Limitations of GCTA as a solution to the missing heritability problem

Siddharth Krishna Kumar[a,1], Marcus W. Feldman[a], David H. Rehkopf[b], and Shripad Tuljapurkar[a]

[a]Department of Biology, Stanford University, Stanford, CA 94305-5020; and [b]School of Medicine, Stanford University, Stanford, CA 94305-5020

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved November 20, 2015 (received for review October 9, 2015)

Genome-wide association studies (GWASs) seek to understand the relationship between complex phenotype(s) (e.g., height) and up to millions of single-nucleotide polymorphisms (SNPs). Early analyses of GWASs are commonly believed to have "missed" much of the additive genetic variance estimated from correlations between relatives. A more recent method, genome-wide complex trait analysis (GCTA), obtains much higher estimates of heritability using a model of random SNP effects correlated between genotypically similar individuals. GCTA has now been applied to many phenotypes from schizophrenia to scholastic achievement. However, recent studies question GCTA's estimates of heritability. Here, we show that GCTA applied to current SNP data cannot produce reliable or stable estimates of heritability. We show first that GCTA depends sensitively on all singular values of a high-dimensional genetic relatedness matrix (GRM). When the assumptions in GCTA are satisfied exactly, we show that the heritability estimates produced by GCTA will be biased and the standard errors will likely be inaccurate. When the population is stratified, we find that GRMs typically have highly skewed singular values, and we prove that the many small singular values cannot be estimated reliably. Hence, GWAS data are necessarily overfit by GCTA which, as a result, produces high estimates of heritability. We also show that GCTA's heritability estimates are sensitive to the chosen sample and to measurement errors in the phenotype. We illustrate our results using the Framingham dataset. Our analysis suggests that results obtained using GCTA, and the results' qualitative interpretations, should be interpreted with great caution.

GCTA | GWAS | heritability | SNP | singular value decomposition

In recent years, genome-wide association studies (GWASs) have become an important tool for investigating the genetic contribution to complex phenotypes. These studies use statistical techniques to find associations between single nucleotide polymorphisms (SNPs) and phenotype(s) (e.g., continuous traits such as height or discrete traits such as presence/absence of a disease). A widely used measure of genetic influence on a phenotype is the (narrow-sense) heritability, defined as the ratio of the additive genetic variance to the total phenotypic variance. A major conundrum revealed by many analyses of GWAS data has been that the small number of significant associations explain much less of the heritability than is estimated from correlations between relatives [i.e., much heritability is "missing" (1–3)]. To address this problem, Yang et al. (4) posited that heritability is not missing but is "hidden." The authors developed a statistical framework [genome-wide complex trait analysis (GCTA)] in which each SNP makes a random contribution to the phenotype, and these contributions are correlated between individuals who have similar genotypes. Applied to many GWASs, GCTA yields estimates of heritability far larger than those obtained using earlier analyses. GCTA has been used to estimate the heritability of many phenotypes from schizophrenia (5) to scholastic achievement (6). Despite its current wide use, recent studies (7, 8) have questioned the reliability of GCTA estimates.

We show here that the results produced using GCTA hinge on accurate estimation of a high-dimensional genetic relatedness matrix (GRM). We show that even when the assumptions in GCTA are satisfied exactly, heritability estimates produced by GCTA will be biased, and it is unlikely that the confidence intervals will be accurate. When there is genetic stratification in the population, we show that GCTA's heritability estimates are guaranteed to be unstable and unreliable, which is especially relevant because stratification is common in human GWASs.

Our analysis has two other important consequences: (*i*) the heritability estimate produced by GCTA is sensitive to the choice of the sample used; and (*ii*) the estimate is sensitive to measurement errors in the phenotype. We argue that this instability and sensitivity are attributable to the fact that GCTA necessarily overfits typical GWASs. We show that a direct approach to eliminating this overfitting leads back to the small SNP heritability estimates derived previously from association studies. We illustrate our results using the Framingham dataset (9, 10) comprising information on 49,214 SNPs in 2,698 unrelated individuals.

We conclude that application of GCTA to GWAS data may not reliably improve our understanding of the genomic basis of phenotypic variability. Even when the assumptions for GCTA all hold, we recommend the use of diagnostic tests, and we describe one such test. We also discuss several ways of moving toward better methods.

The Data and the GCTA Model

The Data. A typical GWAS takes phenotypic values for N individuals and assays the individuals' genotypes at P single nucleotide sites (SNPs). Typically, $N \ll P$ [e.g., our illustrations use the Framingham data on 2,698(=N) unrelated individuals at 49,214(=P) SNPs]. The genetic data can be represented as an $N \times P$ matrix $\mathbf{X}$ whose (i,j) entry takes the value 0, 1, or 2 corresponding to the number of copies of the reference allele of the j^{th} SNP in the i^{th} individual.

Significance

The genetic contribution to a phenotype is frequently measured by heritability, the fraction of trait variation explained by genetic differences. Hundreds of publications have found DNA polymorphisms that are statistically associated with diseases or quantitative traits [genome-wide association studies (GWASs)]. Genome-wide complex trait analysis (GCTA), a recent method of analyzing such data, finds high heritabilities for such phenotypes. We analyze GCTA and show that the heritability estimates it produces are highly sensitive to the structure of the genetic relatedness matrix, to the sampling of phenotypes and subjects, and to the accuracy of phenotype measurements. Plausible modifications of the method aimed at increasing stability yield much smaller heritabilities. It is essential to reevaluate the many published heritability estimates based on GCTA.

Author contributions: S.K.K. and S.T. designed research; S.K.K. conducted much of the analysis with assistance from M.W.F.; D.H.R. and S.T. performed the initial data filtering; and S.K.K., M.W.F., and S.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: sidkk86@stanford.edu.

Mixed-Effect Models. To quantify how genes influence phenotypes, mixed linear models (11, 12) of the form

$$\mathbf{y} = \mathbf{F}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \quad [1]$$

are used in the animal-breeding literature. Here, $\mathbf{y} = \{y_i\}$ is a vector of phenotype values for the N individuals, $\mathbf{F}\boldsymbol{\beta} = \boldsymbol{\mu}$ is a vector of "fixed effects," $\mathbf{u}$ is a vector of random effects of an individual's genotype, $\boldsymbol{\epsilon}$ is a vector of residuals, and $\mathbf{Z}$ is a matrix describing how genetic effects are correlated between individuals. In animal-breeding studies and some human studies, the entries of $\mathbf{Z}$ are estimated from pedigrees.

What GCTA Assumes and Does. GCTA applies Eq. **1** to GWASs, where genetic and phenotypic information are usually measured on unrelated individuals. Each individual's genotype is given by P numbers, so the vector $\mathbf{u}$ of random effects is $P \times 1$ and the matrix $\mathbf{Z}$ is $N \times P$. GCTA estimates $\mathbf{Z}$ by centering and scaling the data matrix $\mathbf{X}$ using Hardy–Weinberg assumptions and assumes that $\mathbf{u} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \alpha^2\mathbf{I})$.

The fixed effects term $\boldsymbol{\mu}$ is commonly dropped from the analysis and the GCTA model written as

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad [2]$$

which we use henceforth. This model can be rewritten more clearly, with the same $\boldsymbol{\epsilon}$, as

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}, \quad [3]$$

where $\mathbf{g} \sim \mathcal{N}(0, V_g\mathbf{A})$ is a random vector of genetic contributions to the phenotypes, $V_g = P\sigma^2$ is the variance of total additive genetic effects, and $\mathbf{A} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}/P$ is the GRM between pairs of individuals.

GCTA obtains maximum-likelihood estimates (MLEs) of the parameters α^2 and V_g and then estimates the heritability as the ratio V_g/V_p, where $V_p = V_g + \alpha^2$ is the observed variance in the phenotype (4).

The relatively simple structure of Eq. **3** (linear, additive, and no environmental or epigenetic effects) relies on assumptions that GCTA has in common with the animal-breeding literature. GCTA assumes that the SNPs used are in linkage equilibrium; in practice, SNPs in linkage disequilibrium are avoided. In our analysis and example, we assume that this selection has been made. However, GCTA makes important additional assumptions: assumption 1, each SNP makes a random contribution to the phenotype independent of the others; assumption 2, the distribution of these random contributions is identical for all SNPs; and assumption 3, there is no genetic stratification in the population.

In this paper, we use a rigorous mathematical analysis of GCTA to answer two key questions. How reliable are the GCTA estimates when the assumptions in the model are satisfied exactly? How robust are the GCTA estimates to violations of assumption 3 (with or without a correction)?

