\\
\mu_{kl,t} &= Corr(\grave{a}^1_{k,t}, \acute{a}^2_{l,t}) = Corr(\grave{a}^2_{k,t}, \acute{a}^1_{l,t}) = Corr(\grave{a}^1_{k,t}, \acute{a}^1_{l,t}) = Corr(\grave{a}^2_{k,t}, \acute{a}^2_{l,t}),
\end{aligned}$$

where $\grave{a}$ and $\acute{a}$ represent haploid loci across mating pairs. $\kappa_{kl,t}$ is then the generation t correlation between loci within gametes, $\ell_{kl,t}$ across gametes, and $\mu_{kl,t}$ across mates. Note that for all $t \in \mathbb{Z}^+$, $\kappa_{kk,t} = 1$ and that $\ell_{kk,t}$ is the coefficient of inbreeding. Further, at $t = 0$, we have $\{\kappa_{kl,0}\}_{kl} = I$ and $\{\ell_{kl,0}\}_{kl} = 0$.

Assume that phenotypic correlation among mates $r \in (0, 1)$ is time-invariant and denote the genetic correlation across mates as $r_{g,t} = rh^2_t$, where $h^2_t = \sigma^2_{g,0}/\sigma^2_{y,0}$ where $\sigma^2_{y,t} = \sigma^2_{g,t} + \sigma^2_e$. Define the standardized allele substitution effects as $\eta_k = \alpha_k\sqrt{q_k(1-q_k)}$ for $k = 1, \ldots, m$ and note that $\sigma^2_{g,0} = 2\eta^T\eta$. Finally assume that all loci are unlinked at $t = 0$ and that parent/offspring environments are independent.

Evolution of the system is determined by the following recurrences for $t \in \mathbb{N}$:

$$\begin{aligned}
\ell_{kl,t} &= \mu_{kl,t-1} \\
\kappa_{kl,t} &= \frac{1}{2}(\kappa_{kl,t-1} + \ell_{kl,t-1}) \llbracket k \neq l \rrbracket + \llbracket k = l \rrbracket, \\
\sigma^2_{g,t} &= 2 \sum_{k,l} \eta_k \eta_l (\kappa_{kl,t} + \ell_{kl,t}), \\
Cov(\acute{Z}\alpha, \grave{Z}\alpha)_t &= r_{g,t} \sigma^2_{g,t} \\
\implies \; 4 \sum_{k,l} \eta_k \eta_l \mu_{kl,t} &= 2r \sum_{k,l} \eta_k \eta_l (\kappa_{kl,t} + \ell_{kl,t}).
\end{aligned}$$

At equilibrium, we have

$$\begin{aligned}
\kappa_{kl,t} &= \ell_{kl,t} = \mu_{kl,t}, \quad \text{for } k \neq l, \\
\kappa_{kk} &= 1, \\
\ell_{kk,t} &= \mu_{kk,t}.
\end{aligned}$$

Thus, $Cov(\acute{Z}\alpha, \grave{Z}\alpha)_t = r_{g,\infty} \sigma^2_{g,t}$ implies

$$\begin{aligned}
4 \sum_{k,l} \eta_k \eta_l \mu_{kl,\infty} &= 2 r_{g,\infty} \sum_{k,l} \eta_k \eta_l (\kappa_{kl,\infty} + \mu_{kl,\infty}) \\
&= 2 r_{g,\infty} \eta^T \eta - 2 r_{g,\infty} \sum_k \eta_k^2 \mu_{kk,\infty} + 4 r_{g,\infty} \sum_{k,l} \eta_k \eta_l \mu_{kl,\infty}.
\end{aligned}$$

Further, the equilibrium genetic variance is

$$\begin{aligned}
\sigma^2_{g,\infty} &= 2 \sum_{k,l} \eta_k \eta_l (\kappa_{kl,\infty} + \ell_{kl,\infty}) \\
&= 4 \sum_{k,l} \eta_k \eta_l \mu_{kl,\infty} - 2 \sum_k \eta_k^2 \mu_{kk,\infty} + 2 \sum_k \eta_k^2.
\end{aligned}$$

Defining the *effective number of loci*

$$\tilde{m} = \frac{\sum_{k,l} \eta_k \eta_l \mu_{kl,\infty}}{\sum_k \eta_k^2 \mu_{kk,\infty}},$$

we then have

$$\begin{aligned}
Cov(\acute{Z}\alpha, \grave{Z}\alpha)_\infty &= 4\tilde{m}\sum_k \eta_k^2 \mu_{kk,\infty} \\
\implies \sigma^2_{g,\infty} &= Cov(\acute{Z}\alpha, \grave{Z}\alpha)_\infty - Cov(\acute{Z}\alpha, \grave{Z}\alpha)_\infty / 2\tilde{m} + \sigma^2_{g,0} \\
&= r_{g,\infty}\sigma^2_{g,\infty} - r_{g,\infty}\sigma^2_{g,\infty}/2\tilde{m} + \sigma^2_{g,0} \\
&= \left(\frac{1}{1 - r_{g,\infty}\mathcal{Q}}\right)\sigma^2_{g,0},
\end{aligned}$$

where we've defined $\mathcal{Q} = (1 - 1/2\tilde{m})$. Thus, we have for highly polygenic traits:

$$\lim_{\tilde{m}\to\infty} \sigma^2_{g,\infty} = \left(\frac{1}{1 - r_{g,\infty}}\right)\sigma^2_{g,0}.$$

Further, we have that

$$h^2_\infty = \left(\frac{1}{1 - r_{g,\infty}\mathcal{Q}}\right)\frac{\sigma^2_{g,0}}{\left(\frac{1}{1-r_{g,\infty}\mathcal{Q}}\right)\sigma^2_{g,0} + \sigma^2_e} = \frac{\sigma^2_{g,0}}{\sigma^2_{g,0} + \sigma^2_e(1 - r_{g,\infty}\mathcal{Q})} = \frac{h^2_0}{1 - r_{g,\infty}\mathcal{Q}(1 - h^2_0)}.$$

Because $r_{g,\infty} = rh^2_\infty$, the above then implies that

$$r_{g,\infty} = \frac{1 - \sqrt{1 - 4h^2_0\mathcal{Q}r(1 - h^2_0)}}{2\mathcal{Q}(1 - h^2_0)}.$$

Substituting this expression into those for the equilibrium genetic variance and the equilibrium heritability and taking the limit as $m \to \infty$ yields expressions (3.1.1) and (3.1.2).

Determining the equilibrium correlation between causal variants requires our linearity assumption, which states that the random variable $g_k = a^1_k\alpha_k$ corresponding to the haploid genic value at the k^{th} locus and the phenotype will have a linear relationship such that

$$\mathbb{E}[g_k - \alpha_k q_k | Y = y] = \zeta_{k,\infty} y \frac{\eta_k}{\sigma_{y,\infty}}, \tag{3.1.4}$$

where $\zeta_{k,\infty}$ is the correlation between the k^{th} haploid genic value and the phenotype. Considering a mating pair, we then have

$$\mathbb{E}[\grave{g}_l - \alpha_k q_l | \acute{Y} = y] = r\zeta_{l,\infty} y \frac{\eta_l}{\sigma_{y,\infty}}.$$

Following [106], the fact that mates' genic values are independent conditioning on either of the phenotypic values yields

$$\begin{aligned}
\mu_{kl,\infty}\eta_k\eta_l &= Cov(\acute{g}_k, \grave{g}_l) \\
&= \mathbb{E}[(\acute{g}_k - \alpha q_k)(\grave{g}_l - q_l)] \\
&= \mathbb{E}\left[\mathbb{E}[(\acute{g}_k - \alpha q_k)(\grave{g}_l - \alpha_l q_l)|\acute{Y} - \mathbb{E}[Y]\right] \\
&= \mathbb{E}\left[\mathbb{E}[(\acute{g}_k - \alpha q_k)|\acute{Y} - \mathbb{E}[Y]]\mathbb{E}[(\grave{g}_l - \alpha_l q_l)|\acute{Y} - \mathbb{E}[Y]]\right] \\
&= r\zeta_{k,\infty}\zeta_{l,\infty}\eta_k\eta_l \frac{\sigma^2_{y,\infty}}{\sigma^2_{y,\infty}} \\
\implies \mu_{kl,\infty} &= r\zeta_{k,\infty}\zeta_{l,\infty}.
\end{aligned}$$

We evaluate the correlation $\zeta_{k,\infty}$ as follows:

$$\begin{aligned}
\zeta_{k,\infty}\eta_k\sigma_{y,\infty} &= Cov(a^1_k\alpha_k, Y) \\
&= Cov(a^1_k\alpha_k, \sum_l a^1_l\alpha_l + a^2_l\alpha_l) \\
&= \sum_l (\mu_{kl,\infty} + \ell_{kl,\infty})\eta_k\eta_l, \\
&= \eta_k\left((1 - r\zeta^2_{k,\infty})\eta_k + 2\zeta_{k,\infty}\sum_l r\zeta_{l,\infty}\eta_l\right) \\
\implies \zeta_{k,\infty} &= (1 - r\zeta^2_{k,\infty})\frac{\eta_k}{\sigma_{y,\infty}} + 2\zeta_{k,\infty} r \sum_l \zeta_{l,\infty}\frac{\eta_l}{\sigma_{y,\infty}}.
\end{aligned}$$

As, at equilibrium, $\mathbb{E}[Z\alpha|Y=y] = h^2_\infty y$, we have that

$$\begin{aligned}
h^2_\infty y &= \mathbb{E}[Z\alpha|Y] \\
\implies h^2_\infty &= 2\sum_k \zeta_{k,\infty}\frac{\eta_k}{\sigma_{y,\infty}} \\
\implies r_{g,\infty} &= 2r\sum_k \zeta_{k,\infty}\frac{\eta_k}{\sigma_{y,\infty}} \\
\implies \zeta_{k,\infty} &= (1-r\zeta^2_{k,\infty})\frac{\eta_k}{\sigma_{y,\infty}} + \zeta_{k,\infty}r_{g,\infty} \\
\implies b_k &= r\zeta^2_{k,\infty}b_k + \zeta_{k,\infty}(1-r_{g,\infty}) \\
\implies \zeta_{k,\infty} &= (2b_kr)^{-1}\left(r_{g,\infty}-1+\sqrt{(1-r_{g,\infty})^2+4b_k^2r}\right),
\end{aligned}$$

where $b_k = \frac{\eta_k}{\sigma_{y,\infty}}$. Additionally, the assumption (3.1.4) implies that

$$h^2_\infty = 2\sum_k \zeta_{k,\infty}\frac{\eta_k}{\sigma_{y,\infty}}2\sum_k \zeta_{k,\infty}b_k$$

Recalling that $\mu_{kl,\infty} = r\zeta_{k,\infty}\zeta_{l,\infty}$ and that

$$r_{g,\infty} = \frac{1-\sqrt{1-4h_0^2\mathcal{Q}r(1-h_0^2)}}{2Q(1-h_0^2)},$$

we obtain the following equation for the correlation between the k^{th} haploid genic value and the phenotype:

$$\zeta_{k,\infty} = (1-r\zeta^2_{k,\infty})\eta_k/\sigma_{y,\infty} + r\zeta_{k,\infty}h^2_\infty.$$

Suppose $\eta_k = \sigma_{g,0}/\sqrt{2m}$ for all k. That is, that each locus contributes equal variance to the heritable component of y. This is the exchangeable loci assumption. We then have

$$\zeta_{k,\infty} = (2r\eta_k/\sigma_{y,\infty})^{-1}\left(r_{g,\infty}-1+\sqrt{(1-r_{g,\infty})^2+4r\eta_k^2/\sigma^2_{y,\infty}}\right), \tag{3.1.5}$$

which implies that $\zeta_{k,\infty} = \zeta_\infty$ for all k. Further, we have that $\tilde{m} = m$:

$$\tilde{m} = \frac{\left(\sum_k \eta_k\zeta_{k,\infty}\right)^2}{\sum_k \eta_k^2\zeta^2_{k,\infty}} = \frac{\left(\zeta_\infty m\sigma_{g,0}/\sqrt{2m}\right)^2}{m\zeta^2_\infty\sigma^2_{g,0}/2m} = m.$$

As a result,

$$\begin{aligned}\mu_\infty &= \frac{r_{g,\infty}}{\tilde{m}(1-r_{g,\infty})+r_{g,\infty}} \\ &\approx \left(\frac{1-\sqrt{1-4h_0^2r(1-h_0^2)}}{(2m-1)\left(\sqrt{1-4h_0^2r(1-h_0^2)}+1-2h_0^2\right)}\right),\end{aligned}$$

as in (3.1.3), noting that $\mathcal{Q} \approx 1$ for large m.

3.1.3 Heritability estimators

3.1.3.1 Haseman-Elston regression

Haseman-Elston (HE) regression [97, 98] is a computationally efficient method for estimating heritability using genome-wide data via ordinary least-squares. Let $y \in \mathbb{R}^n$ denote an observed vector of n individuals' phenotypes standardized to zero expectation and unit variance and let $Z \in \mathbb{R}^{n\times p}$ denote the individuals' standardized genotypes at p loci. Denote the lower triangular components of the phenotypic and genotypic sample covariance matrices as

$$\psi = \mathrm{vec}\,(\{yy^T\}_{i,j:i<j}), \qquad \kappa = \mathrm{vec}\,(\{p^{-1}ZZ^T\}_{i,j:i<j}).$$

The HE regression heritability is estimator is obtained by regressing ψ on to κ:

$$\hat{h}^2_{\mathrm{HE}} = \frac{\widehat{Cov}(\psi,\kappa)}{\widehat{Var}(\kappa)}.$$

Implementations of HE regression or variants thereof are provided by multiple genome-wide data analysis software packages [65, 111]. Theoretical justification for the HE regression estimator under random mating is included in the discussion of lemma 3.1 and 3.2.

3.1.3.2 Genomic relatedness restricted maximum likelihood (REML)

Consider a sample sample of n individuals measured at p SNP loci. We model the phenotype as a random vector with marginal distribution:

$$\tilde{y} \sim \mathcal{MVN}(X\beta, \frac{1}{p}ZZ^T\sigma_g^2 + I\sigma_e^2). \tag{3.1.6}$$

Z consists of the standardized SNP at p loci values for as sample of n individuals:

$$Z = \begin{pmatrix} \vdots & & \vdots \\ \frac{\tilde{z}_{1,i}-2q_1}{\sqrt{2q_1(1-q_1)}} & \cdots & \frac{\tilde{z}_{p,i}-2q_p}{\sqrt{2q_p(1-q_p)}} \\ \vdots & & \vdots \end{pmatrix}.$$

The covariates $X \in \mathbb{R}^{n \times c}$ (we can assume full row rank) and their effects $\beta \in \mathbb{R}^c$ are considered to be nuisance parameters.

The REML estimator is obtained by maximizing the log *residual likelihood* (REML criterion), which is simply the log likelihood of $K^T y$ where $K \in \mathbb{R}^{n \times n-c}$ is such that $K^T : \mathbb{R}^n \to (\text{col } X)^\perp$ and $K^T K = I_{n-c}$. Making the change of variables $\gamma = \sigma_g^2/\sigma_e^2$, denote

$$P_\gamma = V_\gamma^{-1} - V_\gamma^{-1} X (X^T V_\gamma^{-1} X)^{-1} X^T V_\gamma^{-1}, \tag{3.1.7}$$

where $V_\gamma = \frac{1}{p} Z Z^T \gamma + I$. The likelihood is maximized over γ by taking $\hat{\gamma}$ as the solution to the *REML equation*:

$$\frac{y^T P_\gamma Z Z^T P_\gamma y}{\text{tr}\left[P_\gamma Z Z^T\right]} = \frac{y^T P_\gamma P_\gamma y}{\text{tr}\left[P_\gamma\right]}, \tag{3.1.8}$$

and setting

$$\hat{\sigma}_e^2 = \frac{y^T P_{\hat{\gamma}} P_{\hat{\gamma}} y}{\text{tr}\left[P_{\hat{\gamma}}\right]}.$$

Above, the action of K^T occurs through

$$P_\gamma = V_\gamma^{-1} - V_\gamma^{-1} X \left(X^T V_\gamma^{-1} X\right)^{-1} X^T V_\gamma^{-1} = K \left(K^T V_\gamma K\right)^{-1} K^T.$$

The heritability estimate is then

$$\hat{h}^2_{\text{REML}} = \frac{\hat{\sigma}_e^2 \hat{\gamma}}{\hat{\sigma}_e^2 \hat{\gamma} + \hat{\sigma}_e^2} = \hat{\gamma}/(1+\hat{\gamma}).$$

The REML estimator doesn't admit a closed form representation and (3.1.8) must be solved numerically. An in depth discussion of the computational aspects of the REML method can be found in Chapter 2.

