SNP-based heritability estimation: measurement noise, population stratification and stability

Eric R. Gamazon[1,2,*] and Danny S. Park[3]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, TN, USA

[2]Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

[3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA

[*]Correspondence to: Eric R. Gamazon <egamazon@uchicago.edu>

Siddharth Krishna Kumar [1] and co-authors claim to have shown that "GCTA applied to current SNP data cannot produce reliable or stable estimates of heritability." Given the numerous recent studies on the genetic architecture of complex traits that are based on this methodology, these claims have important implications for the field. Through an investigation of the stability of the likelihood function under phenotype perturbation and an analysis of its dependence on the spectral properties of the genetic relatedness matrix, our study characterizes the properties of an important approach to the analysis of GWAS data and identified crucial errors in the authors' analyses, invalidating their main conclusions.

Heritability estimation using genome-wide SNP data is a fundamental research topic with profound implications for studies of the genetic architecture of complex traits. The development of a novel methodology [2,3] in this direction has spurred studies, on a broad spectrum of complex traits, that have reinforced the view that a substantial portion of missing heritability can be accounted for by hitherto undiscovered common variants [4,5] and has led to substantial research that has demonstrated that certain functional categories of SNPs contribute disproportionately to the heritability of complex diseases [6-8]. However, in a recent report [1], Krishna Kumar and co-authors claim to have proved that the method "may not reliably improve our understanding of the genomic basis of phenotypic variability" even when the assumptions of the method are satisfied exactly and that the heritability estimates produced are highly sensitive to the choice of sample used and to measurement errors in the phenotype. We investigated these claims by characterizing the likelihood function and identified crucial analytic errors that seriously undermine the validity of the authors' conclusions.

The GREML model

We consider the following model (Figure 1A) of the phenotype y (which has been simplified, as in Krishna Kumar et al., to exclude any fixed-effects):

$$y = Zu + \varepsilon \qquad [\dagger]$$

where u is a $Px1$ vector of random (genetic) effects, Z is a NxP (standardized genotype) matrix and ε is the (non-genetic) residual. Here

$$u \sim N(0, \sigma^2 I)$$

$$y|u \sim N(Zu, \alpha^2 I)$$

Thus, the distribution of y assumes the following form:

$$y \sim N(0, \alpha^2 I + \sigma^2 ZZ^T)$$

Note that the phenotypic covariance, $var(y)$, is the sum of a genetic covariance and a residual covariance. The Genetic Relatedness Matrix (GRM), which quantifies the genetic similarity between pairs of individuals using the genotype data Z, can be written as follows:

$$A = ZZ^T/P$$

Singularity index and induced quadratic form

We refer to the function $\mathrm{S(Z)} := \log(\det(\alpha^2 I + \sigma^2 ZZ^T))$ as the *singularity index* (because it provides a formal test for the invertibility of the phenotypic covariance matrix $\alpha^2 I + \sigma^2 ZZ^T$) and refer to the function $Q(Z, y_1) := y_1^T(\alpha^2 I + \sigma^2 ZZ^T)^{-1} y_1$ as the *induced quadratic form*. Note the log-likelihood of the observed phenotype data y_1 is given by

$$l(\,\sigma^2, \alpha^2 \mid Z, y_1) = -\frac{N}{2} log(2\pi) - \log(\det(\alpha^2 I + \sigma^2 ZZ^T)) - \frac{1}{2} y_1^T(\alpha^2 I + \sigma^2 ZZ^T)^{-1} y_1 \quad [\dagger\dagger]$$

Using Restricted Maximum Likelihood (REML), GCTA estimates the variances σ^2 and α^2 given the observation y_1, thereby providing an estimate of the SNP-based heritability:

$$h^2_{SNP} = \frac{P\sigma^2}{P\sigma^2 + \alpha^2}$$

Equivalently, the log-likelihood function, now viewed as a function of Z and y_1, can be written as a sum involving the singularity index and the induced quadratic form:

$$f(\,Z, y_1 \mid \sigma^2, \alpha^2) = -\frac{N}{2} log(2\pi) - S(Z) - \frac{1}{2} Q(Z, y_1)$$

