Assortative Mating Biases Marker-based Heritability Estimators

Richard Border, Ph.D.[1†], Sean O'Rourke, Ph.D.[2], Teresa de Candia, Ph.D.[3,4], Michael E. Goddard, Ph.D.[5,6], Peter M. Visscher[7], Ph.D., Loic Yengo, Ph.D.[7], Mathew Jones, Ph.D.[3‡], Matthew C. Keller, Ph.D.[3,4†‡]

[1]Departments of Neurology and Computer Science, University of California, Los Angeles
[2]Department of Mathematics, University of Colorado Boulder
[3]Department of Psychology and Neuroscience, University of Colorado Boulder
[4]Institute for Behavioral Genetics, University of Colorado Boulder
[5]Faculty of Veterinary and Agricultural Science, University of Melbourne
[6]Biosciences Research Division, Department of Economic Development, Jobs, Transport and Resources
[7]Institute for Molecular Bioscience, University of Queensland
[†]Corresponding authors
[‡]Equally contributing authors

Abstract

Many complex traits are subject to assortative mating (AM), with recent molecular genetic findings confirming longstanding theoretical predictions that AM alters genetic architecture by inducing long range dependence across causal variants. However, all marker-based heritability estimators assume mating is random. We provide mathematical and simulation-based evidence demonstrating that both method-of-moments estimators and likelihood-based estimators produce biased estimates in the presence of AM and that common approaches to account for population structure fail to mitigate this bias. Then, examining height and educational attainment in the UK Biobank, we demonstrate that these biases affect real world traits. Finally, we derive corrected heritability estimators for traits under equilibrium AM.

Introduction

Primary phenotypic assortative mating (hereafter simply "AM"; the phenomenon whereby mate-choice is based on phenotypic similarity) has been observed for a variety of heritable traits in human and non-human animals [1–4]. A century ago, Fisher demonstrated that AM induces long-range positive correlations between trait-increasing allele counts at causal loci across the genome, thereby increasing genetic variance across successive generations until it approaches a stable equilibrium [5]. Since Fisher's time, it has been established that many human traits are subject to AM, and that estimates of genetic and environmental variance from twin and family designs, which assume random mating, can be biased in the presence of AM [6]. However, to date there has been no study of how AM influences marker-based heritability estimators. Moreover, many traits that have been focal in the scientific discourse regarding the so-called still-missing heritability—for example, height and educational attainment—are precisely those traits for which both phenotypic and genetic data is consistent with primary phenotypic assortment [4, 7–9], further motivating the need to understand how AM influences marker-based heritability estimates.

Here, we address this gap in knowledge by characterizing the impact of AM on two major families of marker-based heritability estimators: method of moments estimators (MoM; typified by univariate Haseman-Elston (HE) regression [10] but also including PCGC regression and LD score regression [LDSC] [11, 12]), and residual maximum likelihood (REML [13]; typified by GCTA and BOLT-REML [14, 15]). We assume Fisher's classical model of AM, which describes the equilibrium properties of a heritable trait for which mates' genotypes are conditionally independent given their phenotypes, and which has formed the theoretical foundation for recent investigations of AM using measured genetic data [4, 8]. We provide mathematical and simulation-based arguments demonstrating that AM induces a modest but nevertheless non-negligible bias in both classes of estimators that is not addressed by conventional methods of accounting for population structure. In the process, we extend results in random matrix theory and classical quantitative genetics by characterizing the higher-order moments of causal variants and the limiting spectral distribution of the genomic relatedness matrix (GRM) under AM, thereby providing intuition with respect to the observation that genomic principal components do not capture the effects of AM. Additionally, we provide empirical results using data from the UK Biobank that are congruent with our theoretical predictions regarding the influence of AM on

marker-based heritability estimators. Finally, we provide guidelines for using and interpreting the results of marker-based heritability estimators when applied to traits subject to AM.

Results

Our theoretical results depend on several key parameters: r denotes the phenotypic correlation between mates on a phenotype Y; h_0^2 denotes the panmictic heritability, what the heritability of the phenotype would be in the absence of AM; h_∞^2 denotes the true equilibrium heritability under AM; and Z denotes the $n \times m$ matrix of n unrelated individuals' standardized genotypes at m causal loci with effects vector u. We initially assume that all causal variants are present in Z (i.e., that all the narrow-sense heritability is explained by measured variants); we later relax this condition. The rows of Z (individuals' genotypes) are independent random vectors with $m \times m$ covariance matrix Υ, which quantifies the correlation between loci. Under random mating, $\Upsilon := \Upsilon_0$ is approximately block diagonal such that causal variants are largely (aside from linkage disequilibrium between nearby variants) stochastically independent. However, under equilibrium AM, $\Upsilon := \Upsilon_\infty$ is dense due to the presence of positive long-range correlations among trait-increasing allele counts within and across chromosomes. As the elements of Υ_∞ agree in sign with the corresponding elements of uu^T (i.e., trait increasing alleles are positively correlated), the equilibrium genetic variance under AM, $\sigma_{g,\infty}^2$, is considerably greater than the panmictic genetic variance $\sigma_{g,0}^2$. That is, $\sigma_{g,\infty}^2 = u^T \Upsilon_\infty u > \sigma_{g,0}^2 \approx u^T u$ (Figure 1).

Haseman-Elston regression estimates under AM

We first derive the influence of AM on heritability estimates from HE regression [10], which is perhaps the simplest MoM marker-based heritability estimator. Let $\tilde{y}$ denote the standardized (zero-mean unit-variance) phenotype. The HE regression estimator $\hat{h}_{\text{HE}}^2$ of h^2 is the slope of the subdiagonal elements of the phenotypic outer product, $\tilde{y}\tilde{y}^T$, regressed on the subdiagonal elements of the GRM, $m^{-1}ZZ^T$. We demonstrate that $\hat{h}_{\text{HE}}^2$ is upwardly biased relative to both h_0^2 and h_∞^2 under AM. Intuitively, this is because the phenotypic outer product accurately reflects increases in genetic variance due to positive associations across all pairs trait increasing alleles, whereas the effect of AM on the GRM, the elements of which represent individuals' average similarity at homologous diploid loci but not across distinct pairs of loci, is negligible. This latter point is itself notable as some studies have erroneously claimed that AM leads to detectable increases in average

genomic similarity between mates [17], but the actual increase is trivial (of order $O(m^{-1})$ relative to the increase in genetic variance) and all but undetectable for highly polygenic traits [18]. With respect to the influence of AM on HE regression estimates, and assuming that all the narrow-sense heritability is explained by measured variants, we establish the following general result:

$$\mathbb{E}\big[\hat{h}_{\mathrm{HE}}^2\big] = \left(\frac{m \cdot u^T \Upsilon_\infty^2 u}{\sigma_{g,\infty}^2 \cdot \mathrm{tr}[\Upsilon_\infty^2]}\right) h_\infty^2 \geq h_\infty^2. \tag{1}$$

Under AM, the bracketed quantity is greater than one, and thus $\hat{h}_{\mathrm{HE}}^2$ is biased upwards. Under the stricter assumption of exchangeable loci (i.e., each causal variant explains equal variance in the phenotype), we derived the following approximate expression, dependent only on the panmictic heritability and phenotypic correlation between mates:

$$\mathbb{E}\big[\hat{h}_{\mathrm{HE}}^2\big] \approx \left(\frac{1}{1 - r_{g,\infty}}\right) h_\infty^2, \tag{2}$$

where $r_{g,\infty} := r \cdot h_\infty^2$ is the equilibrium genetic correlation between mates (Figure 1; see Supp. Materials S3.1 for greater details and proofs). Under exchangeable loci for known r, we define estimators of the panmictic and equilibrium heritabilities by applying the following transformations to $\hat{h}_{\mathrm{HE}}^2$:

$$\hat{h}_0^2 := \mathbb{E}\big[h_0^2 | \hat{h}_{\mathrm{HE}}^2\big] = \frac{\hat{h}_{\mathrm{HE}}^2}{1 + 2r\hat{h}_{\mathrm{HE}}^2 + r(r-1)\hat{h}_{\mathrm{HE}}^4}, \qquad \hat{h}_\infty^2 := \mathbb{E}\big[h_\infty^2 | \hat{h}_{\mathrm{HE}}^2\big] = \frac{\hat{h}_{\mathrm{HE}}^2}{1 + r\hat{h}_{\mathrm{HE}}^2}. \tag{3}$$

Large scale simulations using realistic genotype data across a variety of scenarios (see Online Methods) demonstrate that the above approximations are accurate even when the exchangeable loci assumption is violated (mean relative error across simulations = -0.009; Figure 1).