3.1.3.3 Linkage disequilibrium score regression

Linkage disequilibrium score regression (LDSC) [50] is a recent method for estimating heritability using summary statistics from genome wide association studies (GWAS) and information regarding patterns of local dependence (linkage-disequilibrium [LD]) between variants. Specifically, the *LD score* of a variant is defined as the sum of its squared correlations with all variants across the genome

$$\ell_j = \sum_{k=1}^{p} r_{jk}^2, \qquad j = 1, \ldots, p.$$

Denoting the GWAS test statistics for each variant by $\{\chi_j^2\}_{j=1}^p$, LDSC assumes a model wherein

$$\mathbb{E}[\chi_j^2] = np^{-1}\ell_j h^2 + 1.$$

Thus, multiplying the estimated slope from a linear regression of the observed χ^2 on to the LD score vector by $n^{-1}p$ provides an unbiased heritability estimate. In practice it is assumed that dependence among variants is spatially limited (i.e., variants that are "far away" from one another are assumed to be uncorrelated), and a local LD score

$$\tilde{\ell}_j = \sum_{k \in \mathcal{B}_{1\text{cM}}^{[j]}} \tilde{r}_{jk}^2,$$

is substituted for ℓ_j, where $\mathcal{B}_{1\text{cM}}^{[j]}$ denotes a the set of indices falling within one centimorgan of the j^{th} variant.

3.2 Results

3.2.1 Haseman-Elston regression

HE regression is biased when there is substantial dependence among causal variants. This is certainly true under the phenotypic assortment model but is also likely to arise under other forms of population stratification. As a result, we characterize this bias with respect to multiple circumstances: generally, as a function of the correlation among causal variants, and specifically, at equilibrium under the phenotypic assortment model assuming exchangeable loci. Proofs are provided in section 3.3.3.

Consider a phenotype y influenced by the additive effects of m casual variants with standardized allele substitution effects $\eta_1, \dots, \eta_m$. Further, suppose that $p = \omega^{-1}m$, $\omega \in (0, 1]$ variants are included in the genomic relatedness matrix, resulting in the consideration $p - m$ non causal variants such that, without loss of generality, $\eta_{m+1}, \dots, \eta_p \equiv 0$. Denote the population correlation matrix of individuals' genotypes by $\Upsilon \in \mathbb{R}^{p \times p}$ such that υ_{kl} indicates the correlation between haplotypes at the k^{th} and l^{th} loci.

Lemma 3.1. *Assuming non-causal variants are independent but allowing for arbitrary substitution effect sizes of causal variants, the HE regression heritability estimator has expectation*

$$\mathbb{E}[\hat{h}^2_{HE}] = \left(\frac{p\eta^T \Upsilon \Upsilon \eta}{tr[\Upsilon \Upsilon] \eta^T \Upsilon \eta} \right) h^2. \tag{3.2.1}$$

Corollary 3.2. *At equilibrium under the phenotypic assortment model assuming causal loci are exchangeable, the HE regression heritability estimator has expectation*

$$\mathbb{E}[\hat{h}^2_{HE}] = \left(\frac{\mu_\infty (2m - 1) + 1}{1 + 4\omega(m-1)\mu^2_\infty + 2\omega\mu_\infty(1 + \mu_\infty)} \right) h^2_\infty, \tag{3.2.2}$$

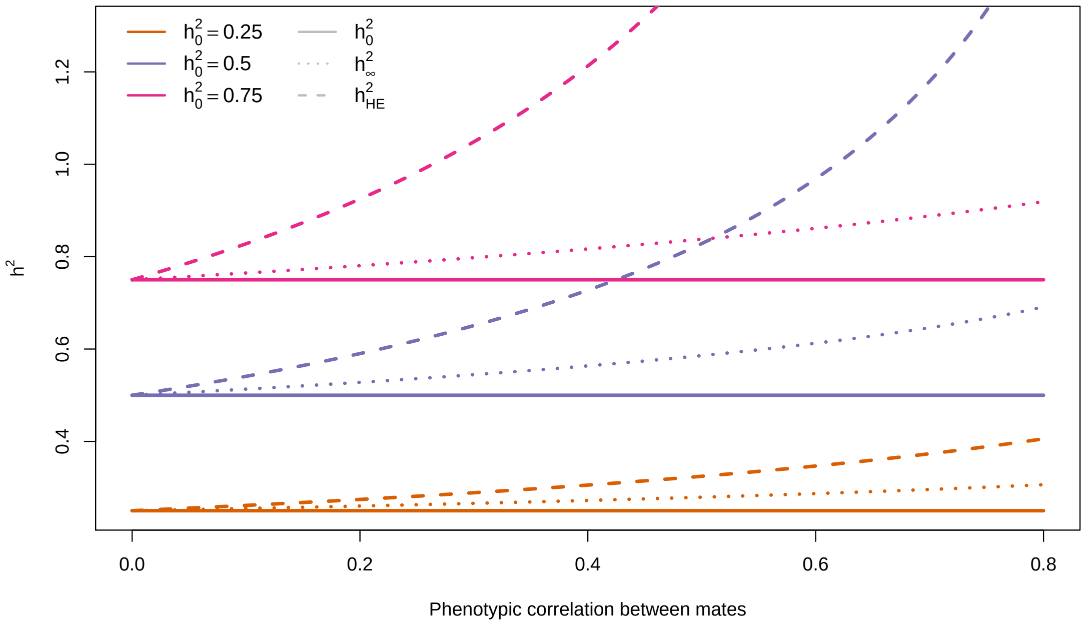
where h^2_∞ and μ_∞ are as in equations (3.1.2) and (3.1.3), respectively. Further, ω is asymptotically ignorable.

Note that for independent loci we simply have $\Upsilon = I_p$ and (3.2.1) reduces to $\mathbb{E}[\hat{h}^2_{\mathrm{HE}}] = h^2$; i.e, the HE regression is unbiased. However, when there is substantial dependence among causal variants consistent with the direction of their effects (i.e., $\operatorname{sgn} \eta_k \eta_l = \operatorname{sgn} \upsilon_{kl}$ for $k, l = 1, \dots, m$), HE regression will substantially overestimate the true heritability. Figure 3.2 demonstrates the substantial bias of the HE regression estimator applied to a polygenic trait under the phenotypic assortment model. Simulation results are congruent with closed form expressions above, as is demonstrated in figure 3.3.

3.2.2 Residual maximum likelihood

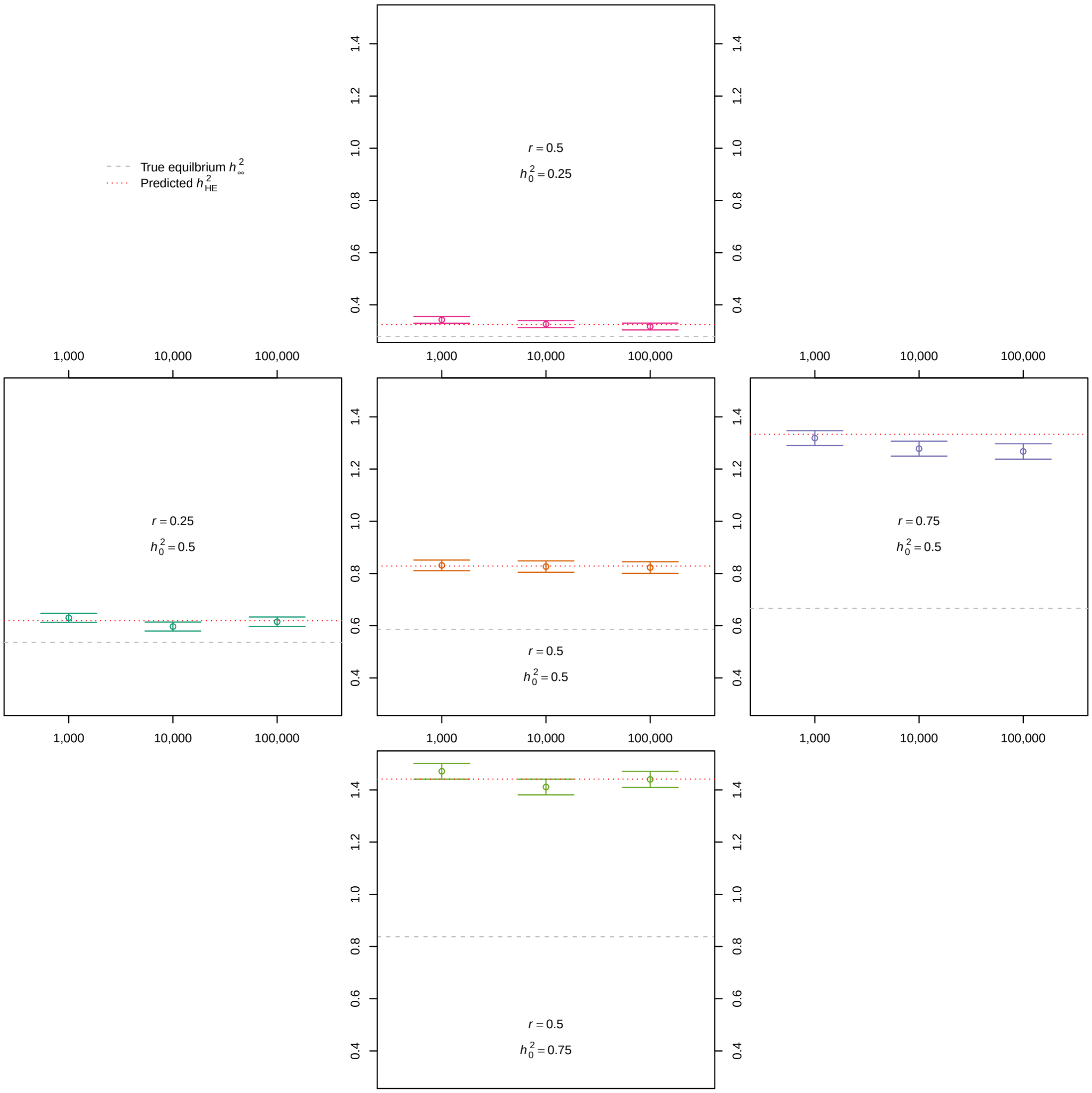
Empirical and theoretical work thus far suggest the REML heritability estimator is not consistent with respect to the equilibrium heritability h^2_∞ under the phenotypic assortment model.

Figure 3.2: Generation zero, equilibrium, and HE regression heritability estimates



Generation zero, equilibrium, and HE regression estimates of heritability for a highly polygenic trait, under varying initial conditions and spousal correlations.

Figure 3.3: Empirical HE regression estimates versus theoretical predictions



Results from simulation studies match closed form expressions for bias of HE regression under the phenotypic assortment model for varying h_0^2, r. Additional simulations are on going.

However, it does appear to be asymptotically unbiased with respect to the generation zero heritability h_0^2, though convergence to the generation zero heritability is slow. Heuristically, this hypothesis can be summarized as follows: If variants are independent of one another, the REML method provides a consistent estimator of the heritability of a phenotype generated under the model described by (3.1.6) [112, 113]. However, assortative mating induces long-range correlations between causal variants, thereby increasing the true genetic variance of the phenotype and simultaneously breaking the conditions required to ensure consistency using existing theory. Still, as the sample size n, number of causal variants m, and total number of measured variants p become large, the REML estimator behaves as if the causal variants were independent and converges to the generation zero heritability h_0^2. We summarize our progress so far with the following lemma characterizing the limiting spectral distribution of the genomic relatedness matrix under the phenotypic assortment model and formalize the above hypothesis in conjecture 3.4.

Lemma 3.3. *Let $S = p^{-1}ZZ^T$ denote the sample relatedness matrix derived from n individuals measured genotypes at p loci, including $m = \omega p$ causal variants, for a constant $\omega \in (0, 1)$. Under the phenotypic assortment model, as $n, p \to \infty$ such that $n/p \to \tau > 0$, the empirical spectral distribution of S converges almost surely to the Marčenko-Pastur law:*

$$F_n^S \overset{a.s.}{\to} \varphi_\tau(x) = \frac{1}{2\pi\tau x}\sqrt{(b_\tau - x)(x - a_\tau)}\,[\![x \in [a_\tau, b_\tau]]\!]\,.$$

Conjecture 3.4. *Under the above conditions, we have that*

1. $\hat{\gamma}_{REML} \overset{p}{\to} \omega\gamma$, *where* $\gamma = \sigma_{g,0}^2/\sigma_e^2$,
2. $\hat{\sigma}_{e,REML}^2 = y^T P_{\hat{\gamma}} P_{\hat{\gamma}} y / tr[P_{\hat{\gamma}}] \overset{p}{\to} \sigma_e^2$,

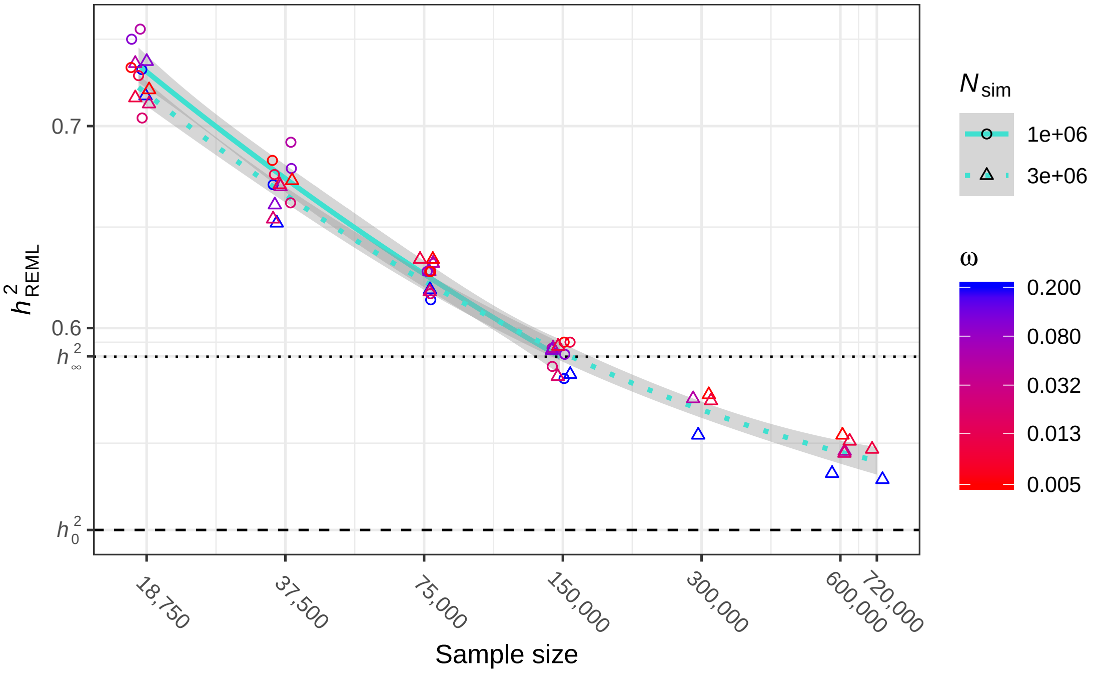
which together imply that $\hat{h}_{REML}^2 \overset{p}{\to} h_0^2$. I.e., the REML heritability estimator is a consistent estimator of the generation zero heritability.