Perturbation of the standardized genotype matrix Z and the GRM A

Because the Z in the GREML model is a standardized genotype matrix (wherein each entry is a function of the number of copies of the reference allele and the reference allele frequency at a SNP), this implies that there are implicit constraints on what is a *valid* perturbed genotype matrix $Z + perturb(Z)$ (i.e., constraints which determine whether $Z + perturb(Z)$ is a realizable or ill-defined standardized genotype matrix). A perturbation matrix $perturb(Z)$ may generate a matrix that departs substantially from a standardized genotype matrix, yielding an ill-defined revised model. To illustrate this, if the original (e.g., independent, real and random) entries in Z have mean 0 and variance 1, a perturbation with elements on the primary diagonal due to the introduction of the phenotype noise $\vartheta \sim N(0, \tau^2)$ would preserve the mean of these elements but alter their variance, possibly quite substantially. In short, not every element of $\mathrm{Matrices(N, P)}$ represents a standardized genotype matrix, and not every perturbation is a reasonable one. For the same reason, a perturbation of the GRM (by an error matrix E, as in the authors' equation [A17] of the Appendix) does not necessarily generate a valid (revised)

GRM. (For example, the resulting perturbed GRM must be symmetric, which implies that the perturbation matrix E must be symmetric as well.) Furthermore, modeling the difference between the true Z and sample Z through an error matrix F via an additive model ($Z_{sample} = Z_{true} + F$) makes some very strong assumptions, including that the two matrices, Z_{true} and Z_{sample}, are of the same dimension (in particular, same number of variants). It is therefore more sound to evaluate the discordance between the true GRM (GRM_{true}) and the estimated GRM (GRM_{sample}). The impact of this discordance (arising, for example, from the imperfect tagging of causal variants [2,9]) on the REML estimate of heritability is indeed a valid subject of research [3]. Interestingly, this issue is related to the classic Horn's conjecture in matrix theory (which was finally settled [10]) on the spectrum of the sum of two Hermitian matrices and on how the eigenvalues of two Hermitian matrices constrain the eigenvalues of their sum.

A critique of the authors' claims

The authors evaluated the sensitivity of the likelihood function, and the resulting GREML estimate, to the GWAS data (specifically, phenotype measurement noise and population stratification). We report here crucial errors in the authors' analyses, on which the main conclusions of the study are based. Furthermore, we highlight a methodological gap, which we address using an approach that may be of interest to future studies in population genetics and GWAS of complex traits.

(We should note a random matrix theory for the Wishart product matrix ZZ^T (or the GRM) generally assumes a Z with independent Gaussian entries, and any application in genetics must demonstrate that the relevant theoretical results apply (robustly) to a (non-Gaussian) matrix (e.g., one consisting of standardized genotype data). The authors appear to claim, clearly incorrectly and rather confusingly, for both Z and its symmetrization ZZ^T a Wishart distribution (e.g., see pages E62 and E68 of the authors' paper [1]). In what follows, we will assume that Z is a standardized genotype matrix (and thus non-Gaussian), and Gaussian-based results that require extension to the non-Gaussian case will be explicitly stated.)

1. Sensitivity of third term of log-likelihood to phenotype noise

The authors sought to show the instability of the induced quadratic form $Q(Z, y_1)$, and thus of the log-likelihood, by showing its sensitivity to the phenotype measurement (i.e., to a perturbation of y_1). In their analysis, this conclusion follows from the instability of the spectral properties of Z even under a "small perturbation." The authors used the following "equivalence" of perturbations (see equation [A10] of their Appendix A) – namely, the perturbation to the phenotype measurement and the induced perturbation of the matrix Z:

$$Z^T\left(y_1 + perturb(y_1)\right) = (Z^T + perturb(Z^T))y_1 \qquad \text{[e1]}$$