In addition, our simulations confirmed that LD score regression [12], which is mathematically equivalent to HE regression when LD scores are exact [19], is analogously biased upwards (Figure 4a). However, the impact of this bias in real world applications depends not only on the extent of AM, but also on the degree to which estimated LD scores reflect the true LD structure in a given population, and therefore no straightforward correction is available.

Residual maximum likelihood estimates under AM

In contrast to HE regression, for which the upward bias is independent of sample size, we show

that the REML estimator, $\hat{h}^2_{\text{REML}}$, exhibits large upward biases in small samples but converges towards h^2_0 (which is less than the true equilibrium value, h^2_∞) from above in large samples. Formally, under the assumption of exchangeable loci, we prove that

$$\hat{h}^2_{\text{REML}} \xrightarrow{P} h^2_0, \quad \text{as } n, m \to \infty, \tag{4}$$

where $n/m \to c \in (0, \infty)$; i.e., for polygenic traits, $\hat{h}^2_{\text{REML}}$ is a consistent estimator of h^2_0. In essence, the parameter values that maximize the residual likelihood function depend only on the eigenvalues of the GRM and the long-range dependence among causal variants induced by AM is "weak" in the sense that the distributions of the eigenvalues (i.e. spectral distributions) of the GRM under random mating and under AM are asymptotically equivalent. However, in *finite* samples, we show via simulation that this convergence can be extremely gradual, requiring samples approaching millions of individuals before estimates reasonably approach h^2_0 (Figure 2a, Figure 2b). Thus, the direction (relative to h^2_∞) and magnitude of the bias of $\hat{h}^2_{\text{REML}}$ depend on sample size in addition to the panmictic heritability and strength of AM. On the other hand, the number of causal variants, the total number of measured SNPs, and the ratios of these with sample size have no apparent influence on the bias of $\hat{h}^2_{\text{REML}}$ (Supp. Figure S1).

Conventional means of addressing population structure do not mitigate AM-induced bias

Inclusion of ancestral principal components as covariates failed to mitigate the AM-induced bias in both the MoM and the REML estimates (Figure 4a). Indeed, we demonstrate that the AM has a negligible effect on the spectral distribution of the GRM in high-dimensional settings (Supp. Materials S2.4). Similarly, these biases are not mitigated by modeling multiple genetic variance components by partitioning SNPs according to LD score and minor allele frequency (Figure 4a) or by partitioning SNPs by chromosome (results not shown).

AM-induced bias persists when not all causal variants are measured

In real-world applications, measured genotypes will often include some fraction of the total number of causal variants. To assess the impact of AM when not all variants are measured, we compared HE regression and REML heritability estimates in synthetic data including 100%, 50%, or 25% of both causal and non-causal SNPs by discarding variants at random (Figure 3b). As expected, this resulted in attenuated estimates commensurate with the fraction of missing data. On

average, heritability estimates were 8.4% (se=0.009%) and 17.6% (se=0.009%) lower after randomly discarding 50% and 75% of SNPs, respectively (the degree of attenuation was smaller than the proportion of SNPs dropped due to linkage disequilibrium between retained and discarded SNPs). Nevertheless, while estimates were lower, the pattern of bias due to AM when some of the heritability was missing appeared roughly the same as when no heritability was missing. To test if $\hat{h}^2$ attenuation varied across methods, we fixed sample size at n=128,000 and regressed $\hat{h}^2$ on the interaction between the fraction of SNPs discarded (0%, 50%, or 75%) and method (HE=1 vs. REML=0). Similarly, to test if $\hat{h}^2$ attenuation varied with sample size, we regressed $\hat{h}^2$ on the interaction between the fraction of SNPs discarded and $\log_{10}$ sample size. We found that the linear relationship between $\hat{h}^2$ and the fraction of SNPs discarded did not depend on method ($\hat{\beta}$=0.036, 95% CI: [-0.011, 0.083]) nor on $\log_{10}$ sample size ($\hat{\beta}$=0.031, 95% CI: [-0.059, 0.120]).

AM in the UK Biobank

In addition to the experiments using synthetic data described above, we sought to verify our theoretical predictions by examining the relationship between sample size and heritability estimates in a sample 335,551 unrelated European-ancestry individuals in the UK Biobank [20]. We *a priori* selected four phenotypes based on evidence (height, years of education) or lack of evidence (body mass index [BMI], bone mineral density [BMD]) for primary phenotypic AM in a previous study [8]. We then computed HE regression and REML heritability estimates in pairs of small (n=16,000) versus large (n=N–16,000) non-overlapping subsamples, where N depended on the available sample size for each phenotype (Table 1). Congruent with theoretical expectations and simulation results, REML and HE estimates diverged with increasing sample size for height and years of education such that the differences between REML estimates in smaller versus larger subsamples were larger than those of HE estimates (mean difference of differences $\bar{\delta}$:=mean δ_{REML}-δ_{HE}=0.024, p=5.24e-4; $\bar{\delta}$=0.015, p=3.94e-4; respectively), but not for BMI or BMD ($\bar{\delta}$=5.84e-3, p=9.42e-2; $\bar{\delta}$=-3.41e-3, p=0.302; respectively; Figure 3). This was consistent with a previous report that quantified the degree of AM for various traits by correlating polygenic scores between odd- versus even-numbered chromosomes also found that height and educational attainment, but not BMI or BMD, showed signatures of AM [8]. Unlike this previous approach, however, our approach is agnostic to variant direction and effect size. Thus, applying the REML estimator across subsamples of varying sizes provides an alternative and independent way to detect

genomic signatures of AM.

We used previously reported estimates of spousal correlations [2, 4] for height and educational attainment in order to correct HE regression heritability estimates (Table 1) via the estimators in Equation (3). Relative to the corrected equilibrium SNP-heritability estimates, HE regression estimates were inflated by 14% and 7% for height and years of education, respectively.

Discussion

Summary of findings

Despite the long-standing understanding that AM alters the genetic architecture of heritable traits, abundant evidence that many phenotypes are subject to AM, and concentrated research activity in marker-based variance component estimation, the effects AM has on these estimators has remained unknown. In the present investigation, we demonstrated that AM biases heritability estimates and that these biases behave differently for MoM estimators versus REML estimators as a function of sample size. In the process, we extended previous results in quantitative genetics and random matrix theory by characterizing the full equilibrium joint distribution of causal variants, demonstrating that the empirical spectral distribution of the resulting relatedness matrix converges to the Marčenko-Pastur law, and thereby proving that REML produces a consistent estimator of the panmictic heritability of polygenic traits in very large samples (see Supp. Materials S2, S3). However, REML estimates of heritability of traits subject to AM behave peculiarly in finite samples, decreasing with larger samples and yielding estimates greater than the true equilibrium heritability in sample sizes typical of those published in the literature (Figure 2b). On the other hand, MoM estimators yield upwardly biased estimates that are higher than REML estimates and remain stable across sample sizes (Figure 2a, Figure 4a). Using UK Biobank data, we observed this differential behavior of estimates for two traits that with previous evidence for AM but not for two negative control traits.

