Noname manuscript No.
(will be inserted by the editor)

Conditions for the validity of SNP-based heritability estimation

James J. Lee · Carson C. Chow

Received: date / Accepted: date

Abstract The heritability of a trait (h^2) is the proportion of its population variance caused by genetic differences, and estimates of this parameter are important for interpreting the results of genome-wide association studies (GWAS). In recent years, researchers have adopted a novel method for estimating a lower bound on heritability directly from GWAS data that uses realized genetic similarities between nominally unrelated individuals. The quantity estimated by this method is purported to be the contribution to heritability that could in principle be recovered from association studies employing the given panel of SNPs (h^2_{SNP}). Thus far the validity of this approach has mostly been tested empirically. Here, we provide a mathematical explication and show that the method should remain a robust means of obtaining h^2_{SNP} under circumstances wider than those under which it has so far been derived.

Keywords heritability · unrelated individuals · quantitative genetics · genome-wide association studies

James J. Lee · Carson C. Chow (✉)
Laboratory of Biological Modeling
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Bethesda, MD 20892, USA
(301) 402-8250
E-mail: carsonc@mail.nih.gov

James J. Lee (✉)
Department of Psychology
University of Minnesota Twin Cities
Minneapolis, MN 55455, USA
(612) 625-4980
E-mail: leex2293@umn.edu

James J. Lee
BGI Cognitive Genomics Lab
Building No. 11, Bei Shan Industrial Zone
Yantian District, 518083 Shenzhen, China

Introduction

A central question in the study of quantitative phenotypic variation is the extent to which such variation is caused by genetic differences. The precise proportion of the phenotypic variance ascribable to genetic differences is formally known as the heritability. Many definitions of heritability have been proposed (Bell, 1977), but in this work we employ the *narrow-sense heritability* commonly denoted by h^2 (Visscher et al, 2008). The concept of heritability was introduced by Fisher (1918) and Wright (1921) in their papers laying the foundations of quantitative genetics, although they did not use the word "heritability" in these early writings.

Its success notwithstanding, the imminent demise of quantitative genetics as a field of research has been repeatedly predicted ever since its first textbook appeared (Falconer, 1960; Hill and Mackay, 2004). Perhaps the prominent role in quantitative-genetic theory of heritability—a macroscopic parameter of a genetic system—has led some to suppose that advancing microscopic knowledge of the genetics underlying a given trait will superannuate the high-level approach. This anticipated obsolescence has not occurred, and indeed the recent explosion of findings from genome-wide association studies (GWAS) has only intensified the spotlight on the concept of heritability. For example, the loci found to be associated with a given trait at a strict threshold of statistical significance typically account for only a small proportion of the trait's heritability (as estimated from traditional studies of the correlations between close relatives), and this discrepancy has led to much discussion of "missing heritability" (Manolio et al, 2009; Eyre-Walker, 2010; Dickson et al, 2010; Wray et al, 2011; Zuk et al, 2012; Gibson, 2012; Hemani et al, 2013).

The estimation of heritability from the correlations between relatives has been substantially augmented by a novel technique that makes use of dense GWAS data from nominally unrelated individuals (Yang et al, 2010; Visscher et al, 2010; Lee et al, 2011). This technique is perhaps the most important innovation in quantitative genetics to have been introduced in the last dozen years, and it has provided what some may regard as decisive evidence for the view that undiscovered common variants account for a substantial portion of missing heritability. We will follow Benjamin et al (2012) and refer to this method as *genomic-relatedness-matrix restricted maximum likelihood* (GREML).

The descriptions of GREML given in the literature suggest that the mathematical basis of this method is not fully understood. For instance, formal justifications of the method that have been offered so far seem unable to account for biased estimates occasionally observed in simulation studies. Here we attempt to further the mathematical understanding of GREML, thereby providing insight into cases where the method has not worked well. By revealing the sense in which these cases are extreme, however, our account conversely shows that GREML estimates are in fact quite robust. We treat the heritability of a single trait, but our account can be generalized to the genetic correlation between two traits.