Applying the Sherman-Morrison-Woodbury identity to the third term of the log-likelihood (equation [††]), one obtains

$$Q(Z, y_1) \coloneqq y_1^T(\alpha^2 I + \sigma^2 ZZ^T)^{-1}y_1 = y_1^T\left(\frac{1}{\alpha^2}I - \frac{\sigma^2}{\alpha^4}Z\left(I + \frac{\sigma^2}{\alpha^2}Z^TZ\right)Z^T\right)y_1$$

$$= \frac{1}{\alpha^2}y_1^Ty_1 - \frac{\sigma^2}{\alpha^4}y_1^TZ\left(I + \frac{\sigma^2}{\alpha^2}Z^TZ\right)\underbrace{Z^Ty_1}$$

Thus, the sensitivity, assuming phenotype perturbation (equation [e1]), depends not only on the factor with an underlying bracket (i.e., the spectral properties of Z), but also on the remaining terms (including $y_1^T y_1$). (The authors highlighted the former and, curiously, disregarded the latter.) Ignoring these remaining terms may yield invalid inferences concerning $Q(Z, y_1)$. Importantly, $Q(Z, y_1)$ is an $\mathbb{R}$-valued continuous (i.e., well-behaved and stable) function at *every* $(Z_0, y_0) \in \mathrm{Matrices(N, P)x}\mathbb{R}^N$, i.e.,

$$\forall \epsilon > 0, \exists \delta > 0 \; so \; that \; |Q(Z,y) - Q(Z_0, y_0)| < \epsilon \text{ whenever } d\big((Z,y),(Z_0,y_0)\big) < \delta$$

where $d\colon \mathrm{Matrices(N, P)x}\mathbb{R}^N \oplus \mathrm{Matrices(N, P)x}\mathbb{R}^N \to \mathbb{R}$ is the distance function defined by:

$$d\big((Z,y),(Z_0,y_0)\big) := \sqrt{(||Z - Z_0||_{Frobenius})^2 + (y - y_0)^T(y - y_0))} \qquad [§]$$

Here, for $M \in \mathrm{Matrices(N, P)}$, $||\mathrm{M}||_{Frobenius} := \sqrt{\sum_{i=i}^{N}\sum_{j=1}^{P}|m_{ij}|^2}$. The metric in equation [§] endows the space $\mathrm{Matrices(N, P)x}\mathbb{R}^N$ with the topology of a Euclidean space (homeomorphic to $\mathbb{R}^{NP+N}$) on which $Q(Z, y)$, consisting of sums and products of continuous functions, is continuous. Similarly, the proper subset $\{(Z, y)|\; Z \text{ a standardized genotype matrix and } y \text{ a phenotype vector}\} \subseteq \mathrm{Matrices(N, P)x}\mathbb{R}^N$, which is embeddable into $\mathbb{R}^{NP+N}$ via the canonical inclusion, gets an induced subspace topology on which $Q(Z, y)$ is continuous.

Given a fixed matrix Z, we ask how a perturbation in y_1 changes $Q(Z, y_1)$. The rate of change in Q with respect to (the vector) y_1 is given by the gradient:

$$\frac{\partial Q}{\partial y_1} = \frac{1}{\alpha^2} y_1^T(2I) + \frac{\sigma^2}{\alpha^4} y_1^T(M + M^T), \text{ with } M = \; Z\left(I + \frac{\sigma^2}{\alpha^2}Z^T Z\right)Z^T = PA + \frac{\sigma^2}{\alpha^2}P^2A^2$$