Implications

Researchers have previously argued that the impact of AM on the heritability of common traits is likely to be "at most modest" [21]. Our results speak to a related but distinct phenomenon: the impact of AM on *heritability estimators* that use molecular genetic data. Notably, this impact is likely to be most salient for "benchmark" traits like height, which has served as a focal point for

the discourse surrounding the missing and still-missing heritability phenomena [9, 22]. Our results suggest that the still-missing heritability may be somewhat larger than currently thought for traits subject to AM. For instance, for a trait subject to AM, the REML estimator will produce higher heritability estimates when applied whole-genome sequence data collected from a smaller number of individuals relative to estimates derived from larger samples for whom only sparse array data are available. On the other hand, twin studies are expected to underestimate the total heritability of an additive trait subject to AM [23], implying that the true discrepancy between the total heritability and that which is currently discernible from dense marker data may be greater than previously believed. Our simulations demonstrate that differences in REML heritability estimates derived in samples of 32,000 versus 128,000 individuals are on the order of those induced by randomly omitting 75% of measured SNPs (Figure 2a , Figure 4a). As such, caution is warranted when comparing heritability estimates across methods or across sample sizes for traits subject to AM.

Interpreting heritability estimates in the presence of AM

As we have demonstrated, existing methods do not produce unbiased heritability estimates for traits subject to primary phenotypic assortment in real world use cases. Although we provide unbiased estimators of both the equilibrium and panmictic heritabilities (by correcting $\hat{h}^2_{\text{HE}}$ to obtain $\hat{h}^2_\infty$ and $\hat{h}^2_0$ as per Equation [3]), these rely on the strong assumption of equilibrium. Likewise, though REML provides an asymptotically unbiased estimator of the panmictic heritability, the sample sizes required to approach unbiasedness are currently unavailable and computationally impractical. Still, it is possible to derive theoretically-sound bounds on the true values of the present day and panmictic heritabilities using existing methods applied to realistic sample sizes in disequilibrium. Specifically, as a population approaches equilibrium over successive generations of assortative mating, the current heritability at generation t, h^2_t, is bounded in expectation from above by $\hat{h}^2_{\text{HE}}$, with equality at $t = 0$ (panmixis), and from below by $\hat{h}^2_\infty = \mathbb{E}[h^2_\infty | \hat{h}^2_{\text{HE}}; t = \infty]$, with equality as $t \to \infty$ (equilibrium). Likewise, the panmictic heritability h^2_0 is bounded in expectation from below by $\hat{h}^2_0 = \mathbb{E}[h^2_0 | \hat{h}^2_{\text{HE}}; t = \infty]$, with equality as $t \to \infty$, and from above by $\hat{h}^2_{\text{REML}}$ (with equality at all generations as $n \to \infty$). Thus, as long as the strength of AM isn't decreasing across generations, $[\hat{h}^2_\infty, \hat{h}^2_{\text{HE}}]$ provide probabilistic bounds encompassing the true present day heritability, and $[\hat{h}^2_0, \hat{h}^2_{\text{REML}}]$ provide probabilistic bounds encompassing the true

panmictic heritability under disequilibrium with $\hat{h}_\infty^2$ and $\hat{h}_0^2$ providing conservative point estimates of h_∞^2 and h_0^2, respectively (Figure 1).

Limitations and future directions

There are several limitations of the current approach. First among these are assumptions inherent to the primary phenotypic AM model. Some of these assumptions, including equilibrium and constancy of the phenotypic mating correlations across generations, provide mathematical tractability and are to some extent inessential to the resulting phenomena. For example, while the problem of characterizing the joint distribution of causal variants becomes substantially more difficult in a population subject to AM that has not reached equilibrium, we observed (results not shown) that estimators behave in a similar, albeit less extreme, fashion relative to their behavior in an equilibrium population. Other assumptions, such as the absence of gene-environment correlation and the conditional independence of mates' genotypes given their phenotypes (which may be violated in structured populations), are more difficult to evaluate and deserve consideration in future investigations. Additional limitations pertain to our theoretical analysis of the REML estimator, which is rooted in a high-dimensional asymptotics framework. The exact causes of the peculiar behavior of the REML estimator in finite samples, particularly regarding the slow rate of convergence to the panmictic heritability, remain unclear to us. Whether this problem is addressable from an asymptotic perspective or instead requires an alternative, non-asymptotic framework, is an open question and is a target for future theoretical work. Finally, when the phenotypic correlation between mates is known, our results provide a prescription for rectifying AM-induced biases (by adjusting $\hat{h}_{HE}^2$ per Equation [3]) under the assumption of equilibrium, which is difficult to verify, and only provide broad bounds on the extent of said biases for populations at disequilibrium. As such, these results should provide motivation and a starting place for the development of new methods that can provide unbiased estimates of genomic variance in the presence of AM.

Online Methods

Theoretical framework

The primary phenotypic assortment model

Here we introduce the model of AM as proposed by Fisher [5] and further developed by Nagylaki

and others [16, 24] (see Supp. Materials S1 for a detailed exposition). Briefly, we consider a phenotype as a random vector composed of independent heritable and non-heritable components:

$$Y = Zu + E, \qquad E \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2), \tag{5}$$

where the rows of Z, representing individuals' standardized genotypes, are independent m-dimensional random vectors following a multivariate discrete distribution with finite moments and finite and which we assume are independent under panmixis. The vector of allele substitution effects u, which we treat as fixed, is such that $u^T u = \sigma_{g,0}^2$. Further, we assume that 1) parent-parent-offspring trios' phenotypes are jointly Gaussian; 2) the phenotypic correlation between mates, r is constant across generations; and 3) there exists $c_0 \in (0, \infty)$ such that $\max\limits_{k=1,...,m} |u_k| \le c_0 \cdot m^{-1/2}$; that is, as traits become increasingly polygenic, the maximal variance attributable to individual variants decreases commensurately.

The equilibrium distribution of causal variants

Over successive generations, the correlation between mates' phenotypes induces positive correlations across trait increasing allele counts independent of physical position on the genome and thereby increases the total genetic variance of the trait. The genetic variance rapidly approaches a stable equilibrium after several generations (typically within ten generations), at which point the within-individual and cross-mate correlations among causal variants are equal to one another. Using the results of Nagylaki [16], we can express the equilibrium covariance matrix between causal variants as a low rank perturbation of a diagonal matrix of the form: $\Upsilon_\infty = D + 2\phi\phi^T$, where ϕ is a known vector-valued function of the substitution effects and mating correlation (Supp. Materials S1.2) with elements $\phi_k = O(m^{-1/2})$ uniformly.

Higher order moments and the limiting spectral distribution of GRM

Employing tools from the study of thermodynamic equilibria, we extend these classical results to bound moments of higher orders (Supp. Materials S2.1, Supp. Materials S2.4). Using these results, we extend the widely-known Marčenko-Pastur theorem, which describes the limiting distribution of the spectrum of sample covariance matrices corresponding to random matrices with independent sub-Gaussian elements [25], to the case of random matrices with independent rows meeting particular moment conditions (Supp. Materials S2.3). Together, these results establish the limiting spectral distribution of the sample GRM (i.e., the distribution of the eigenvalues of the sample

GRM as both sample size and the number of variants examined become large) under AM (Supp. Materials S2.4), providing the necessary theoretical foundation to characterize the asymptotic behavior of the REML estimator. Further, these results explain why controlling for principal components fails to remove AM-induced biases: the impact of AM on the spectrum of the GRM is asymptotically negligible.