We begin with the key point that any observed realization of the matrix $\mathbf{Z}$ in Eq. **2** is a sample from some underlying distribution of possible data. In fact, $\mathbf{Z}$ is a random matrix. Hence, the data (the entries of $\mathbf{Z}$) and the resulting GRM ($\mathbf{A} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}/P$) will have sampling errors. As a consequence, the MLEs produced by GCTA are statistical estimates of the parameters in Eq. **3**. To analyze the precision and stability of these MLEs, we now develop the connection between the MLEs and the geometry of $\mathbf{Z}$ (in terms of the matrix's singular values and singular vectors).

Singular Values and GCTA. The MLEs produced by GCTA depend on the properties of the GRM matrix ($\mathbf{A} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}/P$). We prove here that these MLEs can be expressed in terms of the singular values and associated singular vectors—the spectral properties—of the data matrix $\mathbf{Z}$. Readers will be familiar with spectral properties in the context of principal component analysis (PCA). In PCA, we rank the N eigenvalues (and associated eigenvectors) of the symmetric matrix $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$. PCA is equivalent to a singular value decomposition (SVD) of the matrix $\mathbf{Z}$, which produces a set of $k = \min(N, P)$ real singular values ($\{w_i\}$ for $1 \le i \le k$), a set of left singular vectors ($\{\mathbf{u}_i\}$ for $1 \le i \le k$) of dimension $N \times 1$, and a corresponding set of right singular vectors ($\{\mathbf{v}_i\}$ for $1 \le i \le k$) of dimension $P \times 1$. The eigenvalues of $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ in a PCA are the squares of the singular values of $\mathbf{Z}$ (the PCA eigenvectors are the left singular vectors of $\mathbf{Z}$).

We find (*Appendix A*) that the MLEs computed by GCTA are explicit functions of the singular values and singular vectors of $\mathbf{Z}$. We write the MLEs for α^2 and σ^2 as the sum of three terms, the second of which (Eqs. **A3** and **A8**) is a function of

$$\log\left(\prod_{i=1}^{i=k}\left((1/w_i^2) + (\sigma^2/\alpha^2)\right)\right); \quad [4]$$

note the terms in $(1/w_i^2)$. Using $\mathbf{y}_1$ to represent a realization of $\mathbf{y}$, the third term (Eqs. **A3** and **A9**) is a function of $\mathbf{Z}^{\mathrm{T}}\mathbf{y}_1$ and of the left singular vectors ($\{\mathbf{u}_i\}$ for $1 \le i \le k$) of $\mathbf{Z}$. We use these results to determine how sampling errors influence the estimates produced by GCTA in two cases.

Case 1: GCTA's Assumptions Are Satisfied Exactly. When assumptions 1 through 3 hold, the matrix $\mathbf{Z}$ (asymptotically) has an N variate Wishart distribution with P degrees of freedom. Marčenko and Pastur (13) (also see refs. 14 and 15 for more readable expositions) show that for samples $\mathbf{Z}$ from a Wishart distribution, the empirical distribution of the eigenvalues of $\mathbf{A} = (1/P)\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ (Fig. 1*A*) converges to a limiting form when $N \to \infty$, $P \to \infty$, and P/N is finite. This limit distribution depends only on the value of P/N and, although asymptotic, accurately describes the eigenvalue distribution of $\mathbf{A}$ even for sample sizes as small as $N = 1{,}000$ (Fig. 1*B*). Note from Fig. 1*A* that for all values of P/N, almost all eigenvalues of $\mathbf{A}$ lie within the interval 0–4, and that as P/N approaches 1, the spectrum of $\mathbf{A}$ becomes ill-conditioned (i.e., the ratio of its largest to its smallest eigenvalue becomes large).

In all available GWASs, there are more SNPs than people, so $P/N \gg 1$. For such cases, the known results above imply that the eigenvalues of the GRM are well-conditioned, so that the errors in the second term of the MLE [i.e., in $((1/w_i^2))$] will be small.

In addition to the eigenvalues, the third term in the MLE expression depends on the eigenvectors of the GRM. Because the eigenvectors of the true GRM are unknown, GCTA proceeds by approximating the eigenvectors of the true GRM by the eigenvectors of the sample GRM. This approximation will be valid if the eigenvectors of the true GRM are "similar" to the eigenvectors of the sample GRM. Standard results from perturbation theory show that the eigenvectors will be similar only if the eigenvalues of the sample GRM are not packed close to one another. However, because $N \sim O(10^3)$ eigenvalues of the GRM are packed in the interval 0–4, the eigenvalues of the sample GRM are guaranteed to be packed close to one another (for our simulations in Fig. 1*B*, where $N = 1{,}000$ and $P/N = 4$, the two closest eigenvalues have magnitudes 0.2412 and 0.2417, respectively). As a result of this close packing, the eigenvectors of the sample GRM can be drastically different from the eigenvectors of the true GRM and, these differences bias the MLEs of α^2 and σ^2 by amplifying the sampling errors associated with the GRM (see *Appendix B* for details). If these biases in the MLEs are large, the heritability estimates produced by GCTA will not be representative of the true underlying heritability of the phenotype. Furthermore, because GCTA estimates the SE of the heritability as a function of the MLEs, large biases would make this SE meaningless.

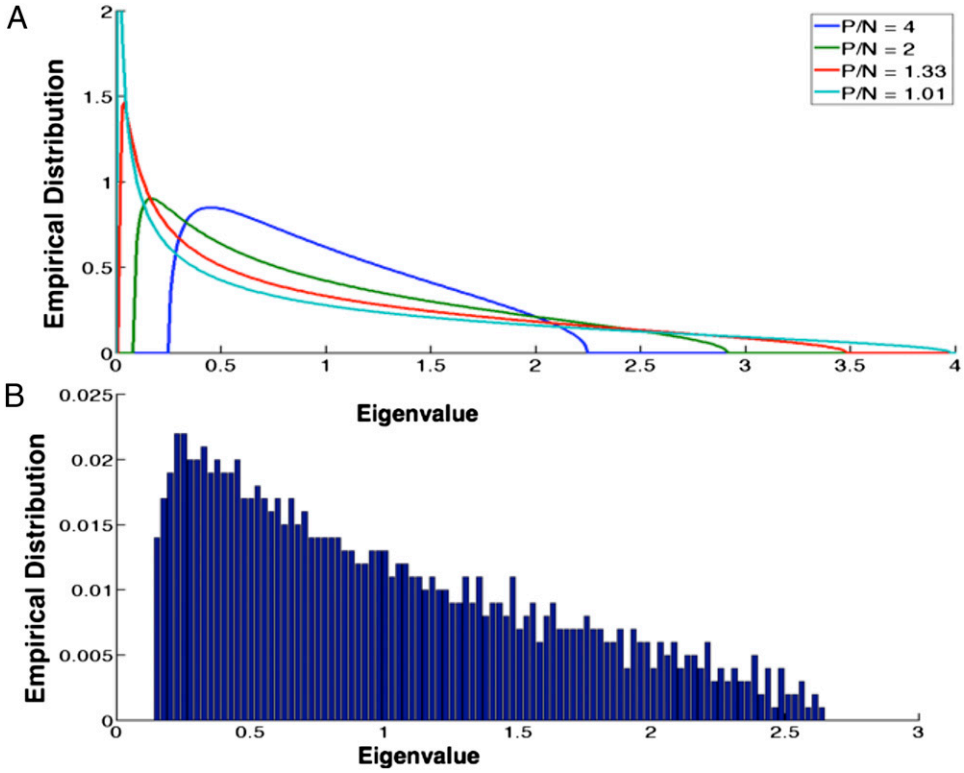

Fig. 1. The M-P distribution. (*A*) Plots of the M-P distribution of eigenvalues for different values of P/N. For P/N values close to 1, the eigenvalues of the GRM are extremely skewed. As P/N increases, the eigenvalues are concentrated on a smaller interval. (*B*) A frequency histogram of the eigenvalues for a sample GRM simulated using $N = 1{,}000$ and $P/N = 4$. One eigenvalue had a magnitude greater than 400 (which is not represented in this plot). Note that even for sample sizes as small as 1,000, the M-P distribution accurately captures the distribution of the eigenvalues of the GRM.

To demonstrate this bias, we simulated a dataset comprising 50,000 SNPs in linkage equilibrium for 2,000 people [using PLINK software (16)] and a phenotype with a heritability of 0.75 (using GCTA). The simulation assumes that the entire additive genetic contribution to the phenotype comes from 45,000 out of the total 50,000 SNPs (the causal SNPs) whose effect sizes are normally distributed with mean 0 and variance 1.