Simulations results so far are consistent with Conjecture 3.4 and the rate of convergence does not not appear to depend on $\omega = m/p$ or $\tau = n/p$ (figure 3.4, figure 3.5, figure 3.6).

3.2.3 Linkage disequilibrium score regression

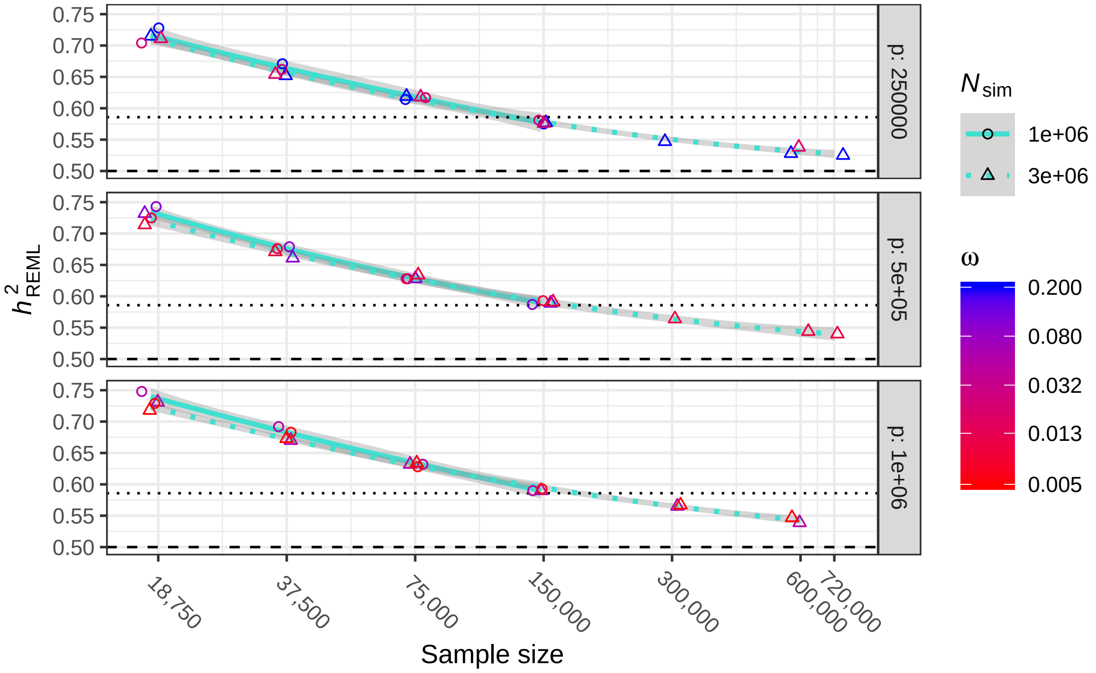
We are currently developing a theoretical characterization of LD score regression under the phenotypic assortment model (see 3.3.6), but we anticipate that both the LD score regression and

Figure 3.4: Empirical REML heritability estimates, aggregated



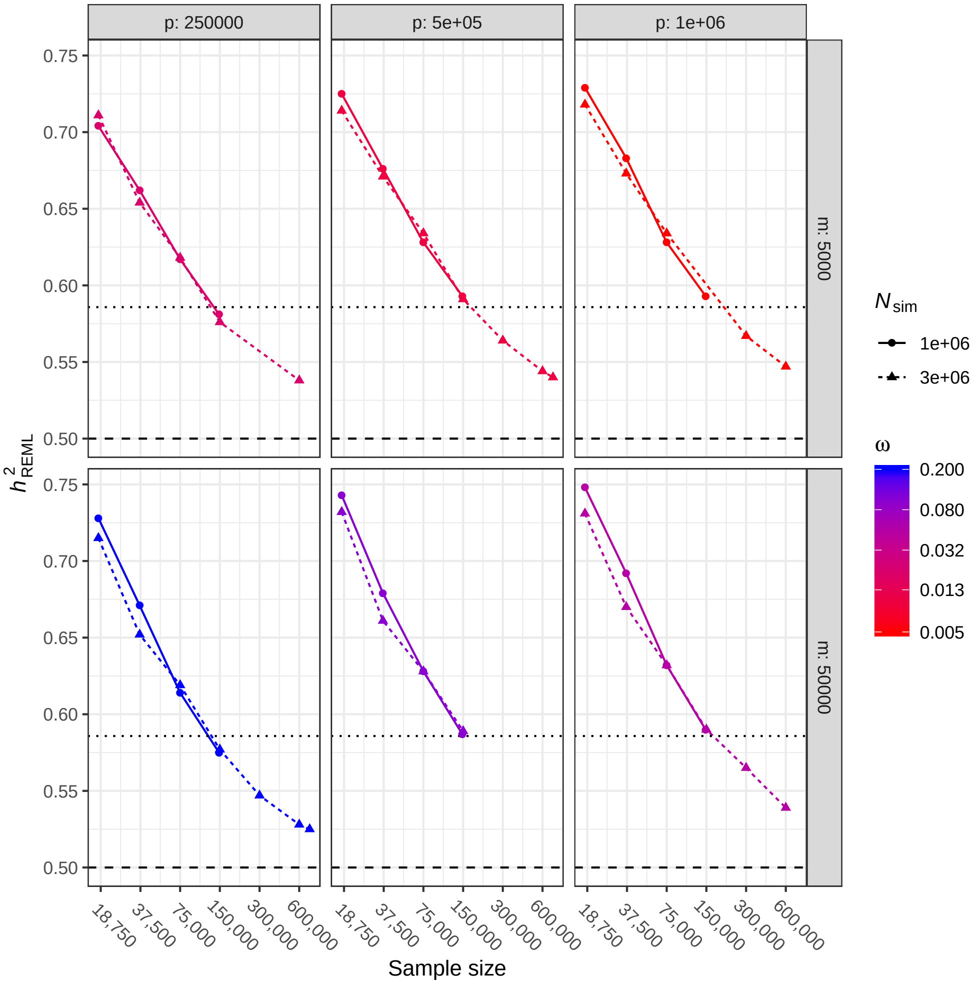
REML heritability estimates in data simulated under the phenotypic assortment model with $r = .5$, $h_0^2 = .5$ at equilibrium for varying m, p, in unrelated samples drawn from populations of sizes $N_{\mathrm{sim}} \in \{1\mathrm{e}{+}6, 3\mathrm{e}{+}6\}$. Rate of convergence does not appear to vary with $\omega = m/p$. The black dashed-line indicates the true equilibrium heritability and the black dotted line indicates the generation zero heritability. Additional simulations are on going.

Figure 3.5: Empirical REML heritability estimates split by p



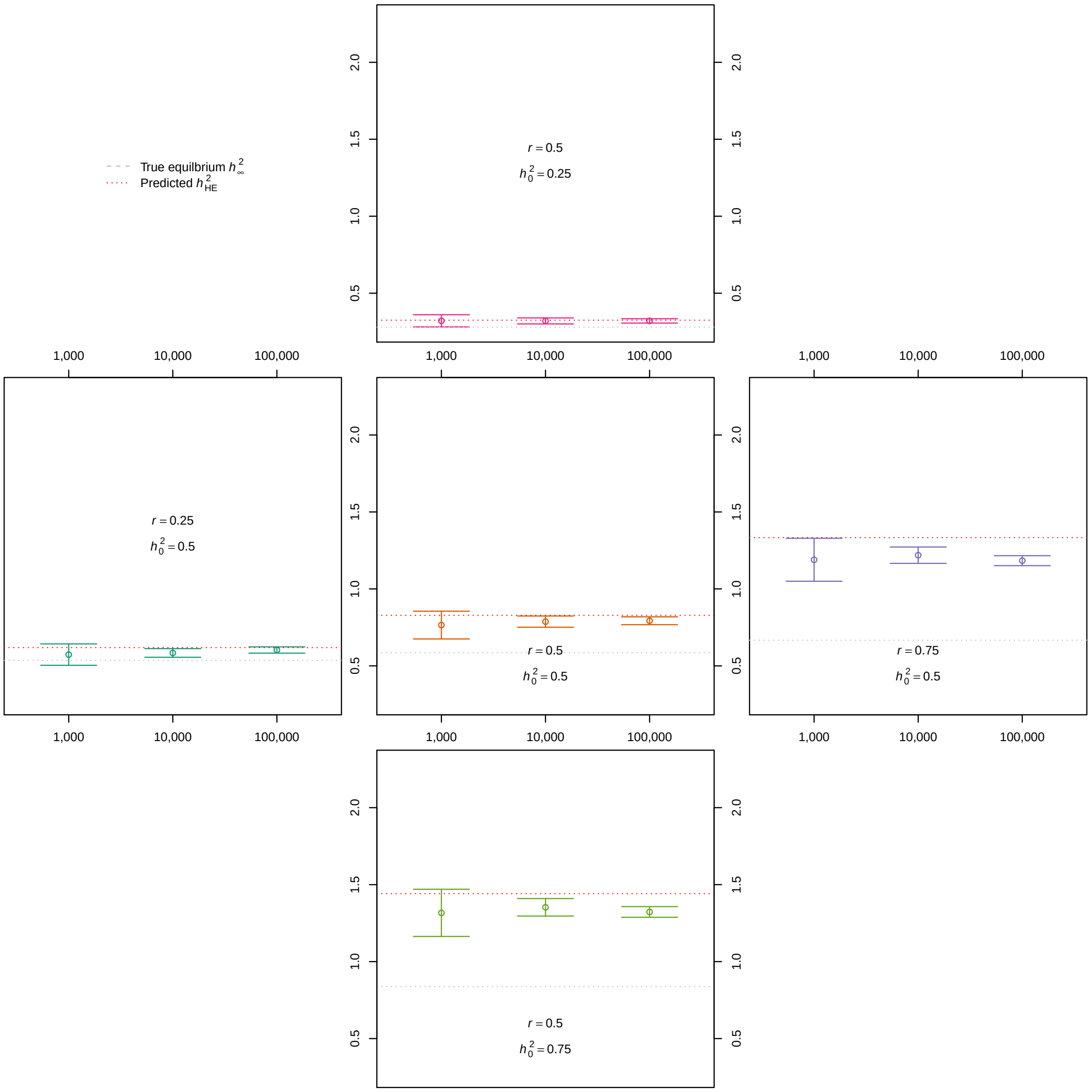
REML heritability estimates in data simulated under the phenotypic assortment model with $r = .5$, $h_0^2 = .5$ at equilibrium for varying m, p, in unrelated samples drawn from populations of sizes $N_{\text{sim}} \in \{1\text{e}{+}6, 3\text{e}{+}6\}$. Rate of convergence does not appear to vary with $\omega = m/p$. The black dashed-line indicates the true equilibrium heritability and the black dotted line indicates the generation zero heritability. Additional simulations are on going.

Figure 3.6: Empirical REML heritability estimates split by $m \times p$



REML heritability estimates in data simulated under the phenotypic assortment model with $r = .5$, $h_0^2 = .5$ at equilibrium for varying m, p, in unrelated samples drawn from populations of sizes $N_{\mathrm{sim}} \in \{1\text{e+}6, 3\text{e+}6\}$. Rate of convergence does not appear to vary with $\omega = m/p$. The black dashed-line indicates the true equilibrium heritability and the black dotted line indicates the generation zero heritability. Additional simulations are on going.

Figure 3.7: Empirical LDSC regression heritability estimates



LDSC regression heritability estimates in equilibrium in data simulated under the phenotypic assortment model at equilibrium for varying m, r, and h^2_∞. Estimates are biased upward, but not to the extent of HE regression estimates. Additional simulations are on going.

intercept will exhibit upward biases bounded above by the expected HE regression estimate established in (3.2.1), and that, like HE regression, these biases are independent of sample size. Initial simulation results support these hypotheses: LD score regression heritability estimates appear to be attenuated relative to HE regression estimates (but still upwardly biased; Figure 3.7) and LD score intercepts demonstrate inflation increasing the strength of assortment.

3.3 Theoretical development

In the following section we motivate previous claims, additionally presenting requisite general results regarding the equilibrium properties of haploid loci under the phenotypic assortment model as necessary.

3.3.1 Correlations among unlinked, exchangeable haploid loci

Consider an individual's standardized diploid genotype[1] Z as a random vector with covariance matrix Υ_t. Suppose we have measured $p \in \mathbb{N}$ genotypes, $m \in [1, p]$ of which are causal. Without loss of generality, order the genotypes such that the first m are causal and the remaining $p - m$ are not causal. Under random mating, we simply have $\Upsilon_0 = I_p$. Under assortment however, assuming exchangeable loci, the equilibrium population covariance matrix is

$$\Upsilon_\infty(m) = \left[\begin{array}{cccc|c} 1+\mu(m) & & & \text{Sym.} & \\ 2\mu(m) & \ddots & & & \text{Sym.} \\ \vdots & \ddots & \ddots & & \\ 2\mu(m) & \cdots & 2\mu(m) & 1+\mu(m) & \\ \hline \multicolumn{4}{c|}{0_{(p-m)\times m}} & I_{p-m} \end{array}\right],$$

where $\mu(m) = \mu_\infty$ for m causal variants given initial conditions $r, h_0^2 \in (0, 1]$. Denote entries of $\Upsilon(m) = \Upsilon_\infty(m)$ by $\upsilon_{ij}(m)$.

Lemma 3.5. *Under the above assumptions, as $m \to \infty$, we have that each $\upsilon_{ij}(m) \to \delta_{ij}$ where δ_{ij} is the Kronecker delta. Further, this convergence is such that the equilibrium correlation among haploid loci has order $\mu(m) \sim \mathcal{O}(m^{-1})$.*

[1] Z is "standardized" in that each $z_k = (\tilde{z}_k - 2q_k)/\sqrt{2p_k q_{kl}}$ where $\tilde{z}_k$ Under random mating, we have $Var(z_k) = 1$; under assortment, however, we have $Var(z_{k,t}) = (1 + \mu_{kk,t})^2 > 1$.

Proof. Note that $\mu(m)$, as a function of three variables,

$$\mu(m, h_0^2, r) = \frac{1 - \sqrt{1 - 4h_0^2 r(1 - h_0^2)}}{(2m - 1)\left(\sqrt{1 - 4h_0^2 r(1 - h_0^2)} + 1 - 2h_0^2\right)}$$

is continuous and monotone increasing with respect to $h_0^2, r \in (0, 1)^2$ (i.e., the equilibrium correlation is largest when both heritability and the correlation among mates are unity. Thus, $\mu(m) \sim \mathcal{O}(m^{-1})$ but not $o(m)^{-1}$, as demonstrated in figure 3.8. □

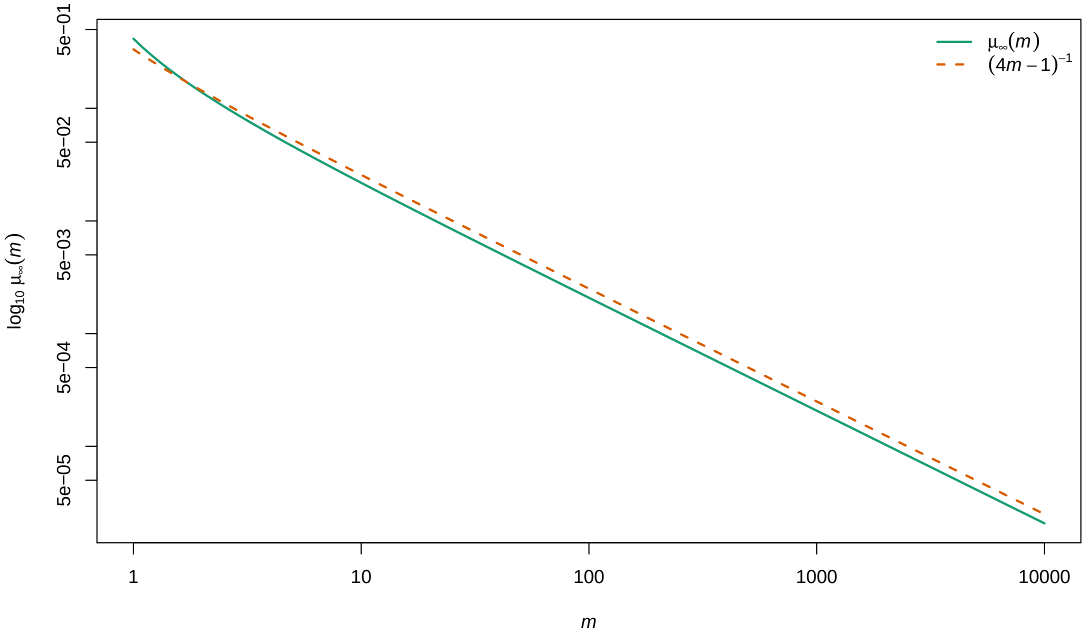


Figure 3.8: Equilibrium correlation among haploid variants for $r = 0.5$, $h_0^2 = 0.5$.