Haseman-Elston regression under AM

The HE regression heritability estimator [26] is obtained by regressing the subdiagonal elements of the standardized phenotypic outer product $\tilde{y}\tilde{y}^T$ on the subdiagonal elements of the GRM $m^{-1}ZZ^T$. Whereas elements of the outcome (the phenotypic outer product) reflect the dependences among all pairs of causal loci:

$$\mathbb{E}\big[\{\tilde{y}\tilde{y}^T\}_{i,j<1}\big] \propto \sum_{k=1}^{m}\sum_{l=1}^{m} u_k u_l \mathbb{E}\big[z_{ik}z_{jl}\big] + e_i e_j, \tag{6}$$

elements of the GRM only capture the dependences among haploid loci that at the same diploid site:

$$\mathbb{E}\big[\{ZZ^T\}_{i,j<1}\big] \propto \sum_{k=l}^{m} \mathbb{E}\big[z_{ik}z_{jl}\big]. \tag{7}$$

As a result, the variance of the outcome increases whereas the variance of predictor remains largely unaffected, leading to overestimation of the true equilibrium heritability, potentially producing estimates greater than one for strong assortment (Figure 1; see Supp. Materials S3.1 for formal further details and proof). In contrast to the REML estimator, the HE regression estimator is upwardly biased irrespective of sample size (Figure 2a).

REML and the spectrum of the GRM under AM

The REML estimator [13] models the phenotype as a random vector with marginal distribution,

$$y \sim \mathcal{MVN}\big(X\beta, m^{-1}ZZ^T\sigma_g^2 + I\sigma_e^2\big), \tag{8}$$

where X is an $n \times c$ matrix of covariates with fixed effects β and the covariance structure is comprised of a heritable component (σ_g^2 times the GRM) and a non-heritable component (σ_e^2 times the identity). The heritability estimator $\hat{h}_{\text{REML}}^2 = \hat{\sigma}_g^2/(\hat{\sigma}_g^2 + \hat{\sigma}_e^2)$ is derived by finding the values of the variance components that satisfy the equation,

$$\nabla\ell\big(\hat{\sigma}_g^2, \hat{\sigma}_e^2 | A^T y\big) = 0, \tag{9}$$

where ℓ denotes the marginal log likelihood of the transformed random variable $A^T y$ for $A^T: \mathbb{R} \to$

$(\text{col } X)^{\perp} \subseteq \mathbb{R}^{n-c}$, $A^T A = I$. The conditional expectation of $\nabla \ell$ given the genotypes is a function of the eigenvalues of the GRM and, as a result, the asymptotic behavior of $\hat{h}^2_{\text{REML}}$ is governed by the asymptotic distribution of the eigenvalues of $m^{-1} Z Z^T$. A foundational result in random matrix theory states that for zero-mean unit-variance sub-Gaussian random matrices $W \in \mathbb{C}^{n \times m}$ with independent elements, the empirical spectral distribution function of $m^{-1} W W^T$ converges almost surely to the Marčenko-Pastur distribution [25]. Employing this result, Jiang and colleagues [27] demonstrated that, in the case of independent causal variants, REML consistently estimates the true heritability in high dimensional settings and is robust to certain forms of model misspecification. In Supp. Materials S2.3, Supp. Materials S2.4, we demonstrate that even though AM induces dependence among causal variants, this dependence is "weak" in the sense that it doesn't change the limiting spectral distribution of the GRM, thereby allowing us to apply arguments in line with those of Jiang and colleagues' (Supp. Materials S3.2). Intuitively our result can be summarized as follows: as the sample size and the number of causal variants become large, the eigenvalues of the GRM under AM behave as if the causal variants were independent (as is largely the case under random mating). The behavior of the REML estimator is determined by the behavior of the eigenvalues of the GRM, and thus $\hat{h}^2_{\text{REML}}$ converges to what the heritability would be if the causal variants were independent, i.e., the panmictic heritability.

Simulation studies

We employed a realistic forward-time simulation framework to generate genotypic and phenotypic data. We then used these data to motivate and verify theoretical results. Below, we describe the general framework and specific simulations we performed.

Simulation framework

Given a recombination map and n_{input} individuals' phased biallelic genotypes at p diploid loci as input, we divided the genome into $k \ll p$ contiguous-within-chromosome, non-overlapping 50kB intervals to obtain a block representation. Recombination events, which occurred with probabilities dictated by the recombination map, were restricted to interval boundaries, thus dramatically reducing the number of haplotypes that had to be tracked while maintaining high genomic resolution. To achieve a target population size $N_{\text{sim}} > n_{\text{input}}$, N_{sim} pairs of the n_{input} individuals were non-monogamously 'mated' (i.e., matched and subject to meiosis), resulting in a new generation of N_{sim} individuals whose genomes were could be represented in terms of the which of

the n_{input} haplotype blocks they inherited at each of the k intervals. We then repeated this random mating procedure for an additional five generations, resulting in N_{sim} chimeric combinations of the original n_{input} genotypes while maintaining the linkage disequilibrium structure of the original data. These discretized genotypes comprised the input for the principal AM simulations.

At the beginning of each particular AM simulation with prespecified panmictic heritability $h_0^2 = \sigma_{g,0}^2/(\sigma_{g,0}^2 + \sigma_e^2)$, phenotypic correlation between mates r, p SNPs, and m diploid causal loci $z_1, \ldots, z_m$, $m \ll p$, the standardized allele substitution effects $u_1, \ldots, u_m$, were independently drawn from a Gaussian distribution with expectation zero and variance $\sigma_{g,0}^2/m$. Unless otherwise stated, all simulations used $p = 10^6$ SNPs. At each generation, phenotypes were constructed via $y = Zu + e$ where e was i.i.d. Gaussian with zero expectation and variance σ_e^2. Next, mates were matched according to their respective phenotypes y_i, y_j such that $\text{corr}(y_i, y_j) \approx r$. This was achieved by drawing N_{sim} independent doubles $\{(w^*, w^{**})^T\}_{k=1}^{N_{\text{sim}}} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right)$ from which N_{sim} pairs of indices $\{(i,j)\}_{k=1}^{N_{\text{sim}}}$ were constructed such that $(i,j)_k$ were the positions of w_k^* and w_k^{**} after concatenating and sorting each element of $\{(w^*, w^{**})^T\}_{k=1}^{N_{\text{sim}}}$. Similarly, N_{sim} indices $l = l_1, \ldots, l_{N_{\text{sim}}}$ were constructed such that l_k indexed the kth largest of the N_{sim} simulated phenotypes. Finally, each kth mating pair was determined by taking the $l_{\lfloor i_k/2 \rfloor}$th and $l_{\lfloor i_j/2 \rfloor}$th replicates. Having chosen mates, meiosis occurred as previously detailed to construct the next generations' genotypes.

Simulations using UK Biobank data

For each simulation, the input data were derived from phased, imputed genotypes at $p = 10^6$ randomly selected imputed SNP loci in a sub-sample of $n_{\text{input}} = 435,301$ European UK Biobank participants [20]. All SNPs were chosen to meet the following criteria: minor allele frequency greater than 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) reference panel [28]. Genotype data were then phased to the 1KG3 reference panel in batches of 40,000 individuals using Eagle v2.4 [29]. This input data was then grown to a population of $N_{\text{sim}} = 10^6$ chimeric genotypes and subjected to an additional five generations of random mating as described in the preceding section.