This simplifies to the following expression (by symmetry of M):

$$\frac{\partial Q}{\partial y_1} = y_1^T\{\left(2\frac{1}{\alpha^2}I\right) + \frac{\sigma^2}{\alpha^4}(M + M^T)\} = 2y_1^T\{\left(\frac{1}{\alpha^2}I\right) + \frac{\sigma^2}{\alpha^4}M\}$$

which allows us to quantify the l^2-norm $||\frac{\partial Q}{\partial y_1}||$ as a function of (the perturbed) y_1. Because $\mathrm{S(Z)}$ does not depend on y_1, this also gives the rate of change of the entire log-likelihood with respect to the phenotype vector. Furthermore, the expression for $\frac{\partial Q}{\partial y_1} = y_1^T \widetilde{M}$ shows that $Q \in$ C^1, i.e., it is actually continuously differentiable as a function of y_1. $\frac{\partial Q}{\partial y_1}$ involves a second-degree polynomial in the GRM A and is therefore continuous as a function of the GRM. Finally:

$$\frac{\partial^2 Q}{\partial^2 y_1} = 2\{\left(\frac{1}{\alpha^2} I\right) + \frac{\sigma^2}{\alpha^4} M\}$$

which does not vary with the phenotype vector.

Consistent with the continuity of the function Q in Z and y and the (stable and “linear”) rate of change in Q with respect to y_1, simulations we performed confirm the stability of the GREML estimate (Figure 1B). We note that, in fact, both terms ($Q(Z, y_1)$ and $\mathrm{S(Z)}$) of the log-likelihood are continuous functions at *every* $(Z_0, y_0) \in \mathrm{Matrices(N, P)x}\mathbb{R}^N$.

The authors’ figure 5, which was intended to show the variation in the GREML estimates from random sampling from repeated measures of a phenotype, is *not* unexpected and, furthermore, does not empirically support the flawed theoretical argument about the instability of the log-likelihood.

2. *Stability of second term of log-likelihood in stratified population*

The authors also sought to show the instability of the singularity index $\mathrm{S(Z)}$ in a stratified population. Using the singular value decomposition (SVD) of Z ($Z = U_1 W_1 V_1^T$) and applying the Matrix determinant lemma, one obtains the following decomposition:

$$\mathrm{S}(\mathrm{Z}) = 2Nlog(\alpha) + \log(\det(W_1^2)) + \log(\det(W_1^{-2} + \frac{\sigma^2}{\alpha^2} I)) \qquad [e2]$$

The last term of equation [e2] can be written in terms of the singular values w_i of Z as $\log(\prod_{i=1}^{k}(\frac{1}{w_i^2} + \frac{\sigma^2}{\alpha^2}))$. From this, the authors concluded (incorrectly, as we will see) that in a stratified population (for which, it is claimed, thousands of the w_i are close to 0), this expression for the last term of [e2] (and thus the entire expression itself) is sensitive to small changes in the values of the w_i. However, one cannot show the instability of the singularity index without also considering the rest of the terms in equation [e2]. Indeed, equation [e2] can be rewritten as follows:

$$\mathrm{S}(\mathrm{Z}) = 2Nlog(\alpha) + \log\left(\prod_{i=1}^{k}(w_i^2)\right) + \log\left(\prod_{i=1}^{k}\left(\frac{1}{w_i^2} + \frac{\sigma^2}{\alpha^2}\right)\right)$$

$$= 2Nlog(\alpha) + \sum_{i=1}^{k}(\log(w_i^2) + \log(\frac{1}{w_i^2} + \frac{\sigma^2}{\alpha^2}))$$

$$= 2Nlog(\alpha) + \textstyle\sum_{i=1}^{k} \log(1 + w_i^2 \frac{\sigma^2}{\alpha^2}) \qquad [e3]$$