3.3.2 Properties of the genomic relatedness matrix

Here we develop the statistical properties of the genetic relatedness matrix under arbitrary patterns of inter-correlations among haploid loci. Denoting diploid genotypic values standardized according to the population allele frequency by $\{z_k\}_{k=1}^p$, we define the sample genomic relatedness matrix (GRM), or kinship matrix, as $K = \frac{1}{p}ZZ^T$. Additionally, for a standardized phenotype $y = \sigma_{\tilde{y}}^{-1}(\tilde{y} - \overline{\tilde{y}})$ where $\tilde{y} = (2p(1-p))^{-1/2}Z\eta$, denote the phenotypic outer product matrix $\Psi_{ij} = yy^T$.

Lemma 3.6. *The variance of individual entries of the genomic related matrix is expressed*

$$Var[vec(K_{ij,i<j})]\frac{1}{p^2}tr[\Upsilon\Upsilon].$$

Additionally, the covariance of the lower triangular elements of the phenotypic outer product and genomic relatedness matrices is expressed:

$$Cov(vec(\Psi_{ij,i<j}), vec(K_{ij,i<j})) = \frac{2}{p\sigma^2_{\tilde{y}}}\eta^T\Upsilon\Upsilon\eta, \quad i<j.$$

Proof. Let ℓ_{kl} denote the population correlation between the the k^{th} and l^{th} haploid loci. The first result of several straightforward computations:

$$\begin{aligned}
\mathbb{E}[z_{ik}z_{jl}] &= \mathbb{E}[z_{ik}z_{jk}] = 0,\\
Cov(z_{ik}, z_{jl}) &= Cov(z_{ik}, z_{jk}) = 0,\\
Var(z_{il}z_{jl}) &= Var(z_{il})Var(z_{jl}) = (1+\ell_{kk})^2,\\
\mathbb{E}[z_{ik}z_{ik}] &= Var(z_{ik}) + \mathbb{E}[z_{ik}]\mathbb{E}[z_{ik}] = (1+\ell_{kk})\\
\mathbb{E}[z_{ik}z_{il\neq k}] &= Cov(z_{ik}, z_{il}) + \mathbb{E}[z_{ik}]\mathbb{E}[z_{il}] = 2\ell_{kl}\\
Cov(z_{il}z_{jl}, z_{ik\neq l}z_{jk\neq l}) &= \mathbb{E}[z_{il}z_{jl}z_{ik}z_{jk}] - \mathbb{E}[z_{il}z_{jl}]\mathbb{E}[z_{ik}z_{jk}] = \mathbb{E}[z_{il}z_{ik}]\mathbb{E}[z_{jl}z_{jk}] = 4\ell^2_{kl},\\
\mathbb{E}[K_{ij}] &= 0,\\
Var[K_{ij}] &= \frac{1}{p^2}Var\left(\sum_{l=1} z_{il}z_{jl}\right) = \frac{1}{p^2}\left(\sum_{l=1}^{p}(1+\ell_{kk})^2 + 4\sum_{k\neq l}^{p}\ell^2_{kl}\right) = \frac{1}{p^2}\text{tr}\,[\Upsilon\Upsilon].
\end{aligned}$$

Noting that

$$\begin{aligned}
\sigma^2_g &= 2\sum_{k}^{m}\eta^2_k(1+\ell_{kk}) + 4\sum_{k\neq l}\eta_k\eta_l\ell_{kl} = 2\eta^T\Upsilon\eta,\\
\tilde{y} &= Z\eta + e\\
\sigma^2_{\tilde{y}} &= 2\eta^T\Upsilon\eta + \sigma^2_e
\end{aligned}$$

Now, note that $\mathbb{E}[\Psi_{ij}] = 0$ and, as $\tilde{e}_i$ and $\tilde{e}_j$ are independent of one another and of all genotypes, the covariance of the lower triangular elements of the phenotypic and genomic relatedness matrices is computed as follows:

$$\begin{aligned}
Cov(\Psi_{ij}, K_{ij})_{i<j} =& \mathbb{E}[\Psi_{ij}K_{ij}] - \mathbb{E}[\Psi_{ij}]\mathbb{E}[K_{ij}] \\
=& \frac{1}{p\sigma^2_{\tilde{y}}}\mathbb{E}\left[\sum_k \left(\eta_k z_{ik}\sqrt{2}\right)\sum_l \left(\eta_l z_{jl}\sqrt{2}\right)\sum_t \left(z_{it}z_{jt}\right)\right] \\
=& \frac{2}{p\sigma^2_{\tilde{y}}}\mathbb{E}\left[\sum_{k,l,t} \eta_k\eta_l z_{ik}z_{jl}z_{it}z_{jt}\right] \\
=& \frac{2}{p\sigma^2_{\tilde{y}}}\eta^T\Upsilon\Upsilon\eta.
\end{aligned}$$

□

3.3.3 Bias of Haseman-Elston regression

The result stated in lemma 3.1 follows immediately from the results in the preceding section as the expected slope from the linear regression of $\mathrm{vec}\left(\Psi_{ij,i<j}\right)$ regressed on $\mathrm{vec}\left(K_{ij,i<j}\right)$ will be $Cov(\Psi_{ij}, K_{ij})/Var(K_{ij})$. Noting that $\sigma^2_g = 2\eta^T\Upsilon\eta$ we then have that

$$\frac{Cov(\Psi_{ij}, K_{ij})}{Var(K_{ij})} = \frac{2p\eta^T\Upsilon\Upsilon\eta}{\sigma^2_{\tilde{y}}\mathrm{tr}\left[\Upsilon\Upsilon\right]}\cdot\frac{2\eta^T\Upsilon\eta}{2\eta^T\Upsilon\eta} = \frac{p\eta^T\Upsilon\Upsilon\eta}{\eta^T\Upsilon\eta\mathrm{tr}\left[\Upsilon\Upsilon\right]}h^2.$$

Specialization to the equilibrium case of exchangeable loci under the phenotypic assortment model, as in (3.2), follows from the following computations. Denote the sets

$$\begin{aligned}
\Omega_{k=t\neq l}(m) &= \{k, t, l \in \{1, \ldots, m\}^3 : k = t \neq l\}, \\
\Omega_{k=l\neq t}(m) &= \{k, t, l \in \{1, \ldots, m\}^3 : k = l \neq t\},
\end{aligned}$$

and so forth, and observe that the cardinalities of the mutually exclusive sets below are computed

$$\begin{aligned}
|\Omega_{k=t\neq l}(m) \cup \Omega_{k\neq l=t}(m)| =& 2m(m-1), \\
\Omega_{k=l\neq t}(m) =& m(m-1) \\
|\Omega_{k\neq l\neq t\neq k}(m)| =& m(m-1)(m-2), \\
|\Omega_{k=l=t}(m)| =& m,
\end{aligned}$$

as can be shown via induction. Thus, we compute the following quantities:

$$
\begin{aligned}
\eta^T \Upsilon\Upsilon\eta &= \sum_{k,l,t} \eta_k \eta_l \upsilon_{kt}\upsilon_{lt} \\
&= \sigma^2_{g,0}/2m \left(\sum_{\substack{k,l,t \\ k=t\neq l}} \upsilon_{kk}\upsilon_{lk} + \sum_{\substack{k,l,t \\ k\neq l=t}} \upsilon_{kl}\upsilon_{ll} + \sum_{\substack{k,l,t \\ k=l\neq t}} \upsilon_{kt}\upsilon_{kt} + \sum_{\substack{k,l,t \\ k\neq l\neq t\neq k}} \upsilon_{kt}\upsilon_{lt} + \sum_{\substack{k,l,t \\ k=l=t}} \upsilon_{kk}\upsilon_{kk} \right) \\
&= \sigma^2_{g,0}/2m \left(\sum_{\substack{k,l,t \\ k=t\neq l}} \frac{2\mu}{1+\mu} + \sum_{\substack{k,l,t \\ k\neq l=t}} \frac{2\mu}{1+\mu} + \sum_{\substack{k,l,t \\ k=l\neq t}} \left(\frac{2\mu}{1+\mu}\right)^2 + \sum_{\substack{k,l,t \\ k\neq l\neq t\neq k}} \left(\frac{2\mu}{1+\mu}\right)^2 + \sum_{\substack{k,l,t \\ k=l=t}} 1 \right) \\
&= \sigma^2_{g,0}/2m \left(\sum_{\substack{k,l,t \\ k=t\neq l}} \frac{2\mu}{1+\mu} + \sum_{\substack{k,l,t \\ k\neq l=t}} \frac{2\mu}{1+\mu} + \sum_{\substack{k,l,t \\ k=l\neq t}} \left(\frac{2\mu}{1+\mu}\right)^2 + \sum_{\substack{k,l,t \\ k\neq l\neq t\neq k}} \left(\frac{2\mu}{1+\mu}\right)^2 + \sum_{\substack{k,l,t \\ k=l=t}} 1 \right) \\
&= \sigma^2_{g,0}/2 \left(2(m-1)\left(\frac{2\mu}{1+\mu}\right) + (m-1)\left(\frac{2\mu}{1+\mu}\right)^2 + (m-1)(m-2)\left(\frac{2\mu}{1+\mu}\right)^2 + 1 \right) \\
&= \sigma^2_{g,0}/2 \left(2(m-1)\left(\frac{2\mu}{1+\mu}\right) + (m-1)^2\left(\frac{2\mu}{1+\mu}\right)^2 + 1 \right) \\
&= \sigma^2_{g,0}/2 \left(1 + (m-1)\left(\frac{2\mu}{1+\mu}\right)\right)^2 \\
\eta^T \Upsilon\eta &= (\sigma^2_{g,0}/2m) \sum_{k,l} \upsilon_{kl} \\
&= (\sigma^2_{g,0}/2) \left(1 + (m-1)\left(\frac{2\mu}{1+\mu}\right)\right),
\end{aligned}
$$

The trace term is computed

$$
\begin{aligned}
\operatorname{tr}\left[\Upsilon\Upsilon\right] &= \operatorname{tr}\left[\begin{pmatrix} \{(1)^{\delta_{ij}}(2\mu_\infty(1-\mu_\infty))^{1-\delta_{ij}}\}^m_{ij} & 0 \\ 0 & I_{p-m} \end{pmatrix}^2 \right] \\
&= (p-m) + \sum_k \sum_l \upsilon_{lk}\upsilon_{kl} \\
&= p + m(m-1)\frac{4\mu^2_\infty}{(1-\mu_\infty)^2},
\end{aligned}
$$

which allows us to compute the expected HE regression slope:

$$\begin{aligned}\mathbb{E}[\hat{h}^2_{\text{HE}}] &= \left(\frac{p\eta^T\Upsilon\Upsilon\eta}{\text{tr}\,[\Upsilon\Upsilon]\eta^T\Upsilon\eta}\right)h^2_\infty \\ &= \left(\frac{\mu_\infty(2m-1)+1}{1+4\omega(m-1)\mu^2_\infty+2\omega\mu_\infty(1+\mu_\infty)}\right)h^2_\infty.\end{aligned}$$

Further, taking the limit as $m \to \infty$ yields the approximation

$$\mathbb{E}[\hat{h}^2_{\text{HE}}] \approx \left(\frac{(\mathcal{V}-2h+2)\,(\mathcal{V}-2h+2)}{4h^2(r+1)-4h\,(3\mathcal{V}+r+2)-6\mathcal{V}+4}\right)h^2_\infty,$$

where $\mathcal{V}$ does not depend on ω or p. What's notable here is the proportion of causal variants $\omega = m/p$ is asymptotically irrelevant, as is demonstrated in Figure 3.9.

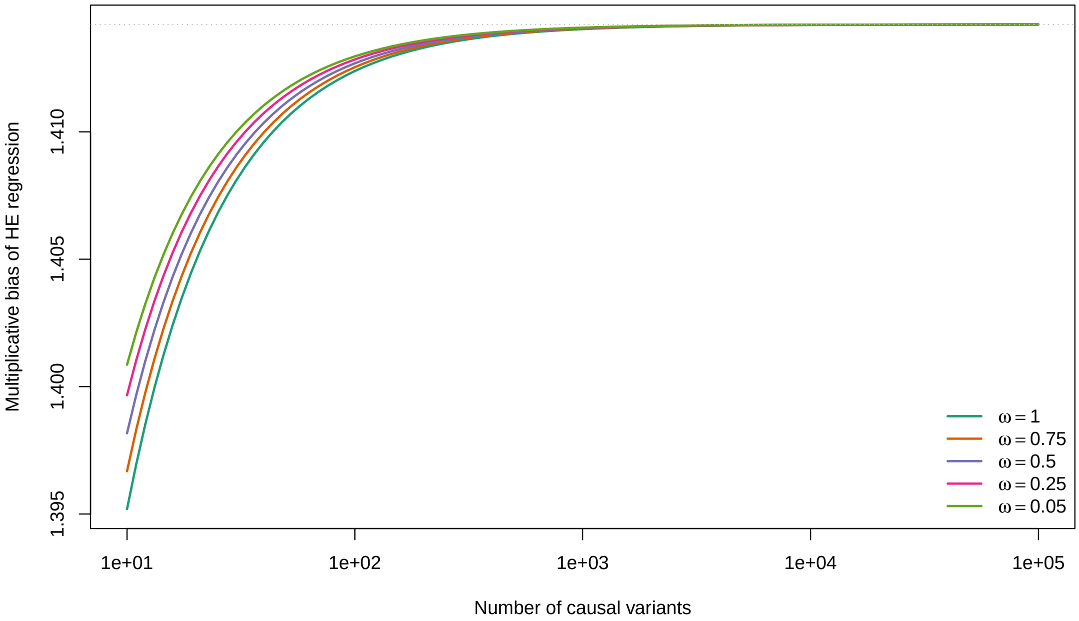


Figure 3.9: Bias of HE regression relative to the true population heritability at $r = 0.5$, $h^2_0 = 0.5$ as a function of the number of causal variants, m, for varying proportions of causal variants $\omega = m/p$.

3.3.4 Higher order moments of unlinked, exchangeable haploid loci

Our results regarding the behavior of the REML estimator rely on the following characterization of the higher order moments of haploid loci. Given the complexity of this topic, the notation in this section is modified and self-contained. Additionally, we state two needed results regarding the properties of integral operators on L^p spaces:

Fact 3.7 (Neumann series representation [114], Ex. 5.16)**.** *Let $K : X \to X$ be a bounded linear operator on a Banach space X with $\|X\| < 1$. Then $[I - K]$ is invertible and the series*

$$[I - K]^{-1} = \sum_{n=0}^{\infty} K^n$$

converges uniformly in the space of bounded linear endomorphisms on X.