We conducted AM simulations for varying mating correlations, $r \in \{0, .25, .5, .75\}$ and numbers

of causal variants, $m \in \{10^4, 10^5\}$, with panmictic heritability fixed at $h_0^2 = .5$. Each simulation consisted of fifteen generations of AM and produced results congruent with classical theory. Prior to heritability estimation, close relatives $\hat{\pi} \geq .05$, were removed using GCTA v1.93.1 [14], resulting in an average sample size of 141,667 across simulated datasets. Additionally, we ran a limited number of larger, more computationally intensive simulations ($N_{\text{sim}} = 3 \times 10^6$) with mating correlations fixed at $r = .5$ to investigate the large sample behavior of the REML estimator, resulting in at least 648,000 unrelated individuals across simulated datasets. There were no apparent differences across simulations as a function of the number of causal variants or the simulated population size.

Heritability estimation in simulated data

We split each simulated genotype-phenotype dataset into collections of random subsamples mutually exclusive within collection but not across collections, yielding 16 samples of 16,000 individuals, 8 samples of 16,000, 4 samples of 32,000 individuals, 2 samples of 64,000 individuals, and 1 sample of 128,000 individuals. We then performed HE regression and single-component REML for each subsample (Figure 2a). We used GCTA v1.91.3b [14] to construct genomic related matrices and perform HE regression. We obtained REML heritability estimates using BOLT-LMM v2.3.4 [30] for computational efficiency; though BOLT-LMM uses a randomized algorithm, its numerical accuracy is comparable to that of the exact algorithm implemented GCTA [31].

We also performed a variety of supplementary analyses for a limited set of simulation parameters ($r = .5$, $h_0^2 = .5$, and $m \in \{10^4, 10^5\}$, $N = 10^6$). To demonstrate that including genomic principal components (PCs) as covariates does not mitigate the impact of AM, we included 10 PCs as covariates in the HE regression and REML analyses. For the former, HE regression was conducted in LDAK v5.0 [32], as the HE regression implementation in GCTA cannot accommodate covariates. To demonstrate that the behavior of LD score regression under AM is equivalent to that of HE regression (assuming that the LD scores accurately reflect the LD structure of the sample), we used PLINK v1.9 [33] to obtain GWAS summary statistics and LDSC v1.0.1 [12] to estimate within-sample LD scores using a one centiMorgan sliding window and to perform LD score regression (Figure 4a). To demonstrate that multiple variance component (also known as partitioned approaches [34, 35]) do not mitigate the impact of AM, we fit multicomponent HE regression and REML after partitioning SNPs by minor allele frequency and LD score (Figure 4a).

Finally, to assess the scenario wherein a non-trivial fraction of causal variants aren't included in the model, we estimated HE regression and REML models after removing 50% or 75% of simulated SNPs at random (Figure 4b).

Empirical results

Sampling procedures

We analyzed 1,211,273 biallelic 1KG3 SNPs with in-sample minor allele frequency greater than 0.01, Hardy-Weinberg p-value greater than 10^{-6}, and INFO scores of at least 0.95, in a sample of 335,551 unrelated European UK Biobank participants [20]. We selected phenotypes *a priori* on the basis of previous evidence for AM; we chose height ($n = 335,551$) and years of education ($n = 331,480$) as traits with previous evidence of AM, whereas we chose BMI ($n = 335,551$) and BMD ($n = 191,330$) as negative control traits [8]. We measured years of education following the procedures detailed in [36].

Analysis/resampling

We tested for evidence of AM by comparing HE and REML heritability estimates in small and large samples. In the presence of AM, our theoretical results imply that HE regression estimates are consistent across all small and large subsamples, whereas REML estimates should decrease with increasing sample size. On the other hand, in the absence of AM, neither HE nor REML estimates should systematically vary with sample size. To this end, we randomly selected ten mutually exclusive subsamples of 16,000 individuals for each trait and compared HE and REML estimates in each subsample to the non-overlapping complementary subsample comprised of the remaining $n - 16,000$ individuals, controlling for sex, age, genotyping batch, testing center, and the first 10 genomic ancestry principal components. To eliminate variance in heritability estimates due to chance differences in covariate effect estimates across subsamples, we adjusted genotypes and phenotypes in the full sample prior to all following analyses. To our knowledge, existing software is incapable of efficient REML analysis using adjusted genotypes (analogous to dosages) in large samples; e.g., BOLT-REML requires hard-calls as input, whereas GCTA and LDAK have cubic complexity in the number of individuals and markers and would require multiple weeks to run on a high thread-count server. We therefore utilized a modified Python implementation of the REML algorithm presented in [31] (code available by request). We used LDAK 5.0 to obtain adjusted HE regression estimates [32]. In order to quantify the divergence of the REML and HE

estimators in large versus small samples we performed the following test, analogous to a t test of the interaction effect in a 2×2 within-subjects experimental design:

$$H_0: \bar{\delta} := \left(\hat{h}^2_{\text{REML}}(S) - \hat{h}^2_{\text{REML}}(S^C) \right) - \left(\hat{h}^2_{\text{HE}}(S) - \hat{h}^2_{\text{HE}}(S^C) \right) = 0,$$
$$H_1: \bar{\delta} \neq 0,$$

where S denotes a given subsample and S^C its complementary subsample. Though this procedure accounts for the dependence among estimates derived in the same subsamples, the individual observations were derived from various partitionings of the same data and do not constitute independent observations. This limitation is not easily avoidable, and the results of this procedure (Figure 3) should be interpreted as descriptive despite our application of inferential procedures.

References

1. Jiang Y, Bolnick DI, Kirkpatrick M (2013) Assortative Mating in Animals. Am Nat 181:E125–E138

2. Stulp G, Simons MJP, Grasman S, Pollet TV (2017) Assortative mating for human height: A meta-analysis. Am J Hum Biol. https://doi.org/10.1002/ajhb.22917

3. Nordsletten AE, Larsson H, Crowley JJ, Almqvist C, Lichtenstein P, Mataix-Cols D (2016) Patterns of Nonrandom Mating Within and Across 11 Major Psychiatric Disorders. JAMA Psychiatry 73:354–361

4. Robinson MR, Kleinman A, Graff M, et al (2017) Genetic evidence of assortative mating in humans. Nat Hum Behav 1:0016

5. Fisher RA (1919) The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth Environ Sci Trans R Soc Edinb 52:399–433

6. Eaves L (1979) The use of twins in the analysis of assortative mating. Heredity 43:399–409

7. Young AI (2019) Solving the missing heritability problem. PLOS Genet 15:e1008222

8. Yengo L, Robinson MR, Keller MC, et al (2018) Imprint of assortative mating on the human genome. Nat Hum Behav 2:948

9. Wainschtein P, Jain DP, Yengo L, et al (2019) Recovery of trait heritability from whole genome sequence data. bioRxiv 588020

10. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

11. Golan D, Lander ES, Rosset S (2014) Measuring missing heritability: Inferring the

contribution of common variants. Proc Natl Acad Sci 111:E5272–E5281

12. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Consortium SWG of the PG, Patterson N, Daly MJ, Price AL, Neale BM (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295

13. Searle SR, Casella G, McCulloch CE (2009) Variance components. John Wiley & Sons, United States

14. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet 88:76–82

15. Loh P-R, Bhatia G, Gusev A, et al (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. Nat Genet 47:1385–1392

16. Nagylaki T (1982) Assortative mating for a quantitative character. J Math Biol 16:57–74

17. Domingue BW, Belsky DW, Fletcher JM, Conley D, Boardman JD, Harris KM (2018) The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health. Proc Natl Acad Sci 115:702–707