Using GCTA on this dataset, we estimate the genotypic variance $V_g = P\sigma^2$ as ≈ 0.685 with a SE of 0.151. This result, along with standard results from large sample theory (17), state that the MLE of σ^2 is approximately normally distributed with mean $0.685/50{,}000 = 1.37\times 10^{-5}$ and SD $1.51/50{,}000 = 3.1\times 10^{-6}$ which forms GCTA's null hypothesis.

To test this hypothesis, we construct 500 genotype matrices, each comprising all 2,000 people but only 5,000 SNPs randomly chosen from the initial 50,000. We ran GCTA using each of these genotype matrices, and in each case, estimated σ^2 (Fig. 2). More than half of these estimates lie outside the 95% confidence interval (marked by red arrows), with the largest of these estimates (4.497×10^{-5}) being more than 10 SDs away from the mean; these results suggest that GCTA's null is almost certainly being violated.

Although our results guarantee that each estimate of σ^2 in Fig. 2 will be biased, our results do not provide any general information about the magnitude of the bias. It is possible that the bias in some of these estimates is small, and some resampling procedure might resolve problems raised in this section; we do not pursue this here.

Case 2: There Is Genetic Stratification. Assumption 3 is typically violated: stratification is widely observed in humans (18, 19) and animals (20, 21) and is a major reason for the high number of false discoveries in GWASs (22). GCTA claims to address this by

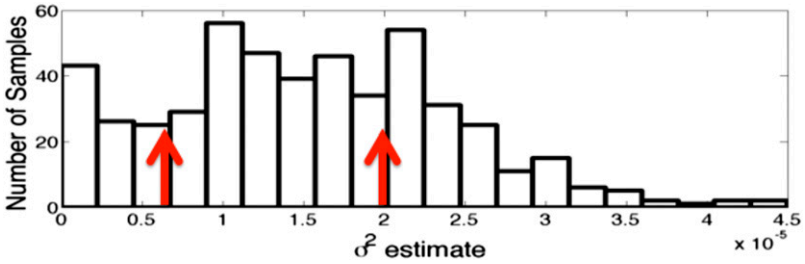

Fig. 2. Large deviations in σ^2 estimates. Estimates of σ^2 are produced by GCTA using 500 randomly sampled genotype matrices, each comprising 5,000 SNPs drawn at random from a simulated dataset (comprising 50,000 SNPs), which satisfies all of GCTA's assumptions. When GCTA is run on the entire dataset, we get a σ^2 estimate of 1.37×10^{-5} and SE of 3.1×10^{-6}; the 95% confidence interval corresponding to this estimate is marked with red arrows. Clearly, the confidence intervals produced by GCTA are grossly underestimating the uncertainty in the σ^2 estimates.

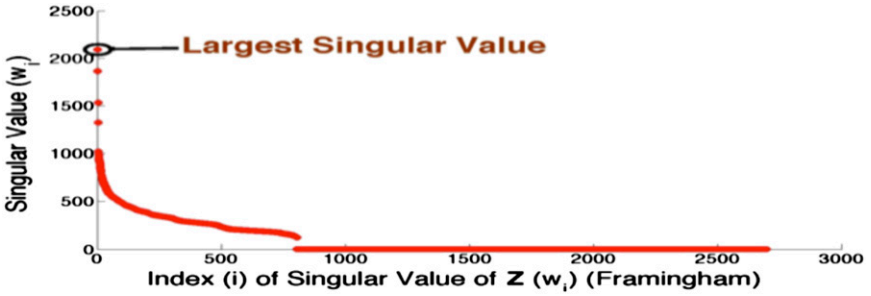

Fig. 3. Skew in the singular values. Notice that **Z** has a long tail of near-zero singular values. The largest singular value of **Z** is $O(10^3)$, and the smallest nonzero singular value of **Z** is $O(10^{-7})$, showing that **Z** has a very high condition number.

incorporating eigenvectors of the GRM as fixed effects [as columns of **F** in Eq. **1**, as suggested by Eigenstrat software (23)]. Surprisingly, GCTA finds that in most cases, the fixed-effect term has little influence (i.e., the heritability estimate from GCTA is nearly independent of the stratification). We will show that the analysis provided by GCTA is flawed and that stratification will induce large errors in the MLEs. We illustrate our points using the Framingham dataset (9, 10), which is known to be stratified.

We begin by analyzing how genetic stratification influences the spectral properties of **Z**. A plot of the singular values of **Z** for the Framingham dataset (Fig. 3) reveals that these values are extremely skewed. The first four singular values are greater than 1,000, and the next 20 or so are between 100 and 1,000, whereas thousands of singular values are close to 0: the largest singular value is $\sim 10^{10}$ times the smallest singular value.

It is well known (see theorem 3 in ref. 24 and section 4.3 in ref. 25) that such skews must occur in stratified populations. In essence, these studies show that if N is large, and the markers are sampled from K different populations, the first $K-1$ singular values of **Z** will be much larger than the remaining $N-K+1$. Because in most cases, $N \gg K > 1$, we expect most of the singular values of **Z** to be close to 0, as in Fig. 3.

This skew, and the long tail of near-zero singular values, have serious implications for the MLEs from GCTA. Recall that the second term in the MLE (Eq. **A8**) is sensitive to the near-zero singular values of **Z**. The third term (Eq. **A9**) is a function of $\mathbf{Z}^T\mathbf{y}_1$ and so is also sensitive to the near-zero singular values when **Z** is ill-conditioned (i.e., the ratio of the largest to the smallest singular value of **Z** is large). Therefore, the accuracy of both terms in the MLE expression hinge on the precise estimation of the near-zero singular values of **Z**. Stewart (26) shows that in the presence of noise, the estimation errors of a singular value whose "true" magnitude is 0 will be larger than the noise by a factor $\sqrt{P}$ (i.e., the estimates of the near-zero singular values are extremely imprecise). We now illustrate several problems with the GCTA estimates in genetically stratified populations, all of which stem from the imprecision of the near-zero singular values.

Sensitivity to the SNPs used in the study. The heritability estimates from GCTA will be sensitive to the SNPs used in the study because the errors associated with the near-zero singular values (and in turn the MLEs) for datasets constructed using different sets of SNPs will be different. To demonstrate this, we construct 2,500 genotype matrices, $\mathbf{X}_i$ for $1 \le i \le 2{,}500$, each comprising 5,000 (of the total 49,214 SNPs) randomly sampled SNPs, and use GCTA to estimate the heritability of systolic blood pressure (BP) (27) using each of these $\mathbf{X}_i$. Contrary to the claim in ref. 4, these heritability estimates show high variability (Fig. 4*A*), because of sampling errors associated with one or more of the near-zero singular values (Fig. 4*B*).

Sensitivity to the measurement errors in the phenotype. Because **Z** is ill-conditioned, small changes in the phenotype vector can cause large changes in the heritability estimates from GCTA. Hence, GCTA violates the "unspoken assumption that imprecision of measurement of phenotype will not have large systematic effects on the location of significant associations in GWAS" (28). To demonstrate this violation, we generate 2,500 noisy samples of the BP phenotype vector [for each sample, the i^{th} entry of the vector is drawn uniformly over the minimum and maximum of the four BP readings available to us for person i; in general, BP readings are much noisier (29)] and use GCTA to estimate the heritability using each of these vectors. Even for the modest errors in this case, GCTA shows high variability in its heritability estimates (Fig. 5).

Saturation of heritability estimates. According to GCTA, each SNP makes a random contribution to the phenotype; therefore, the heritability estimate, $P\sigma^2/V_p$, is necessarily directly proportional to P (because V_p is fixed). Several studies (see figure 5 in ref. 30) using GCTA have found results contradicting this; these studies report a threshold value above which, introducing more SNPs in the analysis produces only marginal increases in the heritability estimates from GCTA. This saturating behavior implies that above a threshold of P, $P\sigma^2$ is nearly independent of P (i.e., σ^2 is inversely proportional to P) which violates GCTA's assumption that the contribution of each SNP to the heritability estimate is independent of the others.

Bias in the heritability estimates. We have shown that for a stratified population, the MLEs produced by GCTA are guaranteed to be biased. The bias arises because thousands of eigenvalues of the GRM are closely packed (near 0) and have large sampling errors associated with their values (*Appendix B*). As a result of the bias, the heritability estimates produced by GCTA are not reflective of the "true" underlying heritability. Furthermore, the SEs reported by GCTA are functions of the MLEs, and so these SEs will also be unreliable.