For singular values w_i of Z that are close to 0, $\log\left(1 + w_i^2 \frac{\sigma^2}{\alpha^2}\right) \approx w_i^2 \frac{\sigma^2}{\alpha^2}$ (based on the Taylor series expansion). Thus, the sampling variability (from the expression for $\mathrm{S}(\mathrm{Z})$; see equation [e3]) for near-zero singular values does not arise from the terms $\frac{1}{w_i^2}$ (as the authors claim), but from w_i^2. Such near-zero singular values should add little to the singularity index and closely-packed singular values (i.e., for which $w_i \approx w$, for some constant w) should affect $\mathrm{S}(\mathrm{Z})$ nearly similarly, and thus the claim that near-zero singular values lead to unreliable estimates of the variance explained by all SNPs ($= P\sigma^2$) remains unfounded. In contrast, very large eigenvalues (such as reflecting non-random population structure) affect the stability of the index, with the rate of change of the index, $\rho_i := \frac{\partial s}{\partial w_i}$, with respect to w_i given by the following expression:

$$\rho_i = \frac{2w_i}{1 + w_i^2 \frac{\sigma^2}{\alpha^2}} (\frac{\sigma^2}{\alpha^2})$$

which, at $w_i = \infty$, is approximately $\frac{1}{w_i}$. Thus, the marginal effect of increasing singular value on the index decays at infinity in a manner inversely proportional to the magnitude of the singular value. The rate-vector ρ is informative about the behavior of the index at extreme singular values. Clearly, $\lim_{w_i \to 0} \rho_i = 0$, which implies that the rate of change becomes almost negligible for singular values near 0.

As we have already noted, the singularity index is also a continuous function at each $(Z_0, y_0) \epsilon \text{ Matrices(N, P)x}\mathbb{R}^N$ and, by projection to the first coordinate, a continuous function of the matrix Z. Related to this, the classical Weyl's inequality [11] implies that, given

$$GRM_{sample} = GRM_{true} + E$$

with $||\text{E}||_{Frobenius} < \varepsilon$, then

$$|w_{i,sample} - w_{i,true}| < \varepsilon$$

i.e., small perturbations in GRM yield only small perturbations in singular values. Equivalently, the following functions:

$$\pi_i: \{ZZ^T/P |\ Z \text{ a standardized genotype matrix}\} \subseteq Sym(N) \to \mathbb{R}$$

$$\pi_i(ZZ^T/P) = w_i(Z)$$

which map a (potential) GRM matrix (which is an element of the set of NxN symmetric matrices $Sym(N)$) to the i-th singular values (for $1 \leq i \leq P$) of the corresponding standardized genotype matrix are continuous. Thus, under an additive model in which the true GRM differs from the sample GRM by a perturbation E whose Frobenius norm is small, the difference in the corresponding singular values between the GRMs will be correspondingly small.

3. *Methodological gap*

What is notably missing from the authors' analysis, given its use of the eigenvalues ($\frac{w_i^2}{P}, 1 \leq i \leq N$) of the GRM (from the SVD) to evaluate the stability of the GREML approach, is a quantification of the degree to which the eigenvalues reflect non-random population structure versus random expectation. A large eigenvalue may well be "within null expectation," and there is thus a need to quantify its significance. (Note this is different from the empirical distribution of the GRM eigenvalues as presented in the authors' figure 1, which aimed to show, despite the small sample sizes considered, concordance of the data with the asymptotic behavior of eigenvalues from the Marchenko-Pastur theory.) Consideration of the null is also missing from the authors' appropriation of the notion of an "ill-conditioned" matrix Z, which is defined in terms of the condition number $\kappa = \frac{\max_i(w_i)}{\min_i(w_{i)}}$, as an approach for investigating the effect on GREML estimates. In addition to these key methodological gaps, it is important to note that κ is a property of the matrix Z rather than of the GREML method. Indeed, a very large κ would also affect effect size estimation in simple linear regression (e.g., equation [†]) that jointly fits multiple SNPs as fixed effects; a very large κ would imply that even a small change in

y could have a destabilizing impact on the estimated SNP effect sizes and that matrix inversion would be unstable with finite-precision numbers.