Table 1 SNPs used in simulations of GREML performance in the case of one nonzero

SNP i	chromosome	MAF (f_i)	$\sum_j \mathbf{C}_{ij}$
very weakly tagged nonzeros			
rs4674229	2	0.0119	2.08
rs4716447	7	0.2460	2.02
rs4968679	17	0.4978	2.09
weakly tagged nonzeros			
rs11039838	11	0.0108	5.01
rs4912830	5	0.2514	5.04
rs2654534	15	0.4955	5.02
moderately tagged nonzeros			
rs7620645	3	0.0176	11.49
rs12692474	2	0.2526	11.42
rs4870308	6	0.4915	11.41
strongly tagged nonzeros			
rs16841231	2	0.0156	16.54
rs6424728	1	0.2505	16.61
rs328890	7	0.4926	16.58
very strongly tagged nonzeros			
rs10138824	14	0.0278	30.36
rs2718306	7	0.2521	30.37
rs8006587	14	0.4910	30.34

Subjects and methods

We emphasize here that some of our mathematical arguments employ restrictive assumptions about sample size, the number of genotyped markers, and the values of the variance components. However, the fact that certain assumptions are sufficient to prove a result does not imply that the assumptions are necessary, and later we provide strong evidence for the generality of our findings.

We illustrate some of our mathematical arguments with numerical simulations, using two GWAS datasets to supply the genetic data. One dataset was used in a GWAS of European Americans reported previously (Chabris et al, 2013). The quality-control filters left 401 individuals and 661,108 markers (although only subsets of markers on chromosome 1 were used). We employed this small-sample dataset when it was necessary to relieve computational burden.

The second dataset was taken from the GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study/Health Professionals Follow-Up Study). We used PLINK to eliminate individuals of reported non-European descent, markers missing more than 5 percent of their calls, markers showing significant deviation from Hardy-Weinberg equilibrium (HWE) ($p < 1 \times 10^{-6}$), markers with minor allele frequency (MAF) < 0.01, individuals missing more than 5 percent of their genotypes, and one individual from any pair with a relatedness (Eq. 5) exceeding 0.025 in absolute value; Zaitlen et al (2013) provide some discussion of the appropriate relatedness cutoff. These filters left 4,975 individuals and 697,709 markers.

We used the software tool LDAK to calculate the extent to which each SNP is tagged by its neighbors (Speed et al, 2012). In particular, we computed the

matrix

$$\mathbf{C}_{ij} = \begin{cases} e^{-\lambda d_{ij}} r_{ij}^2 & \text{if } e^{-\lambda d_{ij}} > .125 \text{ and } r_{ij}^2 > .01, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where d_{ij} is the distance between SNPs i and j in base pairs (equaling ∞ if the SNPs are on different chromosomes), r_{ij}^2 is the standard measure of linkage disequilibrium (LD) between i and j, and λ is chosen so that $\exp(-\lambda d_{ij}) = .125$ when i and j are 3 Mbp apart. SNP i's level of tagging by its neighbors is then the sum of the elements in the ith row of $\mathbf{C}$. Large values of $\mathbf{C}$ correspond to strong tagging (redundancy), whereas small values correspond to a lack of LD with neighbors.

Table 1 lists the markers used in the simulations testing the case of a single marker with a nonzero partial regression coefficient. The mean of $\sum_j \mathbf{C}_{ij}$ over all i was approximately 11.45 and the standard deviation approximately 8.35, and we chose three SNPs with values close to the mean as the "moderately tagged" SNPs. Similarly, we choose three SNPs close to the 3rd percentile (2.05) as the "very weakly tagged" SNPs, three SNPs close to the 20th percentile (5.02) as the "weakly tagged" SNPs, three SNPs close to the 80th percentile (16.58) as the "strongly tagged" SNPs, and three SNPs close to the 97th percentile (30.36) as the "very strongly tagged" SNPs. Within each group of three markers, one was chosen to have an MAF of ~ 0.01, another to have an MAF of ~ 0.25, and the last to have an MAF of ~ 0.50. More specifically, for all markers within a given percentile of the $\sum_j \mathbf{C}_{ij}$ distribution plus/minus 0.05, one random selection was made from the markers with MAF in the interval (0.01, 0.02), another from markers in the interval (0.245, 0.255), and yet another from the interval (0.49, 0.50) to create a set of three markers varying in MAF but matched with respect to LD. The extent of tagging by neighbors is moderately correlated with MAF, and there were no candidates for very strongly tagged markers meeting the initial requirement for low MAF. The right endpoint of the low-MAF interval was therefore extended by increments of .01 until the set of candidates was nonempty.