To demonstrate this unreliability, we first ran GCTA on the Framingham dataset with BP as the phenotype. GCTA reports that $V_g = P\sigma^2$ has an estimate of 0.263 and a SE of 0.048. This result plus large sample theory imply that the MLE for σ^2 will be approximately normally distributed with mean $0.263/49{,}214 = 5.34 \times 10^{-6}$ and SD $0.048/49{,}214 = 9.75 \times 10^{-7}$ which forms GCTA's null hypothesis.

To test this hypothesis, we computed the estimate and SEs of σ^2 for each of the samples used in Fig. 4*A*. The σ^2 for almost all of these samples lies outside the 95% confidence intervals predicted by GCTA's null (marked by the red arrows), with the largest of these estimates (5.46×10^{-5}) being more than 50 SDs away from the mean; these results suggests that GCTA's null is almost certainly being violated.

Resampling techniques like the bootstrap cannot be used to correct for the bias in heritability estimates because every run of the bootstrap will produce a biased estimate of heritability, and there is no way of estimating the magnitude of the bias in any of these samples. Are there other approaches to fixing GCTA when

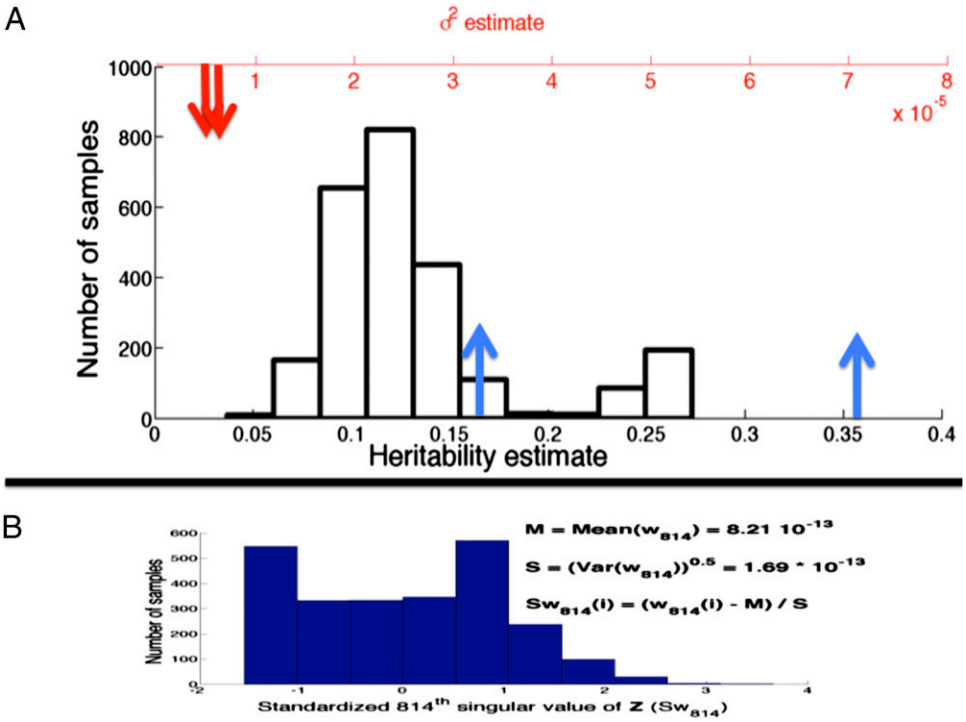

Fig. 4. Sensitivity of GCTA estimates to SNPs retained in GWASs. (*A*) Heritability and corresponding σ^2 estimates from the Framingham dataset produced by GCTA using 2,500 randomly sampled genotype matrices comprising 5,000 SNPs each (from the total 49,214). When GCTA is run on the entire dataset, we get a σ^2 estimate of 5.34×10^{-6} with a SE of 9.75×10^{-7}; the 95% confidence interval for the corresponding σ^2 and heritability estimates are marked with red and blue arrows, respectively. Note how drastically the null hypothesis is violated. (*B*) Histogram of the "standardized" 814th singular value of **Z** (w_{814}) for the 2,500 samples. This singular value has an extremely small magnitude for all of the samples ($\sim O(10^{-13})$) but has high estimation error and therefore can induce large errors in the heritability estimates produced using GCTA (see Eq. **A8** in *Appendix B* for details).

there is genetic stratification in the population? Two common approaches to fixing this problem are (*i*) constraining the random effects associated with only some of the SNPs to be relevant (sparsity) (31) or (*ii*) denoising the matrix **Z** (i.e., setting its lower noisy singular values to 0) before constructing it (32). We do not pursue the first approach because it violates the premise of GCTA that each SNP makes a random contribution to the phenotype. Using the latter approach, we show (*Appendix C*) that the contribution of the random effects term will not be significantly different from 0 (Fig. 6); these results are consistent with our findings on denoising the Framingham dataset. Because the random effects term is the sole driver of the "improved" heritability estimates produced by GCTA, in the term's absence, the heritability estimates will be no better than those obtained using the significant SNPs in association studies.

It is not surprising that the estimates produced by GCTA are sensitive and biased because the method estimates $NP+2$ parameters (NP parameters for the N nonzero singular values and their corresponding left and right singular vectors, plus α^2 and σ^2) from a dataset containing NP entries and therefore overfits the data. In contrast, association studies insist on stringent P values for significance and so greatly reduce the effective number of parameters being estimated (see ref. 33 for details); the resulting effective number of parameters is much smaller than the size of the dataset, so overfitting is not a problem.

The MLEs produced by GCTA will be unreliable irrespective of the number of principal components that are included in the model (see Fig. 7, where the MLEs produced by GCTA are unreliable even when five principal components are used as fixed effects; similar unreliabilities were observed when one and three principal components were used as fixed effects, respectively). Price and coworkers (23, 24) have shown that principal components are useful for identifying population stratification and when used with reliable methods like association studies, are effective in correcting for population stratification. When principal components are used in conjunction with GCTA, the analysis step is unreliable as a result of the overfitting. Therefore, although the principal components are still able to accurately identify population stratification, they only serve to compound the bias of GCTA.

Discussion

GCTA analyses have been widely accepted in large part because they produce heritability estimates that are many times larger than earlier estimates from GWAS data and are closer to those

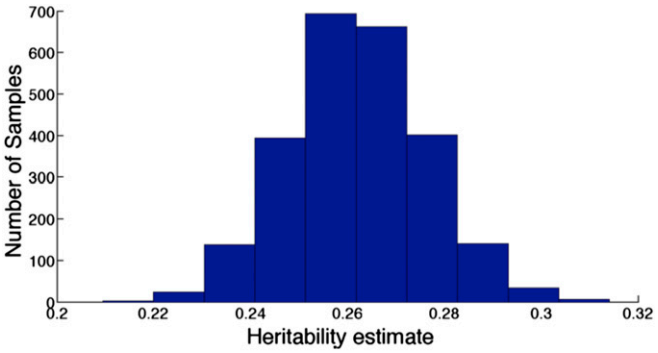

Fig. 5. Sensitivity of GCTA estimates to measurement errors in the phenotype. Heritability estimates produced by GCTA using 2,500 "noisy" estimates of the phenotype (systolic blood pressure) vector; in our study, we used all 2,698 people and 49,214 SNPs.

A

Source	Variance	SE
V(G)	0.154399	0.056085
V(e)	0.095729	0.055262
Vp	0.250128	0.007942
V(G)/Vp	0.617280	0.221665
logL	385.660	
logL0	381.801	
LRT	7.719	
df	1	
Pval	0.002733	
n	2000	

B

Source	Variance	SE
V(G)	0.043226	0.054328
V(e)	0.206905	0.054498
Vp	0.250131	0.007915
V(G)/Vp	0.172813	0.216986
logL	382.137	
logL0	381.801	
LRT	0.673	
df	1	
Pval	0.206	
n	2000	

Fig. 6. Fixing GCTA by denoising the GRM. (*A*) Heritability estimates for a highly heritable phenotype computed using the GRM as prescribed by GCTA (the phenotype was simulated as a quantitative trait using GCTA; the desired heritability of the trait was set to 0.65). Notice that the *P* values suggest a significant contribution from the random effects term. (*B*) Heritability estimates for the same phenotype vector computed using a denoised GRM (obtained by setting the noise terms to 0). Notice that the random effects associated with the SNPs are no longer significant. This result shows that the pathologies of GCTA are general and not dataset-specific (we have verified a similar trend with the Framingham dataset; a significant random effect term loses significance upon denoising the GRM).

obtained from data with reliable pedigrees. The statistical model in GCTA assumes that each SNP makes a small random contribution to the variability in the phenotype. Our analytical and numerical results illustrate the problems with GCTA when (*i*) the assumptions of the model are satisfied exactly or (*ii*) the assumptions are violated as a result of genetic stratification. In both cases, the problems associated with GCTA stem from the fact that a high-dimensional correlation matrix is being estimated from a limited amount of data without dimensionality reduction.