Fact 3.8 (Schur's test; [115], Theorems 8.3.1-2)**.** *Let k be a measurable function on $\mathbb{R}^2$ that satisfies the mixed-norm conditions*

$$\underset{t \in \mathbb{R}}{ess\,sup} \int_{\mathbb{R}} |k(t,s)|\, ds = C_1 < \infty,$$
$$\underset{s \in \mathbb{R}}{ess\,sup} \int_{\mathbb{R}} |k(t,s)|\, dt = C_2 < \infty.$$

Then the integral operator

$$K[f(t)] = \int_{\mathbb{R}} k(t,s)\, f(s)\, ds$$

is a bounded endomorphism on $L^p(\mathbb{R})$ for all $p \in [1, \infty]$ and its operator norm satisfies

$$\|K\| \leq C_1^{1/q} C_2^{1/p},$$

where q is the Hölder conjugate of p.

Let Y^*, Y^{**}, and $\tilde{Y}$ and denote the respective phenotypes of parent-parent-offspring trio. By assumption, the joint distribution of $\tilde{Y}$, Y^*, Y^{**}, is multivariate normal and at equilibrium, $\tilde{Y}$ and Y^* have correlation $\tau \in (0,1)$. Thus,

$$Cov(Y^*, Y^{**}, \tilde{Y}) = \sigma^2_{Y,\infty} \begin{pmatrix} 1 & & \\ r & 1 & \\ \tau & \tau & 1 \end{pmatrix},$$

and

$$
\begin{aligned}
(Y^{**}|\tilde{Y}=w) \overset{D}{=} (Y^{*}|\tilde{Y}=t) \sim \mathcal{N}(\tau \cdot w, 1-\tau^2),\\
(Y^{**}|Y^{*}=t) \sim \mathcal{N}(r \cdot w, 1-r^2).
\end{aligned}
$$

The genetic value at each diploid locus is the weighted sum of that at two haploid loci: $Z_k = \frac{1}{\sqrt{2}}(G_k + G_{k'})$. We can write conditional second moments of the offsprings' haploid effects at distinct diploid loci as

$$
\mathbb{E}[\tilde{G}_k\tilde{G}_{l\neq k,k'}|\tilde{Y}=w] = \int \mathbb{E}[\tilde{G}_k\tilde{G}_l|Y^{*}=u, Y^{**}=v]dP_{Y^{*},Y^{**}|\tilde{Y}}(u,v,w).
$$

Let ω_k^{*} denote the event that the offspring received the haploid allele G_k from the first parent (with phenotype Y^{*}) and so on for ω_l^{**} etc.. Then we can decompose the conditional expectation in the above integrand as

$$
\begin{aligned}
\mathbb{E}[\tilde{G}_k\tilde{G}_l|Y^{*}=u, Y^{**}=v] =& \frac{1}{4}\Big(\mathbb{E}[\tilde{G}_k\tilde{G}_l|\{Y^{*}=u, Y^{**}=v\}\cap\omega_k^{*}\cap\omega_l^{*}]\\
&+\mathbb{E}[\tilde{G}_k\tilde{G}_l|\{Y^{*}=u, Y^{**}=v\}\cap\omega_k^{*}\cap\omega_l^{**}]\\
&+\mathbb{E}[\tilde{G}_k\tilde{G}_l|\{Y^{*}=u, Y^{**}=v\}\cap\omega_k^{**}\cap\omega_l^{*}]\\
&+\mathbb{E}[\tilde{G}_k\tilde{G}_l|\{Y^{**}=v, Y^{**}=v\}\cap\omega_k^{**}\cap\omega_l^{**}]\Big).\\
=& \frac{1}{4}\left(\mathbb{E}[G_k^{*}G_l^{*}|Y^{*}=u] + \mathbb{E}[G_k^{*}G_l^{**}|Y^{*}=u, Y^{**}=v\}\right.\\
&\left.+\mathbb{E}[G_k^{**}G_l^{*}|Y^{*}=u, Y^{**}=v] + \mathbb{E}[G_k^{**}G_l^{**}|Y^{**}=v]\right)\\
=& \frac{1}{4}\left(\mathbb{E}[G_k^{*}G_l^{*}|Y^{*}=u] + \mathbb{E}[G_k^{*}|Y^{*}=u]\mathbb{E}[G_l^{**}|Y^{**}=v\}\right.\\
&\left.+\mathbb{E}[G_k^{**}|Y^{**}=u]\mathbb{E}[G_l^{*}|Y^{*}=v] + \mathbb{E}[G_k^{**}G_l^{**}|Y^{**}=v]\right).
\end{aligned}
$$

Employing our linearity of allelic effects assumption, we have that for $k = 1, \dots, p$,

$$
\mathbb{E}[G_k|Y=y] = \frac{\zeta y}{\sigma_{Y,\infty}}, \qquad \mathbb{E}[G_k|Y^{*}=y^{*}] = \frac{r\zeta y^{*}}{\sigma_{Y,\infty}},
$$

where $\zeta = \sqrt{\mu(m)/r}$. Thus,

$$\mathbb{E}[\tilde{G}_k\tilde{G}_l|Y^* = u, Y^{**} = v] = \frac{1}{4}\left(\mathbb{E}[G^*_kG^*_l|Y^* = u] + \mathbb{E}[G^{**}_kG^{**}_l|Y^{**} = v] + 2\mu uvr^{-1}\sigma^{-2}_{Y,\infty}\right)$$

Defining

$$f_t(u) = \mathbb{E}[G^*_kG^*_l|Y^* = u]_t,$$

where t indexes reproductive generations, we then have

$$f_{t+1}(w) = \frac{1}{4}\iint \left(f_t(u) + f_t(v) + 2\mu uvr^{-1}\sigma^{-2}_{Y,\infty}\right) p_{Y^*,Y^{**}|\tilde{Y}=w}(u,v)\,du\,dv.$$

Using linearity of the integral,

$$\begin{aligned} f_{t+1}(w) =& \frac{1}{4}\iint f_t(u)p_{Y^*,Y^{**}|\tilde{Y}=w}(u,v)\,du\,dv + \frac{1}{4}\iint f_t(v)p_{Y^*,Y^{**}|\tilde{Y}=w}(u,v)\,du\,dv \\ &+ \frac{1}{2}\iint \mu uvr^{-1}\sigma^{-2}_{Y,\infty}p_{Y^*,Y^{**}|\tilde{Y}=w}(u,v)\,du\,dv \\ =& \frac{1}{2}\int f_t(u)p_{Y^*|\tilde{Y}=w}(u)\,du + \frac{1}{2}\mu r^{-1}\iint uv\sigma^{-2}_{Y,\infty}p_{Y^*,Y^{**}|\tilde{Y}=w}(u,v)\,du\,dv. \end{aligned}$$

The right hand integral above is simply the expectation of the product of two zero expectation, variance one, covariance r, jointly normal random variables. Thus,

$$f_{t+1}(w) = \frac{1}{2}\int f_t(u)p_{Y^*|\tilde{Y}=w}(u)\,du + \frac{1}{2}\mu(m).$$

This defines a integral equation recurrence relation of the form:

$$f_{t+1}(w) = h_t(w) - \int f_t(s)k(w,s)\,ds,$$

with known function and kernel:

$$\begin{aligned} h_t(w) &= \frac{1}{2}\mu_t(m), \\ k_t(w,s) &= \frac{1}{2}p_{Y^*|\tilde{Y}}(s,w), \end{aligned}$$

and where $\mu_t(m) \sim o(m^{-1})$ for all $t \in \mathbb{Z}^+$. Explicitly, the kernel function is given by

$$k_{ts}(w,s) = \frac{1}{2\sqrt{2\pi(1-\tau^2)}} \exp\left[-\frac{(s-\tau w)^2}{2(1-\tau^2)}\right], \quad \tau \in (0,1).$$

We now proceed to bound the conditional expectation $\mathbb{E}[\tilde{G}_k \tilde{G}_l | \tilde{Y} = t]$:

Lemma 3.9. *Consider the previously introduced recurrence relation:*

$$\begin{aligned} f_{t+1}(w) &= h_t(w) - \int f_t(s) k(w,s)\, ds, \\ h(w) &= \frac{1}{2}\mu_t(m) \sim \mathcal{O}(m^{-1}), \\ k_t(w,s) &= \frac{1}{2\sqrt{2\pi(1-\tau^2)}} \exp\left[-\frac{(s-\tau w)^2}{2(1-\tau^2)}\right], \quad \tau \in (0,1). \end{aligned}$$

Then,

$$\operatorname*{ess\,sup}_{s\in\mathbb{R}} |f_t(s)| \leq \sim \mathcal{O}(m^{-1}).$$

Proof. For fixed $w = w_0 \in \mathbb{R}$, $k(w_0, s)$, as a function of s, is simply 1/2 times Gaussian density and it's integral is independent of the mean $\tau \cdot w_0$. Hence,

$$\operatorname*{ess\,sup}_{w\in\mathbb{R}} \int_{\mathbb{R}} |k_t(w,s)|\, ds = \frac{1}{2} \equiv C_1.$$

A similar argument yields,

$$\operatorname*{ess\,sup}_{s\in\mathbb{R}} \int_{\mathbb{R}} |k_t(w,s)|\, dw = \frac{1}{2\tau_t} \equiv C_2.$$

Choose the Hölder conjugate pair $p = \infty$, $q = 1$. Then, by (3.8), the integral operator

$$K : \varphi \mapsto \int_{\mathbb{R}} \varphi(s) k_t(w,s)\, ds$$

is a bounded linear endomorphism on $L^\infty(\mathbb{R})$ with operator norm

$$\|K_t\| \leq C_1^1 C_2^0 = \frac{1}{2}, \quad \forall t \in \mathbb{Z}^+.$$

Next, applying (3.7), $[I - K_t]$ is invertible and the series

$$[I - K_t]^{-1} = \sum_{n=0}^{\infty} K_t^n$$

converges uniformly on $\mathscr{B}(L^\infty)$. Since we assume independence of loci and random mating at generation zero we have $f_0 \sim \mathcal{O}(m^{-1})$ and $\mu_0(m) = 0$. Thus,

$$\begin{aligned}
f_{t+1} &= K_t f_t + h_t \\
&\leq K_t f_t + \mu(m)/2 \\
&= K_t[K_{t-1} f_{t-1} + \mu_t(m)/2] + \mu(m/2) \\
&\vdots \\
&= K_t[K_{t-1} \cdots K_0[f_0 + \mu_0(m)/2] \cdots] + \mu(m/2) \\
\implies \|f_{t+1}\|_\infty &\leq \sum \left\|\frac{1}{2}\right\|^t \|\mu_t(m)/2\| \, \mathbb{E}[G_j^{**} G_k^{**} G_l^{**} | Y^{**} = v] \\
&\sim \mathcal{O}(m^{-1}).
\end{aligned}$$

□

Immediately, this yields a bound of the same order on the unconditional expectation

$$\begin{aligned}
\mathbb{E}[\tilde{G}_k \tilde{G}_l] &\leq \sup \mathbb{E}[\tilde{G}_k \tilde{G}_l | Y^* = u, Y^{**} = v] \int dP_{Y^*, Y^{**} | \tilde{Y}}(u, v, w) \\
\implies \mathbb{E}[\tilde{G}_k \tilde{G}_l] &\sim \mathcal{O}(m^{-1}).
\end{aligned}$$

This same argument is used to established bounds for the conditional third moments as follows (and hence unconditional third moments): Let j, k, l index distinct diploid sites. We proceed as above, expanding the conditional third moment

$$\begin{aligned}
\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|Y^*=u,Y^{**}=v] =&\frac{1}{8}\Big(\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^*\cap\omega_k^*\cap\omega_l^*]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^{**}\cap\omega_k^*\cap\omega_l^*]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^*\cap\omega_k^*\cap\omega_l^{**}]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^{**}\cap\omega_k^*\cap\omega_l^{**}]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^*\cap\omega_k^{**}\cap\omega_l^*]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=u,Y^{**}=v\}\cap\omega_j^{**}\cap\omega_k^{**}\cap\omega_l^*]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=v,Y^{**}=v\}\cap\omega_j^*\cap\omega_k^{**}\cap\omega_l^{**}]\\
&+\mathbb{E}[\tilde{G}_j\tilde{G}_k\tilde{G}_l|\{Y^*=v,Y^{**}=v\}\cap\omega_j^{**}\cap\omega_k^{**}\cap\omega_l^{**}]\Big)\\
=&\frac{1}{8}\big(\mathbb{E}[G_j^*G_k^*G_l^*|Y^*=u]\\
&+\mathbb{E}[G_j^{**}|Y^{**}=v]\mathbb{E}[G_k^*G_l^*|Y^*=u]\\
&+\mathbb{E}[G_l^{**}|Y^{**}=v]\mathbb{E}[G_j^*G_k^*|Y^*=u]\\
&+\mathbb{E}[G_k^{**}|Y^{**}=v]\mathbb{E}[G_j^*G_l^*|Y^*=u]\\
&+\mathbb{E}[G_j^*|Y^*=u]\mathbb{E}[G_k^{**}G_l^{**}|Y^{**}=v]\\
&+\mathbb{E}[G_l^*|Y^*=u]\mathbb{E}[G_j^{**}G_k^{**}|Y^{**}=v]\\
&+\mathbb{E}[G_k^*|Y^*=u]\mathbb{E}[G_j^{**}G_l^{**}|Y^{**}=v]\\
&+\mathbb{E}[G_j^{**}G_k^{**}G_l^{**}|Y^{**}=v]\big)\\
&\equiv\frac{1}{8}\mathbb{E}[G_j^*G_k^*G_l^*|Y^*=u]+\phi(u,v)+\frac{1}{8}\mathbb{E}[G_j^{**}G_k^{**}G_l^{**}|Y^{**}=v],
\end{aligned}$$

where $\phi(u,v)$ includes all the terms we have previously bounded and $\phi(u,v)\sim\mathcal{O}(m^{-3/2})$. We can then bound the the remaining terms, which are of the form $f(u)=\mathbb{E}[G_j^*G_k^*G_l^*|Y^*=u]$, by applying the same argument. Briefly,

$$\begin{aligned}
f_{t+1}=K_tf_t+h_t\leq\cdots\leq K_t[K_{t-1}\cdots K_0[f_0+\mathcal{O}(m^{-3/2})\cdots]+\mathcal{O}(m^{-3/2})]\\
\implies\mathbb{E}[G_j^*G_k^*G_l^*]\sim\mathcal{O}(m^{-3/2}).
\end{aligned}$$

Likewise, bounds are established for higher order moments recursively by applying a similar argument to the expanded conditional moments. E.g., 2/16 of the expansion of $\mathbb{E}[\tilde{G}_i\tilde{G}_j\tilde{G}_k\tilde{G}_l|Y^* = u, Y^{**} = v]$ with $|\{i,j,k,l\}| = 4$ are comprised of

$$\begin{aligned}\mathbb{E}[G_i^*G_j^*G_k^*G_l^{**}|Y^* = u, Y^{**} = v] &= \mathbb{E}[G_i^*G_j^*G_k^*|Y^* = u, Y^{**} = v] \cdot \mathbb{E}[G_l^{**}|Y^* = u, Y^{**} = v] \\ &\sim \mathcal{O}(m^{-2}).\end{aligned}$$

When $|\{i,j,k,l\}| = 3$, we simply have fewer possible inheritance patterns to consider. E.g., denoting $\tilde{G}_{jj'kl} = \tilde{G}_j\tilde{G}_{j'}\tilde{G}_k\tilde{G}_l$ and $\mathcal{P} = \{Y^* = u, Y^{**} = v\}$, we have without loss of generality,

$$\begin{aligned}\mathbb{E}\tilde{G}_j\tilde{G}_{j'}\tilde{G}_k\tilde{G}_l[|Y^* = u, Y^{**} = v] =& \frac{1}{8}\Big(\mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{**}\cap\omega_{j'}^{*}\cap\omega_k^{*}\cap\omega_l^{*}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{*}\cap\omega_{j'}^{**}\cap\omega_k^{*}\cap\omega_l^{*}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{**}\cap\omega_{j'}^{*}\cap\omega_k^{**}\cap\omega_l^{*}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{*}\cap\omega_{j'}^{**}\cap\omega_k^{**}\cap\omega_l^{*}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{**}\cap\omega_{j'}^{*}\cap\omega_k^{**}\cap\omega_l^{**}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{*}\cap\omega_{j'}^{**}\cap\omega_k^{**}\cap\omega_l^{**}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{**}\cap\omega_{j'}^{*}\cap\omega_k^{*}\cap\omega_l^{**}] \\ &+ \mathbb{E}[\tilde{G}_{jj'kl}|\mathcal{P}\cap\omega_j^{*}\cap\omega_{j'}^{**}\cap\omega_k^{*}\cap\omega_l^{**}]\Big).\end{aligned}$$