We used GCTA to simulate phenotypes and estimate heritabilities on the basis of GREML. Each simulation scenario was tested with 200 replicates.

Results

Consider a sample of n unrelated (very distantly related) individuals and p biallelic markers. Let $\mathbf{y} \in \mathbb{R}^n$ be the vector of standardized phenotypes, $\mathbf{e} \in \mathbb{R}^n$ the vector of residuals (the sum of non-additive genetic deviations, environmental deviations, and errors of measurement), $\mathbf{Z} \in \mathbb{R}^{n \times p}$ the matrix of standardized genotypes, and $\mathbf{u} \in \mathbb{R}^p$ the vector of partial regression coefficients in the regression of the phenotype on standardized genotypes. If X_{ik} is the count of minor alleles (0, 1, or 2) carried by individual i at marker k, then HWE implies that the standardized count is $Z_{ik} = (X_{ik} - 2f_k)/\sqrt{2f_k(1-f_k)}$, where f_k is the MAF at marker k. Note that the elements of $\mathbf{u}$ are not necessarily

proportional to the average effects of gene substitution (Fisher, 1941; Lee and Chow, 2013), since a non-causal marker may have a nonzero coefficient because it is in LD with a causal locus that has not been genotyped. The phenotype need not be standardized, but it makes our presentation simpler.

The residuals will be uncorrelated with the "chip-based" breeding (additive genetic) values $\mathbf{Zu}$ if gene-environment correlation is absent or properly controlled (Yang et al, 2011b; Browning and Browning, 2011; Goddard et al, 2011; Janss et al, 2012; Speed et al, 2012) and if $\mathbf{Zu}$ and the vector of genotypic means are orthogonal. What the latter condition means is that the chip-based breeding values in $\mathbf{Zu}$ must be uncorrelated with both the discrepancy between true and chip-based breeding values and the non-additive residuals attributable to dominance and epistasis. This condition is difficult to assess, but we assume henceforth that it is met. Then from the basic equation

$$\mathbf{y} = \mathbf{Zu} + \mathbf{e}, \tag{2}$$

we see that the total phenotypic variance can be written as

$$\begin{aligned} \mathrm{Var}(Y) &= \frac{1}{n}\mathrm{E}\left(\mathbf{y}'\mathbf{y}\right) \\ &= \frac{1}{n}\mathrm{E}\left(\mathbf{u}'\mathbf{Z}'\mathbf{Z}\mathbf{u} + \mathbf{e}'\mathbf{e}\right) \\ &= \sigma^2_{A,\mathrm{SNP}} + \sigma^2_{E,\mathrm{SNP}}, \end{aligned} \tag{3}$$

where the expectation is over random residuals. The "SNP-based heritability" is thus $h^2_{\mathrm{SNP}} = \sigma^2_{A,\mathrm{SNP}}/(\sigma^2_{A,\mathrm{SNP}} + \sigma^2_{E,\mathrm{SNP}})$. If we assume that linkage equilibrium (LE) holds approximately, then $\mathbf{Z}'\mathbf{Z} \approx n\mathbf{I}_p$ and the additive genetic variance is approximately $\mathbf{u}'\mathbf{u}$.