When there is genetic stratification in the population, the GRM has a long tail of near-zero eigenvalues; here, GCTA will produce unreliable heritability estimates. GCTA claims that including the first few principal components as fixed effects (following ref. 23) will resolve the problem of stratification, but this is not the case; even after including principal components as fixed effects, the problems associated with the near-zero singular values of **Z** remain. Principal components are useful for dealing with stratification in the context of association studies because principal components reduce the dimensionality of the problem via stringent *P* value criteria. We believe that stratification is responsible for many of the counterintuitive results reported by studies using GCTA (a more detailed discussion of these studies can be found in ref. 8). Furthermore, numerous studies on sensitive subjects like childhood intelligence (34), Tourette syndrome (35), and schizophrenia (36) need to be critically reviewed.

Even when there is no genetic stratification, our analysis strongly suggests that the heritability estimates and their SEs produced by GCTA will be unreliable. These unreliabilities are illustrated by our simulations (Fig. 2). We do not prove that they will apply for all datasets where there is no genetic stratification. Our illustration does suggest a simple test of the reliability of GCTA's estimates by a resampling procedure: first, construct many, say 500, genotype matrices by randomly sampling $(P/10)$ SNPs from the total P SNPs in the dataset and use GCTA to compute the σ^2 estimate corresponding to each of these genotype matrices. Next, compute the estimate (e_1) and SE (s_1) for σ^2 using all P SNPs. Under GCTA's null, ~99.5% of the distribution should lie in the interval $[e_1 - 4s_1,\ e_1 + 4s_1]$; if any of the 500 simulations estimates a σ^2 far outside this range, GCTA's estimates should not be trusted.

Heritability estimates using methods other than GCTA are generally low $\approx$ 3% to 4%. These methods use, for example, either single SNP associations or polygenic scores constructed from a few significant SNPs. We have shown that GCTA grossly underestimates the uncertainties associated with the individual SNP contributions and as a result, the only reliable heritability estimates are the 3–4% produced by these other methods.

The problems in GCTA stem from the overfitting of a high-dimensional GRM. To make progress with GCTA-like mixed models, it is critical that the estimates of this matrix be refined. Some progress has been made in this direction using, for example, methods for covariance smoothing (37). There are several alternative methods of describing the relatedness between individuals (for a survey of these methods, see ref. 38), some of which could prove useful in improving the estimate of the GRM.

We have shown that a brute force approach to estimating the covariance structure of the random effects of SNPs (as in GCTA) does not resolve the problem stemming from the number of SNPs (P) being much larger than the number of subjects genotyped (N). We believe that future studies of GWAS data can make progress by incorporating prior information. Two possible ways of so doing are (*i*) insisting that the basis used for constructing the GRM is sparse (i.e., only some SNPs make random contributions and the rest have fixed contributions) and

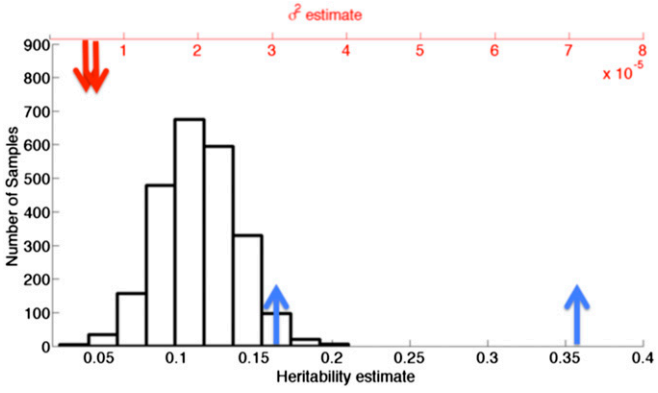

Fig. 7. Using principal components as fixed effects do not resolve problems with stratification. Resampling experiments were identical to those performed in Fig. 4, with five principal components included as fixed effects; GCTA's predicted 95% confidence intervals for the σ^2 and heritability estimates are marked with red and blue arrows, respectively. Note that the SEs are drastically underrepresenting the "true" variation in the σ^2 estimates, despite the inclusion of principal components as fixed effects. Near-identical plots are obtained when the simulations are run with one and three principal components, respectively.

(*ii*) incorporating biological information about the relationships between elements in the random covariance matrix.

Appendices

Appendix A: The Likelihood Function and Sensitivity. Expressing Eq. **2** in probabilistic form, we have

$$P(\mathbf{u}) = \mathcal{N}\left(\mathbf{u}|0, \sigma^2\mathbf{I}\right) \text{ and } P(\mathbf{y}|\mathbf{u}) = \mathcal{N}\left(\mathbf{y}|\mathbf{Zu}, \alpha^2\mathbf{I}\right). \quad \text{[A1]}$$

Therefore, the marginal distribution of $\mathbf{y}$ will be given by

$$P(\mathbf{y}) = \mathcal{N}(\mathbf{y}|0, \mathbf{C}), \text{ where } \mathbf{C} = \alpha^2\mathbf{I} + \sigma^2\mathbf{ZZ}^{\mathrm{T}}. \quad \text{[A2]}$$

We have only one sample of the vector $\mathbf{y}$, namely our observed vector of phenotypes (which we call $\mathbf{y}_1$). Because $\mathbf{y}$ has a multivariate normal distribution, the log likelihood of observing $\mathbf{y}_1$ is

$$\log P\left(\mathbf{y}_1|\alpha^2, \sigma^2\right) = \frac{-N}{2}\log(2\pi) - \log\det(\mathbf{C}) - \frac{1}{2}\mathbf{y}_1^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}_1. \quad \text{[A3]}$$

The Woodbury matrix identity (39) states that the inverse of a matrix $\mathbf{B} = \mathbf{A} + \mathbf{PRQ}$ is given by $\mathbf{B}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{P}(\mathbf{R}^{-1} + \mathbf{QR}^{-1}\mathbf{P})\mathbf{QA}^{-1}$. Hence in Eq. **A3** with $\mathbf{A} = \alpha^2\mathbf{I}$, $\mathbf{P} = \sigma^2\mathbf{Z}$, $\mathbf{Q} = \mathbf{Z}^{\mathrm{T}}$, and $\mathbf{R} = \mathbf{I}$, we have

$$\mathbf{C}^{-1} = \frac{1}{\alpha^2}\mathbf{I} - \frac{\sigma^2}{\alpha^4}\mathbf{Z}\left(\mathbf{I} + \frac{\sigma^2}{\alpha^2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}. \quad \text{[A4]}$$

We now use this formulation to show that for a stratified population, the heritability estimates will be sensitive to the data used in the GWAS.

Instability of the second term in Eq. A3. Using the SVD of $\mathbf{Z}$ ($\mathbf{Z} = \mathbf{U}_1\mathbf{W}_1\mathbf{V}_1^{\mathrm{T}}$) in Eq. **A2**, we have

$$\det(\mathbf{C}) = \alpha^{2N}\det\left(\mathbf{I} + \frac{\sigma^2}{\alpha^2}\mathbf{U}_1\mathbf{W}_1^2\mathbf{U}_1^{\mathrm{T}}\right). \quad \text{[A5]}$$

Sylvester's theorem for determinants (40) states that for invertible matrices $\mathbf{R}_1, \mathbf{R}_2$, the determinant $\det(\mathbf{R}_1 + \mathbf{PR}_2\mathbf{Q}^{\mathrm{T}}) = \det(\mathbf{R}_2^{-1} + \mathbf{Q}^{\mathbf{T}}\mathbf{R}_1^{-1}\mathbf{P})\det(\mathbf{R}_2)\det(\mathbf{R}_1)$. Hence in Eq. **A5**, setting $\mathbf{R}_1 = \mathbf{I}$, $\mathbf{P} = \mathbf{Q} = (\sigma/\alpha)\mathbf{U}_1$, and $\mathbf{R}_2 = \mathbf{W}_1^2$, we obtain

$$\det(\mathbf{C}) = \alpha^{2N}\det\left(\mathbf{W}_1^{-2} + \frac{\sigma^2}{\alpha^2}\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_1\right)\det\left(\mathbf{W}_1^2\right). \quad \text{[A6]}$$

Therefore, because $\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_1 = \mathbf{I}$,

$$\begin{aligned}\log\det(\mathbf{C}) = 2N\log(\alpha) + \log\left(\det\left(\mathbf{W}_1^2\right)\right) \\ + \log\left(\det\left(\mathbf{W}_1^{-2} + \frac{\sigma^2}{\alpha^2}\mathbf{I}\right)\right).\end{aligned} \quad \text{[A7]}$$