The distribution of the largest eigenvalue of the Wishart matrix of a matrix Z with independent Gaussian entries is known [12]. For large values of N and P, if λ denotes the largest eigenvalue, then $\frac{\lambda-\mu(\mathrm{N,P})}{\sigma(N,P)}$ assumes the Tracy-Widom distribution [13]; here both the centering constant $\mu(\mathrm{N,P})$ and the scaling constant $\sigma(N,P)$ depend on only N and P. If the following assumptions are met for the symmetrization $ZZ^T = [s_{ij}]$ in the GREML model (where now Z is the standardized genotype matrix with non-Gaussian entries):

(a) the (independent real random) entries have mean 0 and variance 1

(b) all moments of these random variables are finite

(c) $E(s_{ij})^{2m} \leq m^m$, for some constant m (i.e., the distributions of the entries decay at least as fast as a Gaussian distribution)

Soshnikov's extension theorem [14] implies that the ratio $\frac{\lambda-\mu(\mathrm{N,P})}{\sigma(N,P)}$, for some centering and scaling constants that depend only on N and P, converges in distribution to the Tracy-Widom distribution, just as in the Wishart case. The ratio thus provides a way to assess the significance of the largest eigenvalue of a GRM and to quantify the presence of non-random population structure in the genotype data [15]. (For example, using the Framingham dataset presented in the authors' figure 3, one concludes that the dataset shows extreme population stratification, $p<2.2 \times 10^{-16}$.) Exact expressions for the density and the moments of the distribution of the smallest eigenvalue (in terms of polynomials, exponentials and hypergeometric functions) for a matrix with independent Gaussian entries have been derived, and, interestingly, the form of this distribution depends on whether $P - N$ is odd or even [16]. Additionally, the work of Edelman provides a closed form for the distribution of the condition number κ. Indeed, for Z with independent standard-Gaussian entries and large N [16], we can write

$$P(\kappa(ZZ^T)/N < x) = P(\kappa(Z)^2/N < x) \approx e^{-\frac{2}{x}-\frac{2}{x^2}}$$

providing an asymptotic distribution for $\kappa(Z)$. The claims made by the authors concerning the stability of the GREML estimates such as through their use of the skew in singular values (such as the “Largest Singular Value” of Figure 3 and the discussion thereof in the text) are, as currently presented, statistically problematic without consideration of what is expected under the null.

Conclusions

We investigated the properties of the log-likelihood to evaluate the dependence of the GREML estimate on phenotype perturbation and on the spectral properties of the standardized genotype matrix. We showed the continuity of the singularity index and the induced quadratic form as functions of the standardized genotype matrix and the phenotype vector, supporting the stability of the log-likelihood under perturbation. Furthermore, we derived an explicit expression for the rate of change in the log-likelihood with respect to the phenotype vector. We examined the sensitivity to changes in the singular values, showing that the authors’ claims regarding the impact of sampling variability for near-zero singular values on the GREML estimate were based on an analytic error (and indeed assumed an incorrect view of the structure of genetic relatedness under population stratification). (It should be noted that the observation that population structure, which may be reflected in the largest eigenvalues of the

GRM, may confound heritability estimation, and must thus be adjusted for, has been repeatedly discussed and investigated [17,18].) Finally, we investigated a methodological gap in the authors' study and highlighted an approach to address it, which may be of broad interest to methods development in population genetics and genome-wide association analysis.

References

1 Krishna Kumar, S., Feldman, M. W., Rehkopf, D. H. & Tuljapurkar, S. Limitations of GCTA as a solution to the missing heritability problem. *Proc Natl Acad Sci U S A* **113**, E61-70, doi:10.1073/pnas.1520109113 (2016).

2 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569, doi:10.1038/ng.608 (2010).

3 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).

4 Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44**, 247-250, doi:10.1038/ng.1108 (2012).

5 Davis, L. K. *et al.* Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* **9**, e1003864, doi:10.1371/journal.pgen.1003864 (2013).