We emphasize that $\sigma^2_{A,\mathrm{SNP}}$ is the variance that would be removed from the total phenotypic variance by multiple regression on all markers that happen to be assayed by the genotyping chip, as sample size goes to infinity. Because not all causal variants may be genotyped or represented by LD proxy, the chip-based additive genetic variance denoted here by $\sigma^2_{A,\mathrm{SNP}}$ is smaller than the true additive genetic variance σ^2_A contributed by all causal loci. Similarly, $\sigma^2_{E,\mathrm{SNP}} > \sigma^2_E$ and $h^2_{\mathrm{SNP}} < h^2 = \sigma^2_A/(\sigma^2_A+\sigma^2_E)$. Leaving aside these subtleties of definition, we can see that Eq. 3 holds because $(1/n)\mathrm{E}\left(\mathbf{u}'\mathbf{Z}'\mathbf{Z}\mathbf{u}\right)$ is the variance of chip-based breeding values and hence equal to $\sigma^2_{A,\mathrm{SNP}}$.

GREML estimates the parameters $\sigma^2_{A,\mathrm{GREML}}$ and $\sigma^2_{E,\mathrm{GREML}}$ in the model

$$\begin{aligned} \mathrm{E}\left(\mathbf{y}\mathbf{y}'\right) &= \mathrm{E}\left(\mathbf{Z}\mathbf{u}\mathbf{u}'\mathbf{Z}' + \mathbf{e}\mathbf{e}'\right) \\ &= \mathbf{A}\sigma^2_{A,\mathrm{GREML}} + \mathbf{I}_n\sigma^2_{E,\mathrm{GREML}}, \end{aligned} \tag{4}$$

where, in the notation of Yang et al (2010), $\mathbf{A} = (1/p)\mathbf{Z}\mathbf{Z}'$ is the matrix of realized relatedness coefficients. It is helpful to write out the typical element of $\mathbf{A}$,

$$\mathbf{A}_{ij} = \frac{1}{p}\sum_{k=1}^{p} z_{ik}z_{jk}, \tag{5}$$

which is very analogous to the traditional coefficients of relatedness appearing in the classical formulas for the correlations between close relatives (Crow and Kimura, 1970; Lynch and Walsh, 1998). Preserving this analogy turns out to be important because different possible standardizations of the X_{ik} will lead to different estimates of the variance components. For example, if a matrix differing from $\mathbf{A}$ by a constant factor is used in the place of $\mathbf{A}$, then the estimate of $\sigma^2_{A,\mathrm{GREML}}$ will be multiplied by that constant (Speed et al, 2012). The average of the off-diagonal realized relatedness coefficients over all pairs is zero (Powell et al, 2010), and the diagonal elements of $\mathbf{A}$ converge to unity as p becomes large.

At this point the equality of the first and second lines in Eq. 4 should be regarded not as a derivable fact but rather as *a priori* definitions of the parameters $\sigma^2_{A,\mathrm{GREML}}$ and $\sigma^2_{E,\mathrm{GREML}}$. Note that $\sigma^2_{A,\mathrm{GREML}} + \sigma^2_{E,\mathrm{GREML}} = \sigma^2_{A,\mathrm{SNP}} + \sigma^2_{E,\mathrm{SNP}} = \mathrm{Var}(Y)$.

The emergence of the matrix $\mathbf{A}$ from the action of the expectation operator in Eq. 4 implies that $\mathbf{A}$ is a constant (up to permutations of the sample) characterizing the population from which the n individuals have been drawn. The expectation must thus be interpreted as taken over random samples of size n sharing the same precise histogram of relatedness coefficients $\mathbf{A}_{ij}$ but differing in the specific entries of $\mathbf{e}$. In this way $\mathbf{A}$ is somewhat analogous to the sum of the squared differences between the independent variable and its mean (an ancillary statistic) in Fisher's (1973) discussion of univariate linear regression. Since fixing $\mathbf{Z}$ suffices to fix $\mathbf{A}$, we henceforth assume that $\mathbf{Z}$ is fixed. This interpretation does not seem problematic; across distinct samples of large size n from the same population, genotyped with the same p-variant chip, histograms of relatedness coefficients with a reasonable shared bin width should exhibit little variability.