18. Yengo L, Sidari M, Verweij KJH, Visscher PM, Keller MC, Zietsch BP (2020) No Evidence for Social Genetic Effects or Genetic Similarity Among Friends Beyond that Due to Population Stratification: A Reappraisal of Domingue et al (2018). Behav Genet 50:67–71

19. Bulik-Sullivan B (2015) Relationship between LD Score and Haseman-Elston Regression. https://doi.org/10.1101/018283

20. Sudlow C, Gallacher J, Allen N, et al (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med 12:e1001779

21. Peyrot WJ, Robinson MR, Penninx BWJH, Wray NR (2016) Exploring Boundaries for the Genetic Consequences of Assortative Mating for Psychiatric Traits. JAMA Psychiatry 73:1189–1195

22. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565

23. Verweij KJH, Mosing MA, Zietsch BP, Medland SE (2012) Estimating Heritability from Twin Studies. In: Elston RC, Satagopan JM, Sun S (eds) Stat. Hum. Genet. Methods Protoc. Humana Press, Totowa, NJ, pp 151–170

24. Gimelfarb A (1984) Quantitative characters under assortative mating: Gametic model. Theor Popul Biol 25:312–330

25. Bai Z, Silverstein JW (2010) Spectral analysis of large dimensional random matrices, Second. https://doi.org/10.1007/978-1-4419-0661-8

26. Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19:1–17

27. Jiang J, Li C, Paul D, Yang C, Zhao H (2016) On high-dimensional misspecified mixed model analysis in genome-wide association study. Ann Stat 44:2127–2160

28. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

29. Loh P-R, Danecek P, Palamara PF, et al (2016) Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48:1443–1448

30. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL (2018) Mixed-model association for biobank-scale datasets. Nat Genet 1

31. Border R, Becker S (2019) Stochastic Lanczos estimation of genomic variance components for linear mixed-effects models. BMC Bioinformatics 20:411

32. Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved Heritability Estimation from Genome-wide SNPs. Am J Hum Genet 91:1011–1021

33. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7

34. Evans LM, Tahmasbi R, Vrieze SI, et al (2018) Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat Genet 50:737–745

35. Finucane HK, Bulik-Sullivan B, Gusev A, et al (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47:1228–1235

36. Lee JJ, Wedow R, Okbay A, et al (2018) Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet 50:1112

Phenotype	r	n	$\hat{h}^2_{\text{HE}}$ [se]	$\hat{h}^2_{\infty}$ [se]	$\hat{h}^2_{\text{REML}}$ [se]	$\hat{h}^2_0$ [se]	$\hat{h}^2_{\text{HE}}/\hat{h}^2_{\infty}$ [se]
Height	0.240	334,798	0.567 [4.68e-3]	0.499 [4.23e-3]	0.525 [2.22e-3]	0.467 [3.37e-3]	1.14 [8.70e-4]
Years of Education	0.412	332,198	0.174 [2.64e-3]	0.163 [2.46e-3]	0.155 [2.01e-3]	0.153 [2.06e-3]	1.07 [9.46e-4]

Table 1. Inflation of heritability estimates for select UK Biobank traits

Spousal correlations (as previously reported in British cohorts [2, 4]) and heritability estimates for height and years of education in the UK Biobank, selected *a priori* on the basis of previous evidence for primary phenotypic AM. Assuming equilibrium, $\hat{h}^2_{\infty}$ and $\hat{h}^2_0$ (defined in Equation 3) provide unbiased estimates of the present day and panmictic heritabilites, respectively. Under disequilibrium, they respectively provide probabilistic lower bounds, with $\hat{h}^2_{\text{HE}}$ and $\hat{h}^2_{\text{REML}}$ providing complementary upper bounds. The ratio $\hat{h}^2_{\text{HE}}/\hat{h}^2_{\infty}$ reflects the extent to which HE regression overestimates the true heritability under the assumption of equilibrium.
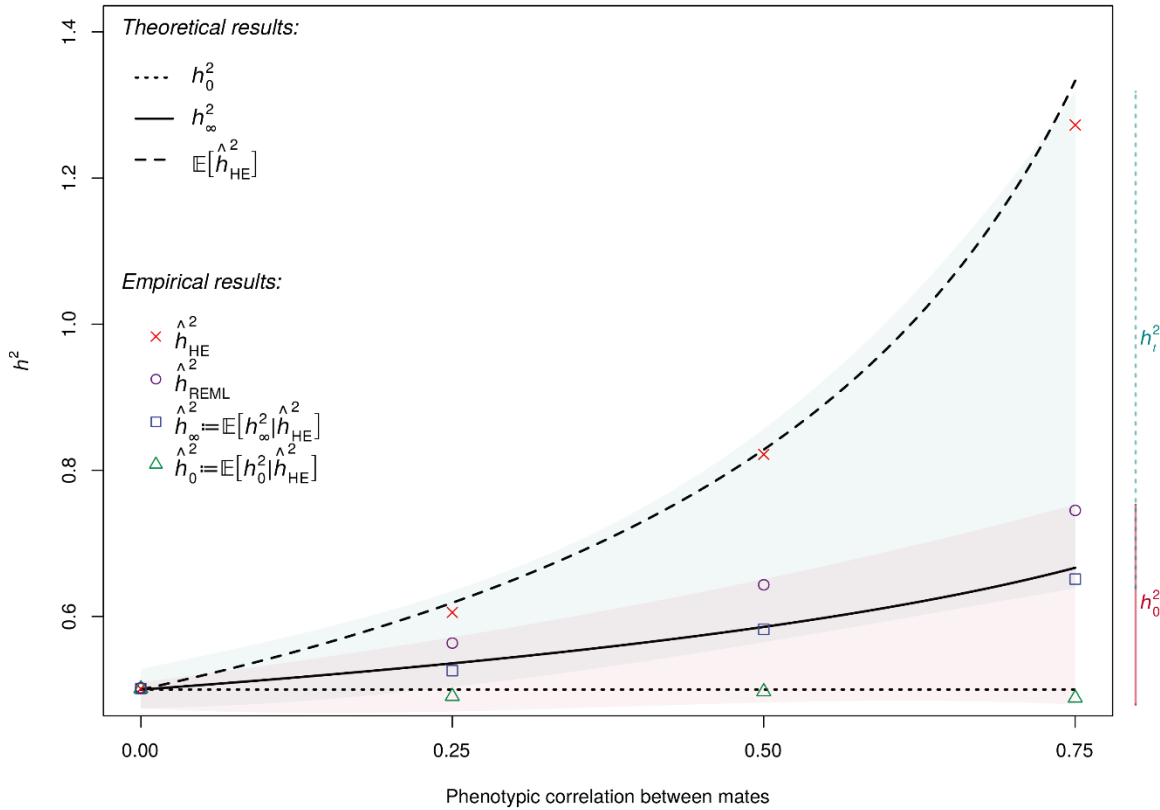
Figure 1. Theoretical and empirical behavior of existing and corrected estimators