None of the above quantities, in terms of available bounds, is asymptotically larger than $\mathcal{O}(m^{-2})$. This result can be formalized as follows:

Lemma 3.10. *For a finite collection of M distinct diploid loci $\{\xi_\iota\}_\iota$, we have*

$$\mathbb{E}\left[\prod_{\iota\in\mathcal{J}}\xi_\iota\right] \sim \mathcal{O}\left(m^{-M/2}\right).$$

Proof. This is an immediate consequence of the previous argument as elements $\left\{\prod_{\iota\in\mathcal{J}}\xi_\iota : \mathcal{J}\in 2^{\{1,\dots,p\}}\right\}$ are simply linear combinations of the elements of $\left\{\prod_{\iota\in\mathcal{J}\cup\mathcal{J}'}G_\iota : \mathcal{J}\in 2^{\{1,\dots,p\}}\right\}$. □

3.3.5 Residual maximum likelihood estimation under AM

Recall that the REML estimator is $\hat{h}^2_{\text{REML}} = \hat{\gamma}(1+\hat{\gamma})^{-1}$ where $\hat{\gamma}$ as the solution to the *REML equation*

$$\frac{y^T P_\gamma Z Z^T P_\gamma y}{\text{tr}\,[P_\gamma Z Z^T]} = \frac{y^T P_\gamma P_\gamma y}{\text{tr}\,[P_\gamma]},$$

and P_γ is as in (3.1.7). This is equivalent to maximum likelihood estimation of the infinitesimal model

$$K^T y = p^{-1/2} K^T Z u + K^T e, \qquad u \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_g^2), \qquad e \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2), \tag{3.3.1}$$

under the change of variables $\gamma = \sigma_g^2/\sigma_e^2$ where $K^T : \mathbb{R}^n \to (\text{col } X)^\perp$ is a maximal rank matrix projecting to the orthogonal complement of the fixed effect covariates $X \in \mathbb{R}^{n \times c}$ and is such that $K^T K = I_{n-c}$. However, implicit in (3.3.1) is the assumption that all p SNPs included in (3.3.1) (and therefore in the GRM) are causal–that is, that $\omega = m/p = 1$. Generally speaking, this unlikely to be true for any complex trait of interest. Fortunately, Jiang and colleagues [112, 113] have demonstrated that $\hat{h}^2_{\text{REML}}$ is a consistent estimator of h^2 in a high dimensional setting, even when $\omega < 1$:

Fact 3.11 (Jiang's consistency theorem [Theorem 3.1 of [112]]). *Consider the model presented in (3.3.1) and assume that the rows of Z are independent realizations of standardized, independent sub-Gaussian random variables with finite fourth moments. Further, suppose that the true values of the variance components σ_g^2, σ_e^2 are positive and that $n, m, p \to \infty$ such that*

$$\frac{n}{p} \to \tau \in (0, 1], \qquad \frac{m}{p} \to \omega \in (0, 1].$$

Then we have that

1. $\hat{\gamma} \xrightarrow{p} \omega\gamma$, *where* $\gamma = \sigma_g^2/\sigma_e^2$
2. $\hat{\sigma}_e^2 = y^T P_{\hat{\gamma}} P_{\hat{\gamma}} y / tr[P_{\hat{\gamma}}] \xrightarrow{p} \sigma_e^2.$

Noting that the true heritability can be expressed as $h^2 = \omega^{-1}\hat{\gamma}(1 + \omega^{-1}\hat{\gamma})^{-1}$, this in turn implies that the REML heritability estimator is consistent despite the misspecification (i.e., the incorrect assumption that $\omega = 1$):

$$\widehat{h}^2_{\text{REML}} = \widehat{\gamma}(1+\widehat{\gamma})^{-1} \xrightarrow{p} \omega\omega^{-1}\sigma_g^2/\sigma_e^2(1+\omega\omega^{-1}\sigma_g^2/\sigma_e^2)^{-1} = h^2.$$

The proof of (3.11) is highly technical thus omitted. However, the following result from random matrix theory is central to Jiang's argument:

Fact 3.12 (Convergence of the ESD to the Marčenko-Pastur law)**.** *Let $Z \in \mathbb{C}^{n\times p}$ be a random matrix with independent entries with zero expectation, unit variance, and finite fourth moments. Define the empirical spectral distribution (ESD) $F^S : \mathbb{R} \to [0,1]$ of the sample covariance matrix $S = p^{-1}ZZ^T$ by*

$$F^S(x) = \frac{1}{n}\sum_k [\![\lambda_k \le x]\!]$$

where $\{\lambda_k\}_1^n$ are the eigenvalues of S. Further assume that $n, m, p \to \infty$ such that

$$n/p \to \tau \in (0,1], \quad ,m/p \to \omega \in (0,1].$$

Then as $p \to \infty$, $F_n^S \xrightarrow{a.s.} \varphi_\tau$ where φ_τ denotes the Marčenko-Pastur law F_τ with density

$$\varphi_\tau(x) = \frac{1}{2\pi\tau x}\sqrt{(b_\tau - x)(x - a_\tau)}\,[\![x \in [a_\tau, b_\tau]]\!]\,,$$

and we've defined $a_\tau = (1-\tau^{1/2})^2$, $b_\tau = (1+\tau^{1/2})^2$.

Under the phenotypic assortment model, that the elements of Z are independent are violated. Specifically, the rows of Z, each of which corresponds to the standardized genotypes of a given individual, can be regarded as i.i.d. multivariate Bernoulli random vectors with covariance matrix

$$\Upsilon(m) = \left[\begin{array}{cccc|c} 1+\mu(m) & & & \text{Sym.} & \\ 2\mu(m) & \ddots & & & \text{Sym.} \\ \vdots & \ddots & \ddots & & \\ 2\mu(m) & \cdots & 2\mu(m) & 1+\mu(m) & \\ \hline \multicolumn{4}{c|}{0_{(p-m)\times m}} & I_{p-m} \end{array}\right],$$

where again the upper left quadrant corresponds to causal variants[2]. However, O'Rourke [personal communication] has extended (3.12) as follows:

[2] Depending on whether or not genotypes are standardized within sample or according to a reference panel, the

Lemma 3.13 (O'Rourke's extension of (3.12) [unpublished])**.** . *Let Z be a random matrix with i.i.d. rows $\{z_k\}_{k=1}^{p}$, each of which has covariance $\Upsilon(m)$. Further assume that*

1. $\mu(m) = \mathcal{O}(m^{-1})$.
2. *There exists $\kappa > 0$ such that $\sup_{1 \leq k \leq p} |z_k| \leq \kappa$ with probability one.*
3. $\mathbb{E}[z_k^2 z_l^2] = 1 + o(1)$ *uniformly for distinct k, l*
4. $\mathbb{E}[z_k^3 z_l] = o(1)$ *uniformly for distinct k, l*
5. *One has*

$$\mathbb{E}[z_k z_l z_r z_s] = \begin{cases} o(m^{-1}) & \text{if } |\{k, l, r, s\}| = 4 \\ o(m^{-1/2}) & \text{if } |\{k, l, r, s\}| = 3 \end{cases}$$

Suppose $n/p \to \tau > 0$ as $n \to \infty$ and $m \geq cp$ for some $c > 0$. Then the ESD of $p^{-1}ZZ^T$ converges almost surely to $\varphi_\tau(x)$ as $n \to \infty$.

We have demonstrated that, under the phenotypic assortment model, the sample GRM satisfies the above conditions. Specifically, condition 1 is demonstrated in section §3.3.1, condition 2 is ensured under the exchangeable loci assumption when applying an MAF threshold to SNPs included in the GRM, and conditions 3-5 are consequences of the results presented in section 3.3.4. Thus we have proven lemma 3.3 and have laid the foundation necessary for a proof of 3.4.

3.3.6 LD score regression under AM

As in [50], consider the phenotype $y = Zu + e$ where all elements of the right hand side are mutually independent random variables. Under the phenotypic assortment model, we have that $u_j \overset{i.i.d.}{\sim} \mathcal{N}(0, m^{-1}\sigma_{g,0}^2)$, $e \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2)$, and $Var(y) = 1 = \sigma_{g,\infty}^2 + \sigma_e^2$ such that $h_{g,\infty}^2 = \sigma_{g,\infty}^2$. Having estimated GWAS statistics $\{\hat{u}\}_{j=1}^{m}$, each defined $\hat{u}_j = z_{\cdot j}^T y$, for a given realization of the above, consider the corresponding test statistics $\{\chi_j^2\}_{j=1}^{m}$, each defined $\chi_j^2 = n\hat{u}_j^2$. Following [50], the expectation of the test statistic is computed as follows

covariance matrix will be respectively as above or $\tilde{\Upsilon}(m) = \left[\begin{array}{ccc|c} 1 & & & \\ \frac{2\mu(m)}{1+\mu(m)} & \ddots & & \text{Sym.} \\ \vdots & \ddots & \ddots & \\ \hline & & 0_{(p-m)\times m} & I_{p-m} \end{array}\right]$.

$$
\begin{aligned}
Var[\hat{u}_j] &= \mathbb{E}\left[Var[\hat{u}_j|Z]\right] + Var\left[\mathbb{E}[\hat{u}_j|Z]\right] \\
&= \mathbb{E}\left[Var[\hat{u}_j|Z]\right], \\
Var[\hat{u}_j|Z] &= \mathbb{E}\left[n^{-2}\sum_{h=1}^{n}\sum_{i=1}^{n} z_{ij}z_{hj}\left(\sum_{k=1}^{m}(z_{ik}u_k) + e_i\right)\left(\sum_{l=1}^{m}(z_{hl}u_l) + e_h\right)|Z\right] \\
&= \mathbb{E}\left[n^{-2}\sum_{h=1}^{n}\sum_{i=1}^{n} z_{ij}z_{hj}\left(\sum_{k=1}^{m}(z_{ik}u_k)\right)\left(\sum_{l=1}^{m}(z_{hl}u_l)\right) + \sum_{i=h}^{n} z_{ij}z_{hj}e_h e_i|Z\right] \\
&= \mathbb{E}\left[n^{-2}\sum_{h=1}^{n}\sum_{i=1}^{n} z_{ij}z_{hj}\left(\sum_{k=l}^{m}(z_{ik}z_{hl}u_k u_l)\right) + \sum_{i=h}^{n} z_{ij}^2 e_i^2|Z\right] \\
&= n^{-2}\sum_{h=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}\left[z_{ij}z_{hj}z_{ik}z_{hk}u_k^2|Z\right] + n^{-1}\sigma_e^2 \\
&= n^{-2}m^{-1}\sigma_{g,0}^2\sum_{h=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{m} z_{ij}z_{hj}z_{ik}z_{hk} + n^{-1}(1-\sigma_{g,\infty}^2) \\
\implies \mathbb{E}\left[Var[\hat{u}_j|Z]\right] &= n^{-2}m^{-1}\sigma_{g,0}^2\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{m} z_{ij}^2 z_{ik}^2\right] + n^{-1}(1-\sigma_{g,\infty}^2) \\
&= m^{-1}\sigma_{g,0}^2\mathbb{E}[\sum_{k=1}^{m}\tilde{r}_{jk}^2] + n^{-1}(1-\sigma_{g,\infty}^2) \\
\implies \mathbb{E}\left[\chi_j^2\right] &= nm^{-1}\sigma_{g,0}^2\mathbb{E}[\sum_{k=1}^{m}\tilde{r}_{jk}^2] + 1 - \sigma_{g,\infty}^2
\end{aligned}
$$

where sample LD measure is defined $\tilde{r}_{jk}^2 = n^{-2}\sum_{k=1}^{m}\sum_{h=1}^{n}\sum_{i=1}^{n} z_{ij}^2 z_{ik}^2$. Applying the delta method obtains the approximation $\mathbb{E}[\sum_{k=1}^{m}\tilde{r}_{jk}^2] \approx \ell_j + n^{-1}(m-\ell_j)$, leading to

$$
\begin{aligned}
\mathbb{E}[\chi_j^2] &\approx nm^{-1}\sigma_{g,0}^2\left(\ell_j + n^{-1}(m-\ell_j)\right) + 1 - \sigma_{g,\infty}^2 \\
&\approx nm^{-1}\sigma_{g,0}^2\ell_j + \sigma_{g,0}^2 + 1 - \sigma_{g,\infty}^2. \qquad (3.3.2)
\end{aligned}
$$

However, both ℓ_j and $\sigma_{g,\infty}^2$ and can be expressed as functions of equilibrium correlation between causal variants, and (3.3.2) will undoubtedly admit a simpler representation. Further complicated matters is that, in practice, the true ℓ_j is replaced by a local LD score $\ell_{[j]} = \sum_{k\in\mathcal{B}_{1\text{cM}}^{[j]}} r_{jk}^2$. Work on expressing $\mathbb{E}[\chi_j^2]$ as an affine function of $\ell_{[j]}$ is ongoing.

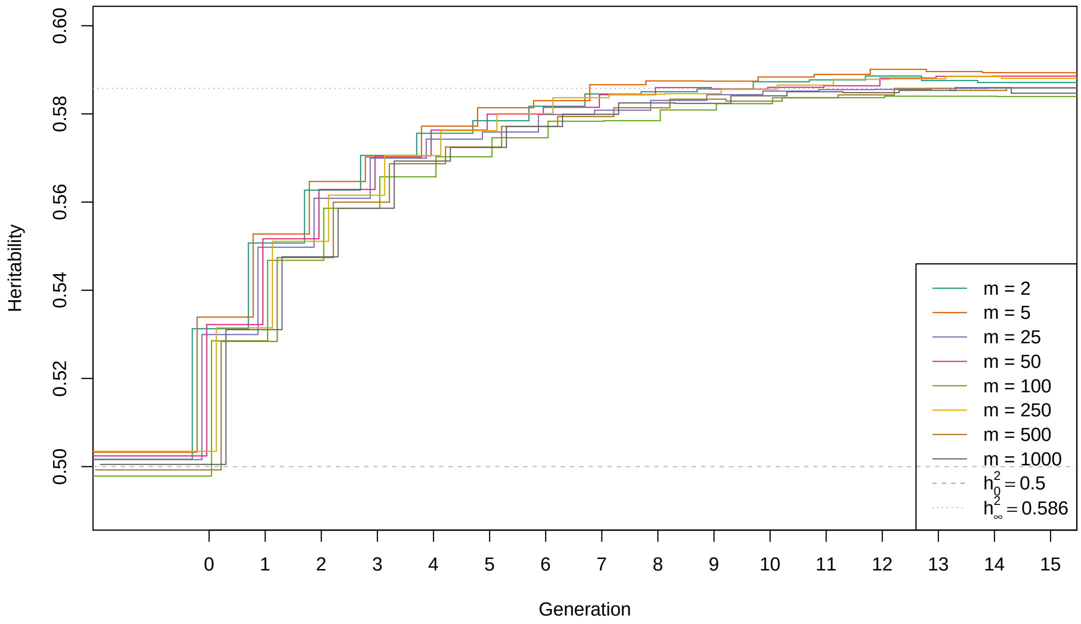
3.4 Computational methods

We conducted extensive simulations to direct and verify theoretical work. Given our interest in high-dimensional asymptotics, the generation of realistic genotype/phenotype data entailed developing new software capable of efficiently modeling populations of millions of individuals and genetic architectures with hundreds of thousands of causal variants. A complete discussion of the novel simulation algorithms, which we have implemented in the *Forward-time Population Structure Simulation* (FPSS) software package, is beyond the scope of this dissertation.