To find the MLEs of α^2, σ^2 using Eq. **A3**, we must differentiate $\log\det(\mathbf{C})$ with respect to α^2 and σ^2 and set the derivatives to 0. The derivative of the first term is independent of $\mathbf{W}_1$, and the derivative of the second term is 0 because $\log(\det(\mathbf{W}_1^2))$ is independent of α^2 and σ^2. The last term in Eq. **A7** is

$$\log\left(\det\left(\mathbf{W}_1^{-2} + \frac{\sigma^2}{\alpha^2}\mathbf{I}\right)\right) = \log\left(\prod_{i=1}^{i=k}\left(\frac{1}{w_i^2} + \frac{\sigma^2}{\alpha^2}\right)\right). \quad \text{[A8]}$$

For a stratified population, thousands of singular values, w_i of $\mathbf{Z}$ will be close to 0, and Eq. **A8** will be extremely sensitive to small changes in the values of w_i

Instability of the third term in Eq. A3. From Eq. **A4**, we have

$$\mathbf{y}_1^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}_1 = \frac{1}{\alpha^2}\mathbf{y}_1^{\mathrm{T}}\mathbf{y}_1 - \frac{\sigma^2}{\alpha^4}\mathbf{y}_1^{\mathrm{T}}\mathbf{Z}\left(\mathbf{I} + \frac{\sigma^2}{\alpha^2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\underbrace{\mathbf{Z}^{\mathrm{T}}\mathbf{y}_1}. \quad \text{[A9]}$$

Consider just the factor with an underlying curly bracket. Suppose, we perturb $\mathbf{y}_1$ to $\mathbf{y}_1 + \gamma\mathbf{y}_2$ for some small γ. Then, that factor becomes

$$\mathbf{Z}^{\mathrm{T}}(\mathbf{y}_1 + \gamma\mathbf{y}_2) = \left(\mathbf{Z}^{\mathrm{T}} + \gamma\mathbf{E}^{\mathrm{T}}\right)\mathbf{y}_1, \quad \text{[A10]}$$

where we chose $\mathbf{E}^{\mathrm{T}}$ such that $\mathbf{E}^{\mathrm{T}}\mathbf{y}_1 = \mathbf{Z}^{\mathrm{T}}\mathbf{y}_2$ (the matrix $\mathbf{E}^{\mathrm{T}}$ can be trivially constructed to have elements only on its primary diagonal). Because a small perturbation of $\mathbf{Z}$ causes a large change in its spectral properties (26), the vector $(\mathbf{Z}^{\mathrm{T}} + \gamma\mathbf{E}^{\mathrm{T}})\mathbf{y}_1$ can be vastly different from $\mathbf{Z}^{\mathrm{T}}\mathbf{y}_1$, and hence $\mathbf{y}_1^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}_1$ is extremely sensitive to measurement errors in the phenotype.

Appendix B: The Likelihood Function and Bias. Here, we reformulate the likelihood function in terms of the eigenvalues [$a_i = (1/P)w_i^2$ for $1 \le i \le N$] and the eigenvectors ($\mathbf{u}_i$ for $1 \le i \le N$) of the GRM, $\mathbf{A} = (1/P)\mathbf{ZZ}^{\mathrm{T}}$. Because $\epsilon \sim N(0, \alpha^2\mathbf{I})$, we have

$$\mathbf{U}_1^{\mathrm{T}}\epsilon \sim N\left(0, \alpha^2\mathbf{I}\right). \quad \text{[A11]}$$

Using the SVD of $\mathbf{Z}$ described in *Singular Values and GCTA*, we have $\mathbf{A} = (1/P)\mathbf{U}_1\mathbf{W}^2\mathbf{U}_1^{\mathrm{T}}$. Now,

$$\mathbf{U}_1^{\mathrm{T}}\mathbf{g} \sim \mathbf{U}_1^{\mathrm{T}}\mathcal{N}\left(0, \sigma^2\mathbf{U}_1\mathbf{W}^2\mathbf{U}_1^{\mathrm{T}}\right) \sim \mathcal{N}\left(0, \mathbf{W}^2\right). \quad \text{[A12]}$$

Premultiplying Eq. **3** by $\mathbf{U}^{\mathrm{T}}$ and using Eq. **A12**,

$$\mathbf{U}_1^{\mathrm{T}}\mathbf{y}_1 = \mathbf{U}_1^{\mathrm{T}}\mathbf{g} + \mathbf{U}_1^{\mathrm{T}}\epsilon \sim \mathcal{N}\left(0, \alpha^2\mathbf{I} + \sigma^2\mathbf{W}\right). \quad \text{[A13]}$$

Setting the diagonal matrix, $\mathbf{S} = \alpha^2\mathbf{I} + \sigma^2\mathbf{W}^2$, we note that $\mathbf{U}_1\mathbf{SU}_1^{\mathrm{T}} = \mathbf{C}$, where $\mathbf{C}$ is defined in Eq. **A2**. Using Eq. **A13**, we can write the likelihood function as

$$\log P\left(\mathbf{y}_1|\alpha^2, \sigma^2\right) = \frac{-N}{2}\log(2\pi) - \log\det(\mathbf{S}) - \frac{1}{2}\mathbf{y}_1^{\mathrm{T}}\mathbf{U}_1\mathbf{S}^{-1}\mathbf{U}_1^{\mathrm{T}}\mathbf{y}_1. \quad \text{[A14]}$$

Because $\mathbf{U}_1$ is an orthogonal matrix, $\det(\mathbf{U}_1)\det(\mathbf{U}_1^{\mathrm{T}}) = 1$. Therefore, we have

$$\begin{aligned}\det(\mathbf{S}) &= \det(\mathbf{U}_1)\det\left(\mathbf{U}_1^{\mathrm{T}}\right)\det(\mathbf{S}) = \det(\mathbf{U}_1)\det(\mathbf{S})\det\left(\mathbf{U}_1^{\mathrm{T}}\right) \\ &= \det\left(\mathbf{U}_1\mathbf{SU}_1^{\mathrm{T}}\right) = \det(\mathbf{C}).\end{aligned} \quad \text{[A15]}$$

Using Eq. **A15** in Eq. **A14** and comparing with Eq. **A3**, we note that the first two terms in the likelihood function are identical. The only term whose derivative with respect to α^2 and σ^2 depends on the eigenvectors is the third term, and therefore we analyze this term in more detail. The eigenvalues of $\mathbf{S}$ will be $s_i = \alpha^2 + \sigma^2 w_i^2$ for $1 \le i \le N$. The last term in Eq. **A14** can be written as

$$J = \frac{1}{2}\mathbf{y}_1^{\mathrm{T}}\mathbf{U}_1\mathbf{S}^{-1}\mathbf{U}_1^{\mathrm{T}}\mathbf{y}_1 = \sum_i \frac{1}{2s_i}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_1\right)^2. \quad \text{[A16]}$$

Now, suppose that the GRM, $\mathbf{A}$, estimated by GCTA differs from the "true" underlying GRM ($\hat{\mathbf{A}}$) by a small sampling error $\mathbf{E}$. Using standard results from perturbation theory, we can express the eigenvectors of $\hat{\mathbf{A}}$ in terms of the spectral properties of $\mathbf{A}$ and the error matrix, $\mathbf{E}$, as

$$\tilde{\mathbf{u}}_i = \mathbf{u}_i + \sum_{j\neq i}\left(\frac{1}{a_i - a_j}\right)\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i = \mathbf{u}_i + \sum_{j\neq i}\left(\frac{P}{w_i^2 - w_j^2}\right)\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i, \quad \textbf{[A17]}$$

where $a_j = (1/P)w_j^2$ is the j^{th} eigenvalue of $\mathbf{A}$, w_j is the j^{th} singular value of $\mathbf{Z}$, and $\tilde{\mathbf{u}}_i$ and $\mathbf{u}_i$ are the i^{th} eigenvectors corresponding to $\tilde{\mathbf{A}}$ and $\mathbf{A}$, respectively. Using Eq. **A17** in Eq. **A16**, we get

$$\begin{aligned} J = &\sum_i \frac{1}{2s_i}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_i\right)^2 \\ &+ \sum_i \frac{1}{2s_i}\sum_{j\neq i}\frac{1}{(a_i - a_j)^2}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_j\right)^2\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i\right)^2 \\ &+ \sum_i \frac{1}{s_i}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_i\right)\sum_{j\neq i}\frac{1}{(a_i - a_j)}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_j\right)\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i\right). \end{aligned} \quad \textbf{[A18]}$$

Note that J is not symmetric in $\mathbf{E}$ (because the second term is not symmetric). Differentiating J with respect to σ^2, we have

$$\begin{aligned} \frac{\partial J}{\partial \sigma^2} = &\sum_i \frac{-w_i^2}{2s_i^2}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_i\right)^2 \\ &+ \sum_i \frac{-w_i^2}{2s_i^2}\sum_{j\neq i}\frac{1}{(a_i - a_j)^2}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_j\right)^2\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i\right)^2 \\ &+ \sum_i \frac{-w_i^2}{s_i^2}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_i\right)\sum_{j\neq i}\left(\mathbf{y}_1^{\mathrm{T}}\mathbf{u}_j\right)\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_i\right). \end{aligned} \quad \textbf{[A19]}$$

Note in Eq. **A19** that unless $\mathbf{E}$ is exactly 0, the MLE of σ^2 will always have a factor that does not average out to 0 (because the second term is not symmetric in $\mathbf{E}$), and therefore σ^2 is guaranteed to be biased. Similar derivations show that the MLE of α^2 are also guaranteed to be biased.