6 Gamazon ER, Im HK, Liu C, Members of the Bipolar Disorder Genome Study (BiGS) Consortium, Nicolae DL, Cox NJ. The Convergence of eQTL Mapping, Heritability Estimation and Polygenic Modeling: Emerging Spectrum of Risk Variation in Bipolar Disorder. *arxiv*, arXiv:1303.6227 (2013).

7 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235, doi:10.1038/ng.3404 (2015).

8 Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* **95**, 535-552, doi:10.1016/j.ajhg.2014.10.004 (2014).

9 Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-1021, doi:10.1016/j.ajhg.2012.10.010 (2012).

10 Knutson, A. & Tao, T. The honeycomb model of GL(n)(C) tensor products I: Proof of the saturation conjecture. *J Am Math Soc* **12**, 1055-1090, doi:Doi 10.1090/S0894-0347-99-00299-4 (1999).

11 Weyl, H. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung. *Mathematische Annalen* **71**, 441-479 (1912).

12 Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* **29**, 295-327, doi:DOI 10.1214/aos/1009210544 (2001).

13 Tracy, C. A. & Widom, H. Level-Spacing Distributions and the Airy Kernel. *Commun Math Phys* **159**, 151-174, doi:Doi 10.1007/Bf02100489 (1994).

14 Soshnikov, A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J Stat Phys* **108**, 1033-1056, doi:Unsp 0022-4715/02/0900-1033/0

Doi 10.1023/A:1019739414239 (2002).

15 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

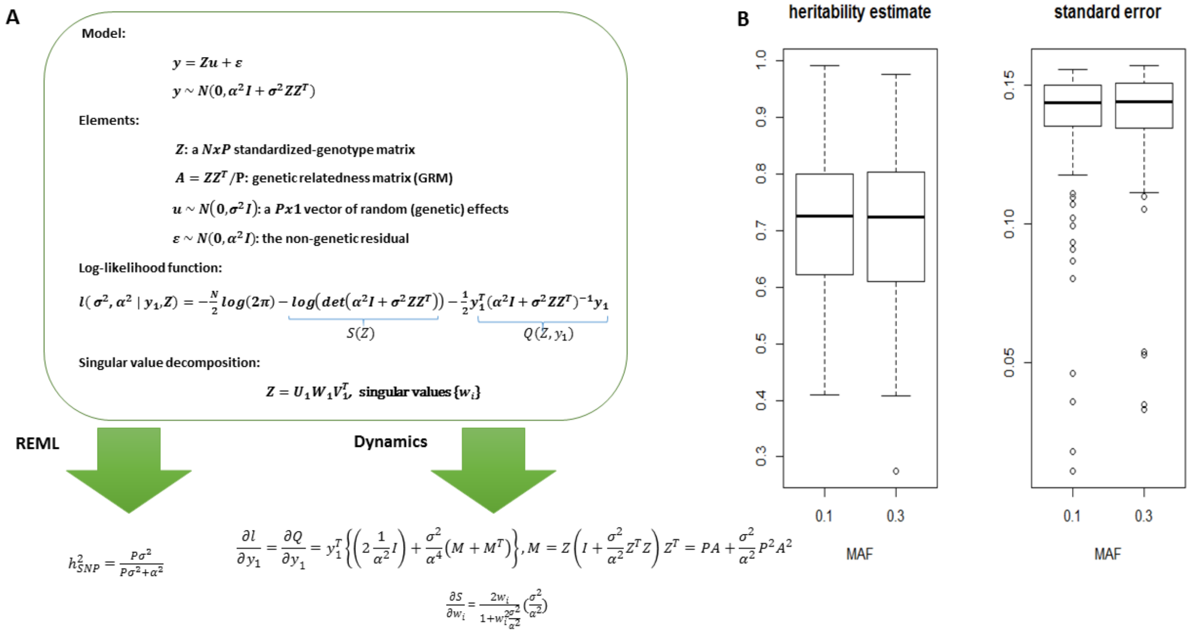
16 Edelman, A. Eigenvalues and Condition Numbers of Random Matrices. *Siam J Matrix Anal A* **9**, 543-560, doi:Doi 10.1137/0609045 (1988).