Founder data consisted of imputed genotypes at one million randomly selected SNP loci in a sub-sample of 435,301 European UK Biobank participants [1]. All SNPs were chosen to meet the following criteria: minor allele frequency greater 0.01, Hardy-Weinberg p-value greater than 10^{-6}, INFO score of at least 0.95, and presence on the 1,000 Genomes Phase 3 (1KG3) three reference panel [116] using Plink v1.9 [117]. Genotype data were then phased to the 1KG3 reference panel in batches of 40,000 individuals using Eagle v2.4.1 [118]. Simulation input data was created by first duplicating founder genotype data at random to created a population of one million individuals and then simulating five generations of random mating. Assortative mating simulations ran for fifteen generations under a variety of conditions (see Section §3.2) and produced results congruent with the theoretical derivations presented in 3.1.2.2 (Figure 3.10). All simulations used a recombination map derived from the 1KG3 data using a 50 kilobase sliding window. Samples of unrelated individuals were produced by removing relatives as close or closer than second cousins using true pedigree information. We also performed a small number of analogous simulations using a population of three million individuals under a limited set of conditions due to computational constraints.

We used GCTA [65] v1.91.3b to construct genomic related matrices and perform HE regression. We obtained REML heritability estimates using Bolt-LMM v2.3.2 [68, 69] for computational efficiency; though Bolt-LMM uses a randomized algorithm, its numerical accuracy is comparable to that of the exact algorithm implemented GCTA [5]. We used Plink v1.9 [117] to obtain GWAS summary statistics and LDSC v1.0.1 [50] to estimate within-sample LD scores using a one centimorgan sliding window and to perform LD score regression.

Figure 3.10: Sample heritability over time under phenotypic assortment



True sample heritabilities in forward time simulation of assortative mating for varying numbers of causal variants m in one million individuals. Spousal correlation was set at $r = 0.5$ and generation zero heritability at $h_0^2 = 0.5$. Results are consistent with previous theory predicting $h_\infty^2 \approx 0.586$. Data are jittered horizontally for clarity.

3.5 Discussion

3.5.1 Summary

In the preceding chapter, we have presented ongoing work characterizing the behavior of three widely-used estimators of SNP heritability under the phenotypic assortment model. Haseman-Elston regression produces dramatic overestimates of the true equilibrium heritability; we have derived closed form expressions for this bias (lemma 3.1, corollary 3.2, figure 3.2) and confirmed our results via simulation (figure 3.3). With respect to the REML estimator, simulation results demonstrate that, though initially biased upward, the estimated heritability is asymptotically unbiased with respect to the generation zero heritability (figure 3.4, figure 3.6), and thus provides an underestimate of the true equilibrium heritability in large samples. We have developed the foundation for a theoretical explanation of this behavior by characterizing the limiting spectral distribution of the sample genomic covariance matrix (lemma 3.3, conjecture 3.4), in the process of which introducing novel theory characterizing the joint distribution of causal variants under the phenotypic assortment model (lemma 3.10). Ongoing simulations are so far congruent with our conjecture (figure 3.4, figure 3.6). Finally, initial simulation results demonstrate that LD score regression produces substantially upwardly biased estimates of the equilibrium heritability (figure 3.7) and inflated intercepts (figure 3.7). We are currently working on a theoretical explanation of this behavior (section 3.3.6)

3.5.2 Limitations and future directions

In addition to the outstanding problems mentioned above, there are several limitations to the present investigation. Some of these limitations are imposed by the assumptions of the phenotypic assortment model itself: independence of parent-offspring environments, independence of heritable and non-heritable influences on the trait under assortment, and independence of mate availability and environments. All of these scenarios are readily handled by the novel simulation software, FPSS, developed in the course of the preceding work, and we intend to examine their effects on heritability estimators in future work. In terms of our results regarding the REML estimator, notable assumptions include the multivariate normality of parent and offspring phenotypes and the independence of non-causal variants in the genotype matrix. Whereas, for the latter, results regarding the limiting spectral distribution of GRM are likely extensible to the case of local LD (see, e.g., [119]), the former assumption is essential to our arguments bounding the higher order moments of causal variants presented in section 3.3.4. Finally, we note that many populations are unlikely to be in equilibrium even if subject to AM as described by the phenotypic assortment model. Thus, for traits subject to AM for a small number of generations, our results are likely to represent upper bounds on the degree of bias to be expected. We plan to address this scenario via simulation in future work.

Bibliography

1. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. en. *PLOS Medicine* **12,** e1001779. ISSN: 1549-1676 (Mar. 2015).

2. Thompson, P. M. *et al.* The ENIGMA Consortium: Large-Scale Collaborative Analyses of Neuroimaging and Genetic Data. *Brain Imaging and Behavior* **8,** 153–182. ISSN: 1931-7557 (2014).

3. Wray, N. R. *et al.* Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression. English. *Nature Genetics* **50.** Citation Key Alias: WrayGenomewideassociationanalyses2017, WrayGenomewideassociationanalyses2018, 668–681. ISSN: 1546-1718 (May 2018).

4. Border, R. *et al.* No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry,* appi.ajp.2018.18070881. ISSN: 0002-953X (Mar. 2019).

5. Border, R. & Becker, S. Stochastic Lanczos Estimation of Genomic Variance Components for Linear Mixed-Effects Models. *BMC Bioinformatics* **20,** 411. ISSN: 1471-2105 (July 2019).

6. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *American Journal of Psychiatry* **157,** 1552–1562. ISSN: 0002-953X (Oct. 2000).

7. McInnes, L. A. & Freimer, N. B. Mapping Genes for Psychiatric Disorders and Behavioral Traits. eng. *Current Opinion in Genetics & Development* **5,** 376–381. ISSN: 0959-437X (June 1995).

8. Ramamoorthy, S. *et al.* Antidepressant- and Cocaine-Sensitive Human Serotonin Transporter: Molecular Cloning, Expression, and Chromosomal Localization. eng. *Proceedings of the National Academy of Sciences of the United States of America* **90,** 2542–2546. ISSN: 0027-8424 (Mar. 1993).

9. Owens, M. J. & Nemeroff, C. B. Role of Serotonin in the Pathophysiology of Depression: Focus on the Serotonin Transporter. eng. *Clinical Chemistry* **40,** 288–295. ISSN: 0009-9147 (Feb. 1994).

10. Heils, A. *et al.* Allelic Variation of Human Serotonin Transporter Gene Expression. eng. *Journal of Neurochemistry* **66,** 2621–2624. ISSN: 0022-3042 (June 1996).

11. Stoltenberg, S. F. & Burmeister, M. Recent Progress in Psychiatric Genetics-Some Hope but No Hype. eng. *Human Molecular Genetics* **9,** 927–935. ISSN: 0964-6906 (Apr. 2000).

12. Buckland, P. R. Genetic Association Studies of Alcoholism–Problems with the Candidate Gene Approach. eng. *Alcohol and Alcoholism (Oxford, Oxfordshire)* **36,** 99–103. ISSN: 0735-0414 (2001 Mar-Apr).

13. Munafò, M. R. Candidate Gene Studies in the 21st Century: Meta-Analysis, Mediation, Moderation. en. *Genes, Brain and Behavior* **5,** 3–8. ISSN: 1601-183X (Feb. 2006).

14. Lander, E. S. & Schork, N. J. Genetic Dissection of Complex Traits. en. *Science* **265,** 2037–2048. ISSN: 0036-8075, 1095-9203 (Sept. 1994).

15. Terwilliger, J. D. & Weiss, K. M. Linkage Disequilibrium Mapping of Complex Disease: Fantasy or Reality? *Current Opinion in Biotechnology* **9,** 578–594. ISSN: 0958-1669 (Dec. 1998).

16. Colhoun, H. M., McKeigue, P. M. & Smith, G. D. Problems of Reporting Genetic Associations with Complex Outcomes. *The Lancet* **361,** 865–872. ISSN: 0140-6736 (Mar. 2003).

17. Caspi, A. *et al.* Role of Genotype in the Cycle of Violence in Maltreated Children. eng. *Science (New York, N.Y.)* **297,** 851–854. ISSN: 1095-9203 (Aug. 2002).

18. Caspi, A. *et al.* Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. English. *Science* **301,** 386–389. ISSN: 0036-8075, 1095-9203 (July 2003).

19. Niculescu, A. B. DISCovery in Psychiatric Genetics. eng. *Molecular Psychiatry* **19,** 145. ISSN: 1476-5578 (Feb. 2014).

20. Duncan, L. E. & Keller, M. C. A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry. eng. *The American Journal of Psychiatry* **168,** 1041–1049. ISSN: 1535-7228 (Oct. 2011).

21. Farrell, M. S. *et al.* Evaluating Historical Candidate Genes for Schizophrenia. eng. *Molecular Psychiatry* **20,** 555–562. ISSN: 1476-5578 (May 2015).

22. Munafò, M. R., Durrant, C., Lewis, G. & Flint, J. Gene × Environment Interactions at the Serotonin Transporter Locus. *Biological Psychiatry. Epigenetic Mechanisms in Psychiatry* **65,** 211–219. ISSN: 0006-3223 (Feb. 2009).

23. Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* Biological Insights from 108 Schizophrenia-Associated Genetic Loci. en. *Nature* **511,** 421–427. ISSN: 1476-4687 (July 2014).

24. Munafò, M. R. & Gage, S. H. Improving the Reliability and Reporting of Genetic Association Studies. *Drug and Alcohol Dependence* **132,** 411–413. ISSN: 0376-8716 (Oct. 2013).

25. Burton, P. R. *et al.* Size Matters: Just How Big Is BIG?: Quantifying Realistic Sample Size Requirements for Human Genome Epidemiology. eng. *International Journal of Epidemiology* **38,** 263–273. ISSN: 1464-3685 (Feb. 2009).

26. Bosker, F. J. *et al.* Poor Replication of Candidate Genes for Major Depressive Disorder Using Genome-Wide Association Data. en. *Molecular Psychiatry* **16,** 516–532. ISSN: 1359-4184 (May 2011).

27. Coleman, J. R. I., Consortium, M. D. D. W. G. o. t. P. G., Consortium, U. B. M. H., Eley, T. C. & Breen, G. Genome-Wide Gene-Environment Analyses of Depression and Reported Lifetime Traumatic Experiences in UK Biobank. en. *bioRxiv,* 247353 (Jan. 2018).

28. Van der Auwera, S. *et al.* Genome-Wide Gene-Environment Interaction in Depression: A Systematic Evaluation of Candidate Genes. en. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **177,** 40–49. ISSN: 1552-485X (Jan. 2018).

29. Johnson, E. C. *et al.* No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes. eng. *Biological Psychiatry.* ISSN: 1873-2402. doi:10.1016/j.biopsych.2017.06.033 (July 2017).

30. Culverhouse, R. C. *et al.* Collaborative Meta-Analysis Finds No Evidence of a Strong Interaction between Stress and 5-HTTLPR Genotype Contributing to the Development of Depression. en. *Molecular Psychiatry* **23,** 133–142. ISSN: 1476-5578 (Jan. 2018).

31. Border, R. *et al.* Imputation of Behavioral Candidate Gene Repeat Polymorphisms in 486,551 Publicly-Available UK Biobank Individuals. en. *bioRxiv,* 358267 (June 2018).

32. Brookes, K. J. The VNTR in Complex Disorders: The Forgotten Polymorphisms? A Functional Way Forward? *Genomics* **101,** 273–281. ISSN: 0888-7543 (May 2013).

33. Wendland, J. R., Martin, B. J., Kruse, M. R., Lesch, K.-P. & Murphy, D. L. Simultaneous Genotyping of Four Functional Loci of Human *SLC6A4*, with a Reappraisal of *5-HTTLPR* and Rs25531. en. *Molecular Psychiatry* **11,** 224–226. ISSN: 1476-5578 (Mar. 2006).

34. Dick, D. M. *et al.* Candidate Gene-Environment Interaction Research: Reflections and Recommendations. *Perspectives on psychological science : a journal of the Association for Psychological Science* **10,** 37–59. ISSN: 1745-6916 (Jan. 2015).

35. Keller, M. C. Gene-by-Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biological psychiatry* **75.** Citation Key Alias: KelleGene2014, keller2014genetexttimes. ISSN: 0006-3223. doi:10.1016/j.biopsych.2013.09.006 (Jan. 2014).

36. Border, R. & Keller, M. C. Commentary: Fundamental Problems with Candidate Gene-by-Environment Interaction Studies – Reflections on Moore et Thoemmes (2016). *Journal of child psychology and psychiatry, and allied disciplines* **58,** 328–330. ISSN: 0021-9630 (Mar. 2017).

37. Munafò, M. R., Zammit, S. & Flint, J. Limitations of Gene × Environment Interaction Models in Psychiatry. *Journal of child psychology and psychiatry, and allied disciplines* **55,** 1092–1101. ISSN: 0021-9630 (Oct. 2014).

38. Assary, E., Vincent, J. P., Keers, R. & Pluess, M. Gene-Environment Interaction and Psychiatric Disorders: Review and Future Directions. *Seminars in Cell & Developmental Biology. Arc/ARg3.1* **77,** 133–143. ISSN: 1084-9521 (May 2018).

39. Moore, S. R. Commentary: What Is the Case for Candidate Gene Approaches in the Era of High-Throughput Genomics? A Response to Border and Keller (2017). en. *Journal of Child Psychology and Psychiatry* **58,** 331–334. ISSN: 1469-7610 (Mar. 2017).

40. Moffitt, T. E. *Letter to Culverhouse* June 2012.

41. Uher, R. Gene-Environment Interactions in Severe Mental Illness. eng. *Frontiers in Psychiatry* **5,** 48. ISSN: 1664-0640 (2014).

42. National Center for Biotechnology Information, US National Library of Medicine & National Institutes of Health. *PubMed* en. https://www.ncbi.nlm.nih.gov/pubmed/.

43. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. en. *Perspectives on Psychological Science* **11,** 702–712. ISSN: 1745-6916 (Sept. 2016).

44. Cock, P. J. A. *et al.* Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. en. *Bioinformatics* **25,** 1422–1423. ISSN: 1367-4803 (June 2009).

45. Biobank, U. *Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers: Interim Data Release* http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf. 2015.

46. Conger, R. D., Schofield, T. J. & Neppl, T. K. Intergenerational Continuity and Discontinuity in Harsh Parenting. *Parenting, science and practice* **12,** 222–231. ISSN: 1529-5192 (Jan. 2012).

47. Derringer, J. *et al.* Genome-Wide Association Study of Behavioral Disinhibition in a Selected Adolescent Sample. eng. *Behavior Genetics* **45,** 375–381. ISSN: 1573-3297 (July 2015).

48. Stallings, M. C. *et al.* A Genome-Wide Search for Quantitative Trait Loci Influencing Substance Dependence Vulnerability in Adolescence. *Drug and alcohol dependence* **70,** 295–307 (2003).

49. Border, R. *et al.* Imputation of Behavioral Candidate Gene Repeat Variants in 486,551 Publicly-Available UK Biobank Individuals. eng. *European journal of human genetics: EJHG* **27,** 963–969. ISSN: 1476-5438 (June 2019).

50. Bulik-Sullivan, B. K. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. en. *Nature Genetics* **47.** Citation Key Alias: BulikLD2015, bulik-sullivan2015ldscore, 291–295. ISSN: 1546-1718 (Mar. 2015).

51. Center for Open Science. *Open Science Framework* https://osf.io/.

52. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11,** e1004219. ISSN: 1553-7358 (Apr. 2015).

53. Uher, R. *et al.* Serotonin Transporter Gene Moderates Childhood Maltreatment's Effects on Persistent but Not Single-Episode Depression: Replications and Implications for Resolving Inconsistent Results. *Journal of affective disorders* **135,** 56–65. ISSN: 0165-0327 (Dec. 2011).