Theoretical and empirical behavior of HE regression REML at equilibrium for varying phenotypic correlations among mates (r) and fixed panmictic heritability ($h_0^2 = .5$) across simulated datasets with sample size fixed at $n = 64{,}000$. HE regression ($\widehat{h}_{HE}^2$) produces upwardly biased estimates consistent with our closed-form approximation under the assumption of exchangeable loci, $\mathbb{E}[\widehat{h}_{HE}^2]$. Further, our corrected estimators, $\widehat{h}_\infty^2$ and $\widehat{h}_0^2$, recover the true equilibrium and panmictic heritabilities for a trait at equilibrium. For a trait at disequilibrium, the present day heritability h_t^2 is bounded in expectation between $\widehat{h}_{HE}^2$ and $\widehat{h}_\infty^2$ (represented by the teal shaded region), whereas the panmictic heritability h_0^2 is bounded in expectation between $\widehat{h}_{REML}^2$ and $\widehat{h}_0^2$ (represented by the red shaded region).
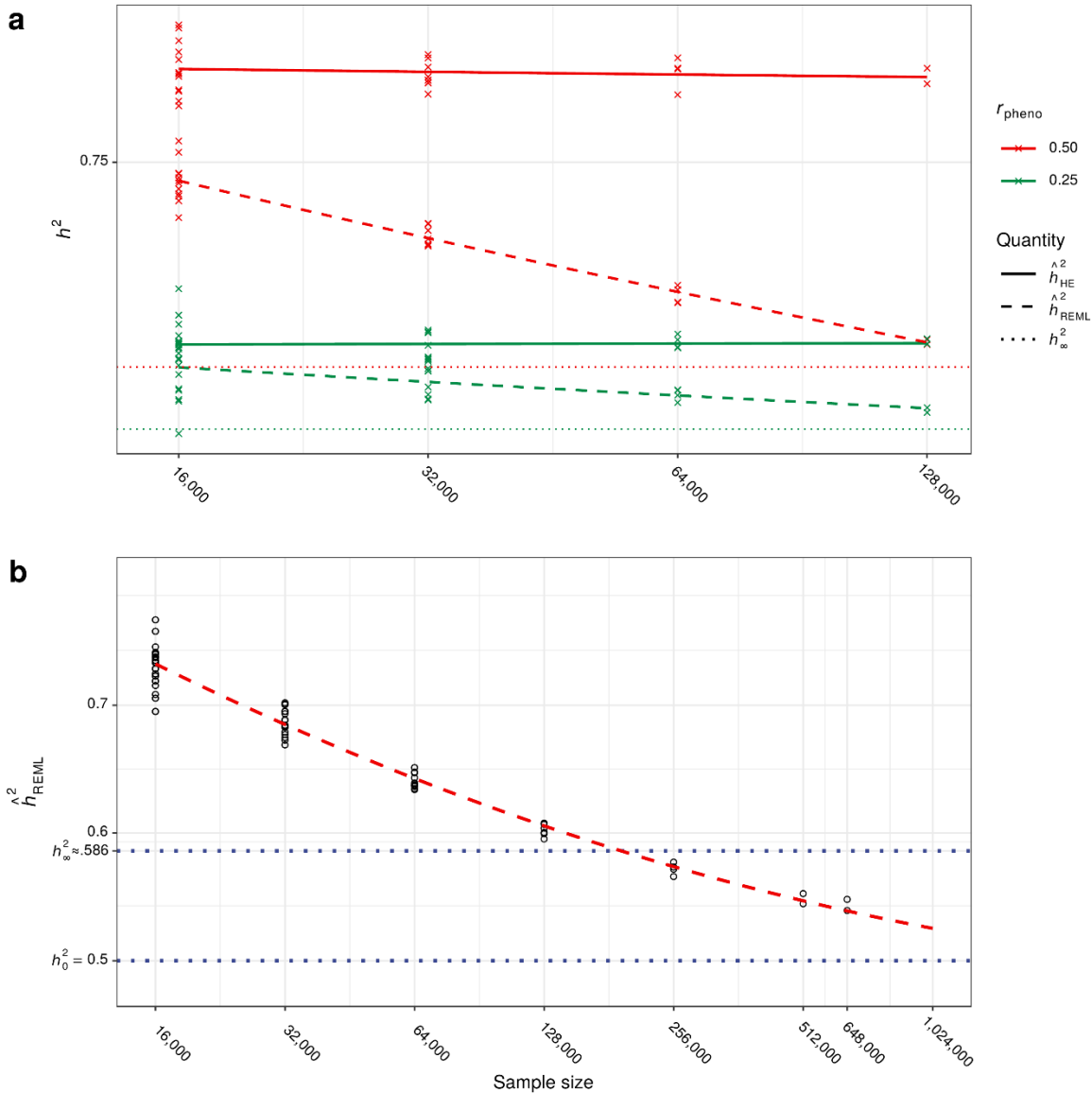
Figure 2. REML and HE estimates across varying sample sizes in simulated data

(a) Comparison of HE regression and REML heritability estimates as functions of sample size for varying phenotypic mating correlation (r) and fixed panmictic heritability ($h_0^2 = .5$) in simulated data. We computed multiple estimates per sample size for each estimator and parameter combination by applying estimators to independent sub-samples. Whereas HE regression estimates are upwardly biased independent of sample size, REML estimates slowly converge to the panmictic heritability as sample sizes increase.

(b) Extended simulations demonstrating high-dimensional behavior of the REML estimator as a function of sample size for fixed phenotypic mating correlation ($r = .5$) and panmictic heritability ($h_0^2 = .5$). Forward time simulations required a larger population size ($N_{\text{sim}} = 3e6$) to obtain samples of up to $n = 648{,}000$ unrelated individuals. Obtaining REML estimates for samples larger than this was not computationally feasible, but the dashed red line shows predicted values for larger sample sizes extrapolated from a regression model including first and second order log-linear components. Results are consistent with theoretical predictions that the REML estimator converges to the panmictic heritability in very large samples (e.g., $>1e6$).
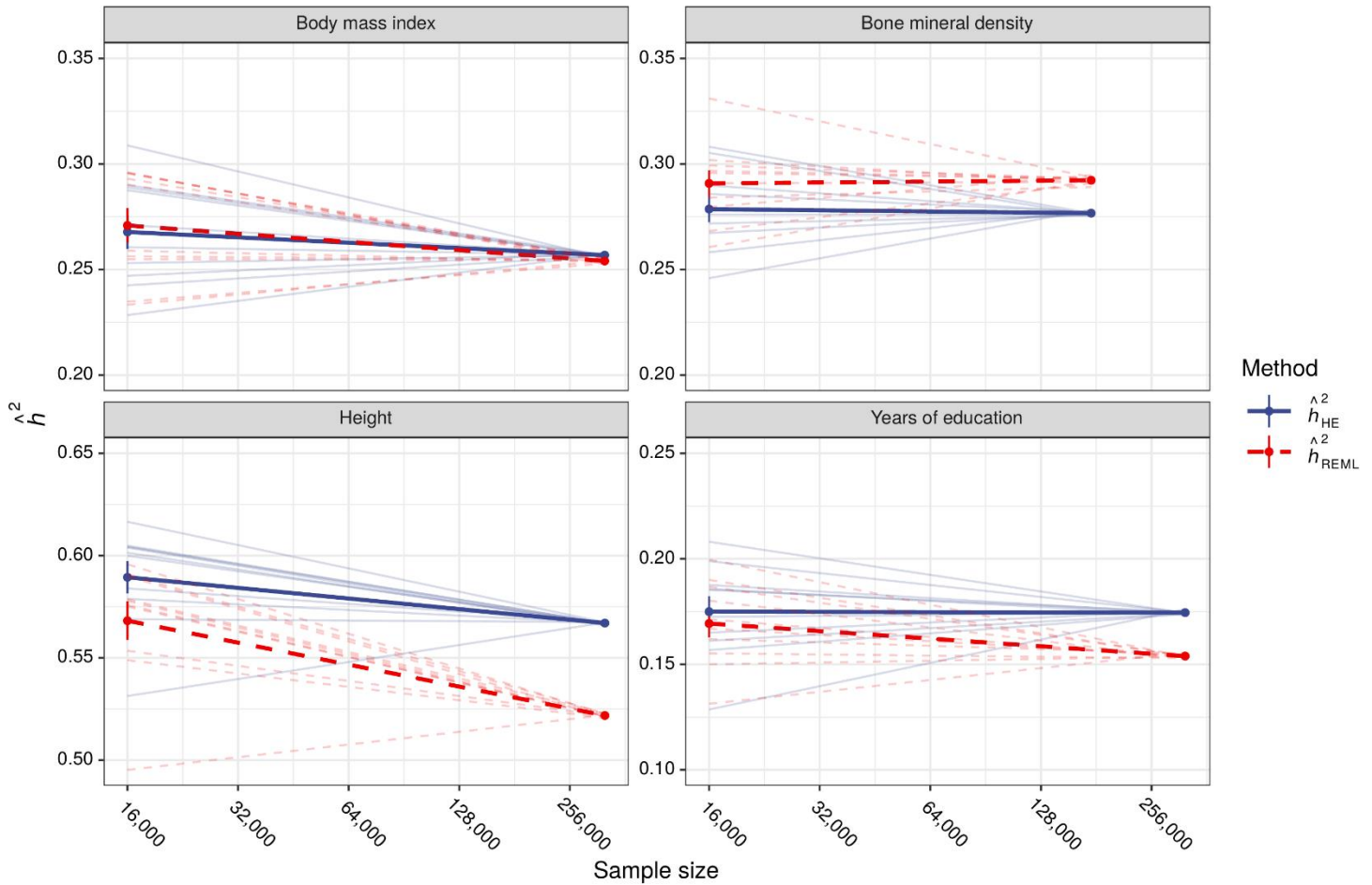
Figure 3. REML and HE estimates across varying sample sizes in UK Biobank data