17 Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520, doi:10.1371/journal.pgen.1003520 (2013).

18 Browning, S. R. & Browning, B. L. Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet* **89**, 191-193; author reply 193-195, doi:10.1016/j.ajhg.2011.05.025 (2011).

Figure Legend

Figure 1. **A)** The GREML model (which underlies the GCTA software implementation) has been simplified here, as in the Krishna Kumar et al. study, to exclude fixed effects. Note that the phenotypic covariance decomposes into a genetic covariance and a residual covariance. The Genetic Relatedness Matrix (GRM), which quantifies the genetic similarity between pairs of individuals using the genotype data Z, can be written as $A = ZZ^T/P$. Using Restricted Maximum Likelihood (REML), GCTA provides estimates of σ^2 and α^2 and thus of the SNP-based heritability: $h_{SNP}^2 = \frac{P\sigma^2}{P\sigma^2+\alpha^2}$. The dynamics of the log-likelihood can be investigated by considering the rate of change with respect to a perturbation in the phenotype vector (i.e., the gradient) or a perturbation in the GRM (through an analysis of the singular values). **B)** We performed simulations, assuming $\mathrm{N} = 2{,}000$ unrelated individuals, $\mathrm{P} = 50{,}000$ independent SNPs and $h^2 = 0.75$. For each value of the minor allele frequency (MAF) ϵ {0.10, 0.30}, we generated the matrix Z by drawing from the binomial distribution $Bin(2, maf)$ and standardizing (i.e., by centering and scaling) the entries. We simulated 100 phenotypes for each MAF. The genetic effects u were drawn from the standard normal $N(0,1)$. We used the generative model described in (A) and the necessary residual to arrive at the required level of heritability. The distribution of GREML estimates for h^2 and corresponding standard error is shown for each MAF.



Model:

$$y = Zu + \varepsilon$$

$$y \sim N(0, \alpha^2 I + \sigma^2 ZZ^T)$$

Elements:

Z: a NxP standardized-genotype matrix

$A = ZZ^T/\mathrm{P}$: genetic relatedness matrix (GRM)

$u \sim N(0, \sigma^2 I)$: a $Px1$ vector of random (genetic) effects

$\varepsilon \sim N(0, \alpha^2 I)$: the non-genetic residual

Log-likelihood function:

$$l(\sigma^2, \alpha^2 \mid y_1, Z) = -\frac{N}{2} log(2\pi) - \underbrace{log\left(det\left(\alpha^2 I + \sigma^2 ZZ^T\right)\right)}_{S(Z)} - \frac{1}{2}\underbrace{y_1^T(\alpha^2 I + \sigma^2 ZZ^T)^{-1} y_1}_{Q(Z, y_1)}$$

Singular value decomposition:

$$Z = U_1 W_1 V_1^T, \text{ singular values } \{w_i\}$$

$$h_{SNP}^2 = \frac{P\sigma^2}{P\sigma^2 + \alpha^2}$$

$$\frac{\partial l}{\partial y_1} = \frac{\partial Q}{\partial y_1} = y_1^T\left\{\left(2\frac{1}{\alpha^2} I\right) + \frac{\sigma^2}{\alpha^4}(M + M^T)\right\}, M = Z\left(I + \frac{\sigma^2}{\alpha^2} Z^T Z\right) Z^T = PA + \frac{\sigma^2}{\alpha^2} P^2 A^2$$

$$\frac{\partial S}{\partial w_i} = \frac{2w_i}{1 + w_i^2\frac{\sigma^2}{\alpha^2}}\left(\frac{\sigma^2}{\alpha^2}\right)$$

Author contributions

E.R.G. designed the study, performed the research and wrote the paper. D.S.P. performed the research. Both authors reviewed and approved the final manuscript.

Acknowledgments

E.R.G. acknowledges support from R01 MH101820, R01 MH090937 and R01 CA157823.