Let us compare Eqs. 3 and 4. The expectation of $\mathbf{ee}'$ alone is $\mathbf{I}_n\sigma^2_{E,\mathrm{SNP}}$. Therefore, $(\sigma^2_{A,\mathrm{GREML}}, \sigma^2_{E,\mathrm{GREML}}) = (\sigma^2_{A,\mathrm{SNP}}, \sigma^2_{E,\mathrm{SNP}})$ implies that $\mathbf{Zuu}'\mathbf{Z}' = \mathbf{A}\sigma^2_{A,\mathrm{SNP}}$, which in turn implies that $\mathbf{Zuu}'\mathbf{Z}'$ is proportional to $\mathbf{u}'\mathbf{uZZ}'$. Such proportionality does not hold, however, as a matter of mathematical necessity. Therefore the question is posed: under what circumstances is

$$h^2_{\mathrm{GREML}} = \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{A,\mathrm{GREML}} + \sigma^2_{E,\mathrm{GREML}}} \quad \text{(the quantity estimated by GREML)}$$

approximately equal to

$$h^2_{\mathrm{SNP}} = \frac{\sigma^2_{A,\mathrm{SNP}}}{\sigma^2_{A,\mathrm{SNP}} + \sigma^2_{E,\mathrm{SNP}}} \quad \text{(what we wish to know)?}$$

It is important to note that no assumption-free adjustment of h^2_{GREML} can yield a reliable estimate of the true heritability h^2 if the genotyping chip assays a limited panel of markers. For example, it may be that rare causal variants have large phenotypic effects, and such variants may be absent from the genotyping chip and poorly represented by LD proxy. Therefore, if it turns

out that h^2_{GREML} bears no exact relationship to h^2, this should not necessarily be construed as a fault of GREML. The most that can be reasonably demanded from the method is that $h^2_{\mathrm{GREML}} \approx h^2_{\mathrm{SNP}}$, and this is the issue that we address here.

Yang et al (2010) assume that each element of $\mathbf{u}$ can be regarded as an independent draw from a normal distribution with mean zero and variance $\sigma^2_{A,\mathrm{SNP}}/p$. The desired equality between $\sigma^2_{A,\mathrm{GREML}}$ and $\sigma^2_{A,\mathrm{SNP}}$ then follows if we further suppose that the treatment of $\mathbf{u}$ as a vector of independent random variables justifies the replacement of $\mathbf{uu}'$ with $\mathbf{I}_p\sigma^2_{A,\mathrm{SNP}}$ under the action of the expectation operator. There are two aspects of this assumption, however, that seem rather nonbiological. The first is that the number of markers with nonzero regression coefficients ("nonzeros") is typically believed to be much smaller than the total number of genotyped markers (Park et al, 2011; Stahl et al, 2012), which is inconsistent with a normal distribution. Secondly and more importantly, the partial regression coefficients in $\mathbf{u}$ represent the average effects of gene substitution (or LD proxies for such effects) and thus cannot be said to vary randomly across individuals. Hence, while the spectrum of the coefficients could be described by a normal distribution or some other distribution, the exterior product $\mathbf{uu}'$ cannot be averaged over this distribution characterizing *markers* when given as an input to an expectation operator over random residuals disturbing the phenotypes of *individuals*. This implies that $\mathbf{uu}'$ is not proportional to the identity matrix.

Note the contrast between the meaning of $\mathbf{u}$ in the GREML literature and in the treatment of linear mixed models by Lynch and Walsh (1998). For instance, in the example of Lynch and Walsh's Chapter 26, the elements of $\mathbf{u}$ are the breeding values of the pedigree founders and thus can properly be said to vary randomly across different realizations of the pedigree structure.

For the reasons just given, we refrain from the Yang et al (2010) assumption and treat $\mathbf{u}$ as an arbitrary fixed constant rather than a random variable. Fixing both $\mathbf{Z}$ and $\mathbf{u}$ implies that the population targeted for inference consists of random samples sharing the same *sample* value of $\sigma^2_{A,\mathrm{SNP}}$ (Eq. 3). This limitation does not seem unduly restrictive since, for values of n often used in GREML applications (∼10,000), different samples from the same population (e.g., Northwest Europeans) will scarcely differ in their realized values of $\sigma^2_{A,\mathrm{SNP}}$.