54. Gonda, X. *et al.* Financial Difficulties but Not Other Types of Recent Negative Life Events Show Strong Interactions with 5-HTTLPR Genotype in the Development of Depressive Symptoms. *Translational Psychiatry* **6,** e798. ISSN: 2158-3188 (May 2016).

55. Miller, S. *et al.* in *B31. INFLAMMATION AND COPD* A2714–A2714 (American Thoracic Society, May 2015). doi:`10.1164/ajrccm-conference.2015.191.1_MeetingAbstracts.A2714`.

56. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider Scope: When Selection Bias Can Substantially Influence Observed Associations. eng. *International Journal of Epidemiology* **47.** Citation Key Alias: MunafoColliderscopewhen2018a, Munafo-Colliderscopewhen2018b, 226–235. ISSN: 1464-3685 (Jan. 2018).

57. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: Validity of a Brief Depression Severity Measure. eng. *Journal of General Internal Medicine* **16,** 606–613. ISSN: 0884-8734 (Sept. 2001).

58. McClellan, J. & King, M.-C. Genetic Heterogeneity in Human Disease. English. *Cell* **141,** 210–217. ISSN: 0092-8674, 1097-4172 (Apr. 2010).

59. Kapur, S., Phillips, A. G. & Insel, T. R. Why Has It Taken so Long for Biological Psychiatry to Develop Clinical Tests and What to Do about It? eng. *Molecular Psychiatry* **17,** 1174–1179. ISSN: 1476-5578 (Dec. 2012).

60. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic Architectures of Psychiatric Disorders: The Emerging Picture and Its Implications. *Nature reviews. Genetics* **13,** 537–551. ISSN: 1471-0056 (July 2012).

61. Hyman, S. E. & Krystal, J. H. *Report of the National Advisory Mental Health Council Workgroup on Genomics* https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/reports/report-of-the-national-advisory-mental-health-council-workgroup-on-genomics.shtml#roster.

62. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and Pitfalls in the Application of Mixed Model Association Methods. *Nature genetics* **46,** 100–106. ISSN: 1061-4036 (Feb. 2014).

63. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *arXiv preprint arXiv:1406.5823* (2014).

64. Zhou, X. & Stephens, M. Genome-Wide Efficient Mixed Model Analysis for Association Studies. *Nature genetics* **44,** 821–824. ISSN: 1061-4036 (June 2012).

65. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-Wide Complex Trait Analysis. *American Journal of Human Genetics* **88,** 76–82. ISSN: 0002-9297 (Jan. 2011).

66. Lippert, C. *et al.* FaST Linear Mixed Models for Genome-Wide Association Studies. en. *Nature Methods* **8,** 833–835. ISSN: 1548-7105 (Oct. 2011).

67. Zhou, X. & Stephens, M. Efficient Multivariate Linear Mixed Model Algorithms for Genome-Wide Association Studies. *Nature methods* **11,** 407 (2014).

68. Loh, P.-R. *et al.* Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts. en. *Nature Genetics* **47,** 284–290. ISSN: 1546-1718 (Mar. 2015).

69. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-Model Association for Biobank-Scale Datasets. *Nature genetics,* 1 (2018).

70. Searle, S. R., Casella, G. & McCulloch, C. E. *Variance Components* (John Wiley & Sons, United States, 2009).

71. Graser, H.-U., Smith, S. P. & Tier, B. A Derivative-Free Approach for Estimating Variance Components in Animal Models by Restricted Maximum Likelihood. en. *Journal of Animal Science* **64,** 1362–1370. ISSN: 0021-8812 (May 1987).

72. Björck, A. *Numerical Methods in Matrix Computations* (Springer, Switzerland, 2015).

73. Atkinson, K. E. *An Introduction to Numerical Analysis* ISBN: 780471624899 (John Wiley & Sons, United Kingdom, 2008).

74. O'Leary, D. P. The Block Conjugate Gradient Algorithm and Related Methods. *Linear Algebra and its Applications. Special Volume Dedicated to Alson S. Householder* **29,** 293–322. ISSN: 0024-3795 (Feb. 1980).

75. Frommer, A. & Maass, P. Fast CG-Based Methods for Tikhonov-Phillips Regularization. en. *SIAM Journal on Scientific Computing* **20,** 1831–1850 (1999).

76. Sogabe, T. A Fast Numerical Method for Generalized Shifted Linear Systems with Complex Symmetric Matrices. en. *Recent Developments of Numerical Analysis and Numerical Computation Algorithms,* 13 (2010).

77. Hutchinson, M. F. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics - Simulation and Computation* **19,** 433–450. ISSN: 0361-0918 (Jan. 1990).

78. Avron, H. & Toledo, S. Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-Definite Matrix. *J. ACM* **58,** 8:1–8:34. ISSN: 0004-5411 (Apr. 2011).

79. Golub, G. H. & Meurant, G. *Matrices, Moments and Quadrature with Applications:* en. ISBN: 978-1-4008-3388-7. doi:10.1515/9781400833887 (Princeton University Press, Princeton, Jan. 2009).

80. Ubaru, S., Chen, J. & Saad, Y. Fast Estimation of Tr(f(A)) via Stochastic Lanczos Quadrature. *SIAM Journal on Matrix Analysis and Applications* **38,** 1075–1099. ISSN: 0895-4798 (Jan. 2017).

81. Chen, J. & Saad, Y. A Posteriori Error Estimate for Computing Tr(f(A)) by Using the Lanczos Method. en. *arXiv:1802.04928 [math].* arXiv: `1802.04928 [math]` (Feb. 2018).

82. Border, R. Stochastic Lanczos Likelihood Estimation of Genomic Variance Components. *Applied Mathematics Graduate Theses & Dissertations* (Nov. 2018).

83. Zhu, S. & Wathen, A. J. Essential Formulae for Restricted Maximum Likelihood and Its Derivatives Associated with the Linear Mixed Models. *arXiv:1805.05188 [stat].* arXiv: `1805.05188 [stat]` (May 2018).

84. McCulloch, C., Searle, S. R. & Neuhaus, J. M. *Generalized, Linear, and Mixed Models* ISBN: 978-0-470-07371-1 (John Wiley & Sons, Hoboken, New Jersey, 2008).

85. Loh, P.-R. *et al.* Contrasting Genetic Architectures of Schizophrenia and Other Complex Diseases Using Fast Variance Components Analysis. *Nature genetics* **47,** 1385–1392. ISSN: 1061-4036 (Dec. 2015).

86. Schling, B. *The Boost C++ Libraries* ISBN: 978-0-9822191-9-5 (XML Press, CA, United States, 2011).

87. Wang, E. *et al.* in *High-Performance Computing on the Intel® Xeon Phi™* 167–188 (New York, 2014).

88. Oliphant, T. NumPy: A Guide to NumPy (2006).

89. Jones, E., Oliphant, T., Peterson, P., *et al.* SciPy: Open Source Scientific Tools for Python (2001).

90. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C. & Sorensen, D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLoS genetics* **9,** e1003608 (2013).

91. Evans, L. M. *et al.* Comparison of Methods That Use Whole Genome Data to Estimate the Heritability and Genetic Architecture of Complex Traits. eng. *Nature Genetics* **50,** 737–745. ISSN: 1546-1718 (May 2018).

92. Stulp, G., Simons, M. J., Grasman, S. & Pollet, T. V. Assortative Mating for Human Height: A Meta-analysis. *American Journal of Human Biology* **29.** ISSN: 1042-0533. doi:`10.1002/ajhb.22917` (2017).

93. Alford, J. R., Hatemi, P. K., Hibbing, J. R., Martin, N. G. & Eaves, L. J. The Politics of Mate Choice. *The Journal of Politics* **73,** 362–379 (2011).

94. Nordsletten, A. E. *et al.* Patterns of Nonrandom Mating Within and Across 11 Major Psychiatric Disorders. *JAMA psychiatry* **73,** 354–361. ISSN: 2168-622X (Apr. 2016).

95. Robinson, M. R. *et al.* Genetic Evidence of Assortative Mating in Humans. en. *Nature Human Behaviour* **1,** 0016. ISSN: 2397-3374 (Jan. 2017).

96. Yengo, L. *et al.* Imprint of Assortative Mating on the Human Genome. En. *Nature Human Behaviour* **2,** 948. ISSN: 2397-3374 (Dec. 2018).

97. Elston, R. C., Buxbaum, S., Jacobs, K. B. & Olson, J. M. Haseman and Elston revisited. fr. *Genetic Epidemiology* **19,** 1–17. ISSN: 1098-2272 (2000).

98. Haseman, J. K. & Elston, R. C. The Investigation of Linkage between a Quantitative Trait and a Marker Locus. eng. *Behavior Genetics* **2,** 3–19. ISSN: 0001-8244 (Mar. 1972).

99. Patterson, H. D. & Thompson, R. Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika* **58,** 545. ISSN: 00063444 (Dec. 1971).

100. Jiang, J. REML Estimation: Asymptotic Behavior and Related Topics. *The Annals of Statistics* **24,** 255–286. ISSN: 0090-5364 (1996).

101. Bulik-Sullivan, B. *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. en. *Nature Genetics* **47,** 1236–1241. ISSN: 1546-1718 (Nov. 2015).

102. Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. en. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh* **52,** 399–433. ISSN: 2053-5945, 0080-4568 (1919/ed).

103. Crow, J. F. & Felsenstein, J. The Effect of Assortative Mating on the Genetic Composition of a Population. eng. *Eugenics Quarterly* **15,** 85–97. ISSN: 0097-2762 (June 1968).

104. Nagylaki, T. The Correlation between Relatives with Assortative Mating. eng. *Annals of human genetics* **42.** Citation Key Alias: NagylCorrelation1978a, NagylCorrelation1978b, 131–137. ISSN: 0003-4800 (July 1978).

105. Gimelfarb, A. A General Linear Model for the Genotypic Covariance between Relatives under Assortative Mating. en. *Journal of Mathematical Biology* **13,** 209–226. ISSN: 1432-1416 (Dec. 1981).

106. Nagylaki, T. Assortative Mating for a Quantitative Character. en. *Journal of Mathematical Biology* **16,** 57–74. ISSN: 1432-1416 (Dec. 1982).

107. Gimelfarb, A. Quantitative Characters under Assortative Mating: Gametic Model. *Theoretical Population Biology* **25,** 312–330. ISSN: 0040-5809 (June 1984).

108. Gimelfarb, A. Is Offspring—Midparent Regression Affected by Assortative Mating of Parents? en. *Genetics Research* **47,** 71–75. ISSN: 1469-5073, 0016-6723 (Feb. 1986).

109. Lee, J. J. *et al.* Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals. En. *Nature Genetics* **50,** 1112. ISSN: 1546-1718 (Aug. 2018).

110. Yengo, L. & Visscher, P. M. Assortative Mating on Complex Traits Revisited: Double First Cousins and the X-Chromosome. *Theoretical Population Biology* **124,** 51–60. ISSN: 0040-5809 (Dec. 2018).

111. Golan, D., Lander, E. S. & Rosset, S. Measuring Missing Heritability: Inferring the Contribution of Common Variants. en. *Proceedings of the National Academy of Sciences* **111,** E5272–E5281. ISSN: 0027-8424, 1091-6490 (Dec. 2014).

112. Jiang, J., Li, C., Paul, D., Yang, C. & Zhao, H. On High-Dimensional Misspecified Mixed Model Analysis in Genome-Wide Association Study. en. *The Annals of Statistics* **44,** 2127–2160. ISSN: 0090-5364 (Oct. 2016).

113. Jiang, J. *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems* (Chapman and Hall/CRC, 2017).

114. Hunter, J. K. & Nachtergaele, B. *Applied Analysis* (World Scientific Publishing Company, 2001).

115. Heil, C. *Metrics, Norms, Inner Products, and Operator Theory* (Springer, 2018).

116. Consortium, T. 1. G. P. A Global Reference for Human Genetic Variation. en. *Nature* **526,** 68–74. ISSN: 1476-4687 (Oct. 2015).

117. Chang, C. C. *et al.* Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. English. *GigaScience* **4,** 7. ISSN: 2047-217X (2015).

118. Loh, P.-R. *et al.* Reference-Based Phasing Using the Haplotype Reference Consortium Panel. en. *Nature Genetics* **48,** 1443–1448. ISSN: 1546-1718 (Nov. 2016).

119. Polaczyk, B. Concentration of the Empirical Spectral Distribution of Random Matrices with Dependent Entries. en. *arXiv:1809.05393 [math].* arXiv: `1809.05393 [math]` (Sept. 2018).

120. Seber, G. A. *A Matrix Handbook for Statisticians* (John Wiley & Sons, 2008).

121. Lewis, T. O. & Newman, T. G. Pseudoinverses of Positive Semidefinite Matrices. *SIAM Journal on Applied Mathematics* **16,** 701–703. ISSN: 0036-1399 (1968).

Appendix A

Quadratic form lemma

Let A^+ denote any generalized inverse and let $A^\dagger$ denote the Moore-Penrose pseudoinverse.

Fact A.1 (Seber [120], page 457)**.** *Let $S \in \mathbb{R}^{n\times n}$ be an orthogonal projection operator. Then for any $B \in \mathbb{R}^{m\times n}$,*

$$S(BS)^\dagger = (BS)^\dagger$$

and

$$(SB)^\dagger S = (SB)^\dagger$$

Fact A.2 ([121], Theorem 6)**.** *Let $A \in \mathbb{R}^{n\times n}$ be such that $A = A^T \succeq 0$, $C \in \mathbb{R}^{n\times c}$, and define $\tilde{A} = A + C^T C$. Let $A^+ \succeq 0$ be a generalized inverse of A. Then*

$$\tilde{A}^+ = A^+ - A^+C^T(I + CA^+C^T)^{-1}CA^+$$

is a generalized inverse of $\tilde{A}$ if and only if $\ker A \subseteq \ker C$*. Further, if this is the case, we have*

$$\begin{gathered}\tilde{A}^+\tilde{A} = A^+A,\\ \tilde{A}\tilde{A}^+ = AA^+,\\ \tilde{A}^+ \succeq 0,\\ \textit{and } \tilde{A}^+ = \tilde{A}^\dagger \iff A^+ = A^\dagger.\end{gathered}$$

Lemma A.3. *Write the full QR decomposition of $X \in \mathbb{R}^{n\times c}$ as $X = QR = [Q_X|Q_{X^\perp}]R$ such that $Q_X \in \mathbb{R}^{n\times(n-c)}$, $Q_{X^\perp} \in \mathbb{R}^{n\times c}$ and define $S = I_n - Q_XQ_X^T$. Then,*

$$y^TS(S(H+\sigma I)S)^\dagger Sy = y^TS(SHS+\sigma I)^{-1}Sy$$

Proof. Applying the first fact,

$$y^T S(S(H + \sigma I)S)^\dagger Sy = y^T (S(H + \sigma I)S)^\dagger y.$$

Rewrite the left hand side as

$$\begin{aligned} y^T (S(H + \sigma I)S)^\dagger y &= \sigma^{-1} y^T (\underbrace{\sigma^{-1} SHS}_{\equiv A} + SS)^\dagger y \\ &= \sigma^{-1} y^T \tilde{A}^\dagger y, \end{aligned}$$

where we define $\tilde{A} = A + S^T S$. Note that the first fact also yields

$$SA^\dagger S = SA^\dagger = A^\dagger S = A^\dagger$$

Using the second fact, we then have

$$\begin{aligned} \sigma^{-1} y^T \tilde{A}^\dagger y &= \sigma^{-1} y^T (A^\dagger - A^\dagger S(I + SA^\dagger S)^{-1} SA^\dagger) y \\ &= \sigma^{-1} y^T (A^\dagger - A^\dagger (I + A^\dagger)^{-1} A^\dagger) y \\ &= \sigma^{-1} y^T (A + I)^{-1} y \\ &= y^T S(SHS + \sigma I)^{-1} Sy \end{aligned}$$

□