Case 1: There is no stratification in the population. When the assumptions of GCTA are met exactly, $\mathbf{Z}$ asymptotically has an N variate Wishart distribution with P degrees of freedom (15). Marčenko–Pastur theory (13) provides an empirical distribution (which we henceforth refer to as the M-P distribution) for the eigenvalues of the variance-covariance matrix (which in our case will be the GRM, $\mathbf{A}$) as a function of P/N, which although asymptotic, works well even for sample sizes as small as $N = 1{,}000$ (Fig. 1*B*). The distribution shows that most of the eigenvalues of the GRM will lie in 0–4. Note from Fig. 1*A* that when P/N is close to 1, the eigenvalues will be skewed and as the value of P/N becomes smaller, the eigenvalues become concentrated on smaller intervals. For example, when $P/N = 4$, the eigenvalues lie within 0.25–2.25 (Fig. 1*A*).

Because the M-P distribution is continuous, we assume without loss of generality that the eigenvalues are unique (i.e., there are no repeated eigenvalues). Because N is large, the eigenvalues of $\mathbf{A}$ are necessarily packed extremely close to one another. To be specific, for $N = 2{,}000$ and $N/P = 0.25$, the maximum value of the minimum separation between the eigenvalues will be $2/2{,}000 = 0.001$. This upper bound is not tight; our ballpark estimate assumes a (best-case) uniform spread of eigenvalues on the interval. In reality, the distribution is peaked near 0.5 (Fig. 1*A*), and therefore we expect the eigenvalues near 0.5 to be a lot closer than 0.001 (for our simulation in Fig. 1*B*, the minimum spacing between the eigenvalues is 5×10^{-4}).

This small separation causes the eigenvectors of the estimated GRM to be drastically different from those of the true GRM. To see why, consider an eigenvalue (say the p^{th} one) of $\mathbf{A}$, which is closely packed to another eigenvalue (say the q^{th} one). From Eq. **A17**, the angle between $\mathbf{u}_p$ and $\tilde{\mathbf{u}}_p$ will have terms of the form $(1/(a_p - a_q))\mathbf{u}_p^{\mathrm{T}}\mathbf{E}\mathbf{u}_q$, which will be large because $a_p - a_q$ is close to 0. This is just one of the terms; all eigenvalues that are "sticking" close to the p^{th} eigenvalue will induce large differences between the estimated and true eigenvectors.

These differences are amplified in the expressions for the MLE and its derivatives (see the second and third summations in Eqs. **A18** and **A19**). Therefore, the bias in the heritability estimates produced by GCTA can be large.

Case 2: The population is stratified. In a stratified population, there are two sources of bias in the GCTA estimates. The first source of bias is identical to that described in the case when there is no stratification for a stratified population, thousands of eigenvalues of the GRM are tightly packed near 0, and therefore from Eqs. **A18** and **A19**, there can be large errors associated with the MLEs produced by GCTA.

The second source of bias comes from the large errors associated with the eigenvalues of the GRM (26). Specifically, suppose there are sampling errors Δa_p and Δa_q associated with two closely packed eigenvalues whose true magnitudes are a_p and a_q, respectively. We have

$$\begin{aligned} \frac{1}{a_p + \Delta a_p - (a_q + \Delta a_q)} &= \frac{1}{a_p - a_q - (\Delta a_q - \Delta a_p)}, \\ &\approx \frac{1 + \kappa(p,q)}{a_p - a_q} \end{aligned} \quad \textbf{[A20]}$$

where $\kappa(p,q) = (\Delta a_q - \Delta a_p)/(a_p - a_q)$. Because the errors in the near-zero eigenvalues can be large, κ need not be close to 0. As a result, all of the error terms in Eqs. **A18** and **A19** will have additional amplification factors of the form $(1 + \kappa(p,q))$ for every pair (p and q) of near-zero eigenvalues of the GRM.

The bootstrap and GCTA. Here, we show that resampling techniques (e.g., the bootstrap, jackknife, etc.) cannot be used to improve the estimates produced by GCTA. Loosely put, the bootstrap estimates the parameter in question (in this case, the heritability) by resampling from the original sample and relying on the fact that the sampling errors "average out" to 0. We have shown in *Appendix B* that the GCTA estimates are overly (erroneously) biased by local information.

For run i of a bootstrap, let the heritability estimate obtained by GCTA be $\theta_g(i)$, the biasing error be $\kappa(i)$, the sampling error (with mean 0) be $\pi(i)$, and the "true" heritability estimate be θ_0. The i^{th} bootstrap estimate can be expressed as

$$\theta_g(i) = \theta_0 + \kappa(i) + \pi(i). \quad \textbf{[A21]}$$

Taking the mean in Eq. **A21** over sufficient bootstrap samples, we have

$$\mathrm{E}(\theta_g) = \theta_0 + \mathrm{E}(\kappa). \quad \textbf{[A22]}$$

The bootstrap will estimate $\mathrm{E}(\theta_g)$, which is not helpful because we have no handle on $\mathrm{E}(\kappa)$. More importantly, $\mathrm{E}(\theta_g)$ provides no useful information about θ_0.

Appendix C: Dynamics of GCTA and More Problems with Stratification. Consider the most general case, where N random people and P random SNPs are used to construct the random genotype matrix $\tilde{\mathbf{X}}$ (of which one realization is the observed data matrix $\mathbf{X}$). From $\tilde{\mathbf{X}}$, we construct the matrix $\tilde{\mathbf{Z}}$ by centering and scaling so that the entries of matrix $\tilde{\mathbf{Z}}$ will be i.i.d., with mean 0 and variance 1. As a result, $\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}$ and $\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}}$ will be random Wishart matrices (15), whose eigenvectors are known to be uniformly distributed over the unit sphere in N dimensions (41) which implies that with high probability, the eigenvectors of $\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}$ ($\{\tilde{\mathbf{u}}_i\}$ for $1 \le i \le k$) and $\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}}$ ($\{\tilde{\mathbf{v}}_i\}$ for $1 \le i \le k$) satisfy $|\tilde{\mathbf{u}}_i|_\infty = O(\sqrt{\log N/N})$ and $|\tilde{\mathbf{v}}_i|_\infty = O(\sqrt{\log P/P})$, respectively (42) (the ∞ norm of a vector $\mathbf{x} = \{x_1, x_2 \ldots x_n\}$ is

defined as $|\mathbf{x}|_\infty = \max(|x_1|, |x_2| \ldots |x_n|))$ [i.e., the maximum entries in the eigenvectors are very small (for $n = 2{,}698$ and $P = 49{,}214$, $(\log N/N) \approx 0.001$ and $(\log P/P) \approx 0.0001$)].