Comparison of HE regression and REML estimators as a function of sample size for real traits in a sample of unrelated European ancestry UK Biobank participants. Points connected by thin lines represent estimates derived from pairs of complementary disjoint subsamples of size 16,000 and $N - 16{,}000$, whereas thick lines reflect average log-linear trends. Two negative control traits (body mass index and bone mineral density) and two traits with previous evidence for AM (height and years of education) were selected for analysis *a priori*. Consistent with theoretical predictions, height and years of education demonstrated significant estimator divergence with increasing sample size (p=5.24e-4, p=3.94e-4, respectively), whereas body mass index and bone mineral density did not (p=9.42e-2, p=0.302, respectively).
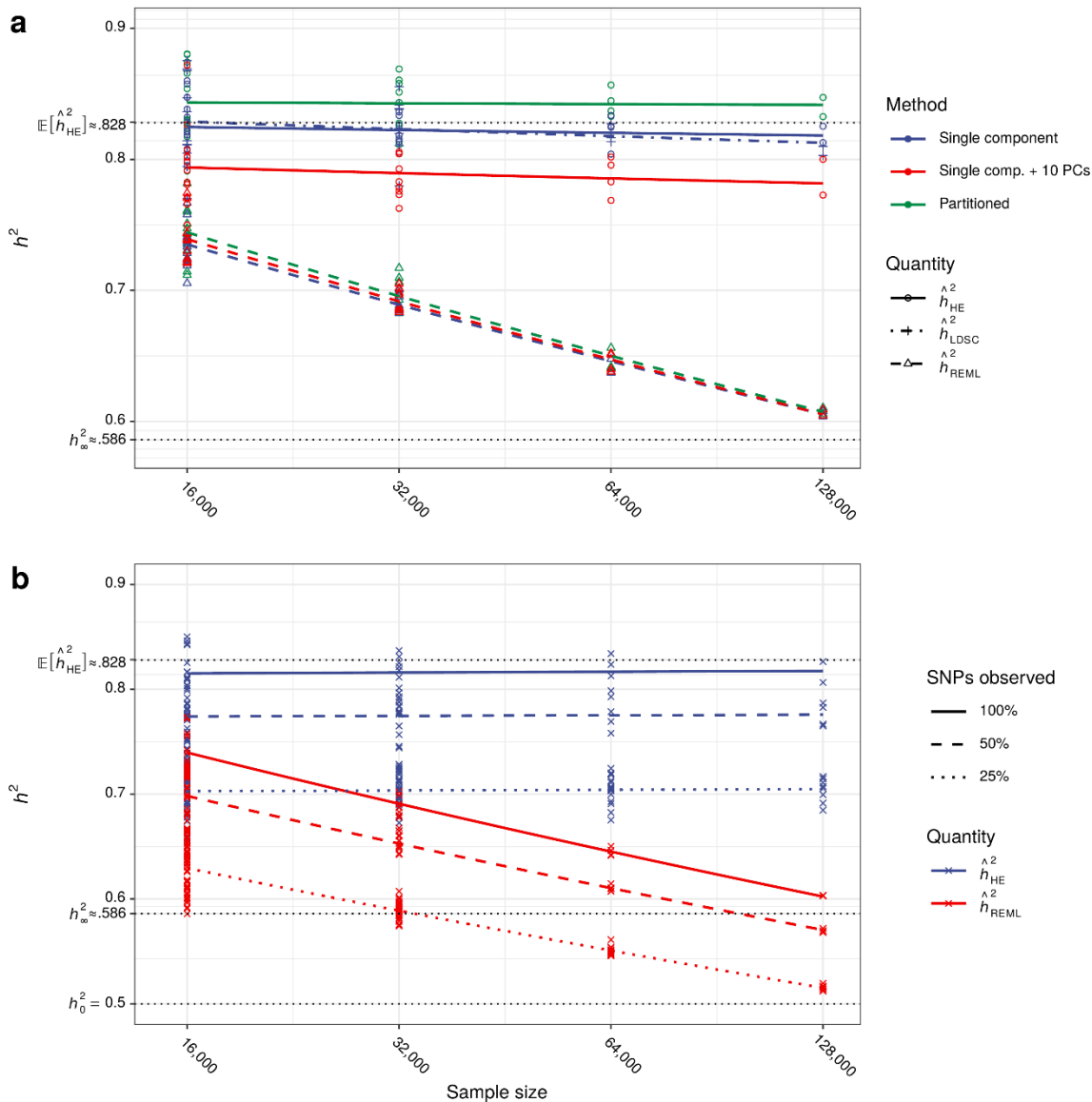
Figure 4. Naïve approaches to addressing AM induced bias and the impact of missing data

(a) Simulations demonstrate that neither partitioned heritability nor principal component adjusted methods mitigate the impact of assortative mating on HE regression and REML estimates. Additionally, simulations confirm that LD score regression (LDSC), which is mathematically equivalent to HE regression, is subject to equivalent biases. "Single component" refers to standard infinitesimal single genomic variance component models, "Single comp. + 10 PCs" included the first ten within-sample principal components as covariates, and "Partitioned" included four annotation-based variance components generated by median splits of within-sample minor allele frequencies and linkage disequilibrium scores. Phenotypic mating correlation and panmictic heritability were fixed ($r = .5$, $h_0^2 = .5$, respectively) across simulations. (b) Simulations demonstrate that conclusions regarding estimator bias do not change when only a portion of the genetic variance is explained by measured SNPs. Shown are HE regression and REML estimators when 100%, 50%, or 25% of randomly selected SNPs (both causal and non-causal) were included in the model under the same simulation conditions described in (a). As expected, estimates were attenuated when SNPs were missing but overall patterns remained consistent.