Suppose the first k_1 singular values of $\widetilde{\mathbf{Z}}$ are much larger than the others. We analyze $\mathbf{g} = \widetilde{\mathbf{Z}}\mathbf{u}$, the term primarily responsible for the high heritability estimates obtained using Eq. **3**. First, we express $\mathbf{g}$ as a function of the singular values and singular vectors of $\mathbf{Z}$, namely

$$\begin{aligned}\mathbf{g} &= \sum_{i=1}^{i=k_1} \widetilde{w}_i \widetilde{\mathbf{u}}_i \left(\widetilde{\mathbf{v}}_i^{\mathrm{T}} \mathbf{u}\right) + \sum_{i=k_1+1}^{i=k} \widetilde{w}_i \widetilde{\mathbf{u}}_i \left(\widetilde{\mathbf{v}}_i^{\mathrm{T}} \mathbf{u}\right) \\ &= \widetilde{Z}_1 \mathbf{u} + \sum_{i=k_1+1}^{i=k} \widetilde{w}_i \widetilde{\mathbf{u}}_i \left(\widetilde{\mathbf{v}}_i^{\mathrm{T}} \mathbf{u}\right).\end{aligned} \quad \textbf{[A23]}$$

In such cases, it is best to discard the near-zero singular values (26) while constructing the GRM. Accordingly, Eq. **A23** becomes

$$\mathbf{g} = \sum_{i=1}^{i=k_1} \widetilde{w}_i \widetilde{\mathbf{u}}_i \left(\widetilde{\mathbf{v}}_i^{\mathrm{T}} \mathbf{u}\right) = \widetilde{\mathbf{Z}}_1 \mathbf{u}. \quad \textbf{[A24]}$$

In this expression, note that $\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{u}$ will be very close to 0 (because the ∞ norm of $\widetilde{\mathbf{v}}_i$ is close to 0) for all $1 \le i \le k_1$, and therefore the random effect term will never be significant. Because the ∞ norm of $\widetilde{\mathbf{u}}_i$ for $1 \le i \le k_1$ is also close to 0, we expect $\widetilde{\mathbf{Z}}_1$ to be close to the 0 matrix, which is observed to be true in most cases.

GCTA tries to explain the variance of the phenotype vector $\mathbf{y}$ by k random projections onto a plane defined by the columns of $\mathbf{Z}$. In Eq. **A23**, we expressed these random projections using the left singular vectors of $\mathbf{X}$ as a basis. When N is fixed, $P \to \infty$ (i.e., as we collect more genotypic information on a fixed set of individuals) and one (or a few) of the singular values is much larger than the rest, the singular vector corresponding to these singular values will be consistent and the singular vectors corresponding to most of the remaining nonzero singular values will be strongly inconsistent (32) (informally put, consistency describes whether the singular vector estimates from the data matrix approach the "true" singular vector as more data are collected). This implies that the subspace defined by the higher singular vectors of the data matrix contain little biological information about the genotype matrix. [Similar inconsistency results hold for the case where $N \to \infty$, $P \to \infty$, and $P/N \to c$ for some constant c (43)].

In Eq. **A23**, we showed that the random projection of $\tilde{\mathbf{y}}$ onto the first k_1 singular vectors (given by $\widetilde{\mathbf{Z}}_1\mathbf{u}$) is almost surely 0. Because the subspace defined by the higher singular vector estimates $(\widetilde{\mathbf{u}}_{k_1+1}, \widetilde{\mathbf{u}}_3, \ldots . \widetilde{\mathbf{u}}_k)$ is inconsistent, there is little biological connection between $\sum_{i=k_1+1}^{i=k} \widetilde{w}_i \widetilde{\mathbf{u}}_i (\widetilde{\mathbf{v}}_i^{\mathrm{T}} \mathbf{u})$ in Eq. **A23** and the heritability estimate. It appears that it is this term that is responsible for the high values of the GCTA estimate (because the first term makes nearly 0 contribution to the projection). For every set of people and SNPs that are chosen, the data matrix generates a new set of $O(P)$ arbitrary singular vectors, which in turn generate arbitrary α^2 and σ^2 estimates.

ACKNOWLEDGMENTS. We thank Chiara Sabatti, Chris Gignoux, David Golan, David Steinsaltz, Edgar Dobriban, Jonathan Pritchard, and Kenneth Wachter for useful comments on earlier drafts of this paper. This project is funded by National Institutes of Health Grant AG22500 (to S.T.) and the Morrison Institute for Population and Resource Studies. S.K.K. is funded by the Stanford Center for Computational, Evolutionary and Human Genomics. D.H.R. is supported by National Institute on Aging Grant K01AG047280. Our use of the Framingham data has been approved by the Stanford Institutional Review Board.

1. Manolio TA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.
2. Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456(7218):18–21.
3. Weedon MN, et al.; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium; Cambridge GEM Consortium (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40(5):575–583.
4. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.
5. Lee SH, et al.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS) (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 44(3):247–250.
6. Davies G, et al. (2011) Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol Psychiatry* 16(10):996–1005.
7. Boardman JD, et al. (2014) Is the gene-environment interaction paradigm relevant to genome-wide studies? The case of education and body mass index. *Demography* 51(1):119–139.
8. Charney E (September 19, 2013) Still chasing ghosts: A new genetic methodology will not find the "missing heritability". *Independent Science News*.
9. Govindaraju DR, et al. (2008) Genetics of the Framingham heart study population. *Adv Genet* 62:33–65.
10. Splansky GL, et al. (2007) The third generation cohort of the national heart, lung, and blood institute's Framingham heart study: Design, recruitment, and initial examination. *Am J Epidemiol* 165(11):1328–1335.
11. VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91(11):4414–4423.
12. Robinson GK (1991) That blup is a good thing: The estimation of random effects. *Stat Sci* 6(1):15–32.
13. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Sbornik Mathematics* 1(4):457–483.
14. Wachter KW (1978) The strong limits of random matrix spectra for sample matrices of independent elements. *Ann Probab* 6(1):1–18.
15. Johnstone IM (2006) High dimensional statistical inference and random matrices. arXiv:math/0611589.
16. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
17. Casella G, Berger RL (2002) *Statistical Inference* (Duxbury, Pacific Grove, CA), Vol 2.
18. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. *Lancet* 361(9357):598–604.
19. Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *J Natl Cancer Inst* 92(14):1151–1158.
20. Heaton MP, et al. (2002) Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mamm Genome* 13(5):272–281.
21. Beraldi D, et al. (2007) Quantitative trait loci (QTL) mapping of resistance to strongyles and coccidia in the free-living Soay sheep (Ovis aries). *Int J Parasitol* 37(1):121–129.
22. Tian C, Gregersen PK, Seldin MF (2008) Accounting for ancestry: Population substructure and genome-wide association studies. *Hum Mol Genet* 17(R2):R143–R150.
23. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
24. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
25. Bryc K, Bryc W, Silverstein JW (2013) Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theor Popul Biol* 89:34–43.
26. Stewart GW (1998) *Perturbation theory for the singular value decomposition* (Institute for Advanced Computer Studies, University of Maryland, College Park, MD).
27. Kannel WB (2000) Risk stratification in hypertension: New insights from the Framingham Study. *Am J Hypertens* 13(1 Pt 2):3S–10S.
28. Barendse W (2011) The effect of measurement error of phenotypes on genome wide association studies. *BMC Genomics* 12(1):232.
29. Pickering TG, et al.; Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research (2005) Recommendations for blood pressure measurement in humans and experimental animals: Part 1: Blood pressure measurement in humans: A statement for professionals from the Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research. *Hypertension* 45(1):142–161.
30. Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: A comparison of pedigree and genomic approaches. *Mol Ecol* 23(14):3434–3451.
31. Johnstone IM, Lu AY (2009) On consistency and sparsity for principal components analysis in high dimensions. *J Am Stat Assoc* 104(486):682–693.
32. Jung S, Marron J (2009) PCA consistency in high dimension, low sample size context. *Ann Stat* 37(6B):4104–4130.
33. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning* (Springer, New York), Vol 2.
34. Trzaskowski M, Yang J, Visscher PM, Plomin R (2014) DNA evidence for strong genetic stability and increasing heritability of intelligence from age 7 to 12. *Mol Psychiatry* 19(3):380–384.
35. Davis LK, et al. (2013) Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* 9(10):e1003864.
36. de Candia TR, et al.; International Schizophrenia Consortium; Molecular Genetics of Schizophrenia Collaboration (2013) Additive genetic variation in schizophrenia risk is

shared by populations of African and European descent. *Am J Hum Genet* 93(3): 463–470.
37. Crossett A, Lee AB, Klei L, Devlin B, Roeder K (2013) Refining genetically inferred relationships using treelet covariance smoothing. *Ann Appl Stat* 7(2):669–690.
38. Speed D, Balding DJ (2015) Relatedness in the post-genomic era: Is it still useful? *Nat Rev Genet* 16(1):33–44.
39. Horn RA, Johnson CR (2012) *Matrix Analysis* (Cambridge Univ Press, Cambridge, UK).
40. Harville DA (1997) *Matrix Algebra from a Statistician's Perspective* (Springer, New York), Vol 157.
41. Bai Z, et al. (2007) On asymptotics of eigenvectors of large sample covariance matrix. *Ann Probab* 35(4):1532–1572.
42. Wang K (2013) Optimal upper bound for the infinity norm of eigenvectors of random matrices. PhD thesis (Rutgers, The State University of New Jersey, New Brunswick, NJ).
43. Johnstone IM, Lu AY (2004) Sparse principal components analysis. arXiv:0901.4392.