We can now write the typical off-diagonal element of the matrix $\mathbf{Zuu}'\mathbf{Z}'$ in component form as

$$\left(\mathbf{Zuu}'\mathbf{Z}'\right)_{ij} = \left(\sum_{k=1}^{p} z_{ik}u_k\right)\left(\sum_{k=1}^{p} z_{jk}u_k\right) = \sum_{k=1}^{p} z_{ik}z_{jk}u_k^2 + \sum_{k\neq\ell} z_{ik}u_k z_{j\ell}u_\ell. \quad (6)$$

Now suppose that the constants u_k^2 were all equal to $u^2 = \sigma^2_{A,\mathrm{SNP}}/p$. Then the first sum in the expression above would become $\mathbf{A}_{ij}\sigma^2_{A,\mathrm{SNP}}$. However, since it is surely false that each marker has the same squared coefficient in the regression on standardized genotypes, we decompose each u_k^2 into the sum $u^2 + \Delta_k = \sigma^2_{A,\mathrm{SNP}}/p + \Delta_k$, where $\sum \Delta_k = 0$. If there are many genotyped

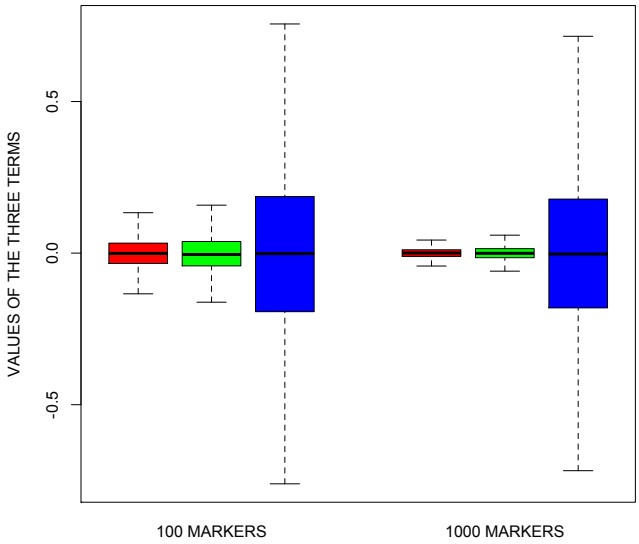

Fig. 1 Boxplots of the three terms in Eq. 7 for pairs of simulated individuals and **u**. Red corresponds to Term 1, green to Term 2, and blue to Term 3. The plots on the left display the results for 100 markers in LE, whereas those on the right display the results for 1,000 markers in LE. Values lying beyond 1.5 times the interquartile range past either the 25th or 75th percentile are omitted. Note that the average of each term is close to zero.

markers, then u^2 is a rather small quantity. If the nonzeros are a small fraction of the total, then most of the Δ_k are equal to $-u^2$. If the kth marker has a nonzero coefficient, however, then its Δ_k has a relatively large positive value.

Then we have, for the typical off-diagonal element of $\mathbf{Zuu'Z'}$,

$$\underbrace{\mathbf{A}_{ij}\sigma^2_{A,\mathrm{SNP}}}_{\text{Term 1}} + \underbrace{\sum_{k=1}^{p} z_{ik}z_{jk}\Delta_k}_{\text{Term 2}} + \underbrace{\sum_{k\neq\ell} z_{ik}u_k z_{j\ell}u_\ell}_{\text{Term 3}}. \tag{7}$$

We have already stated that the sum of Term 1 over all pairs of individuals is zero (i.e., the sum of $\mathbf{A}_{ij}$ over $i \neq j$ is zero). Since the Δ_k are constants, the sum of Term 2 over all pairs is also zero. The sum of Term 3 over all pairs is zero as well because knowledge of randomly chosen individual i's genotype at marker k cannot provide any information about randomly chosen individual j's genotype at marker ℓ, even if k and ℓ are in perfect LD; only individual j's own genotype at k can provide this information.

It might seem that Terms 2 and 3 must be extremely small compared to Term 1 for each pair of individuals i and j in order for $\sigma^2_{A,\mathrm{GREML}} \approx \sigma^2_{A,\mathrm{SNP}}$ to hold. In genetic data, however, this condition will rarely be fulfilled. To

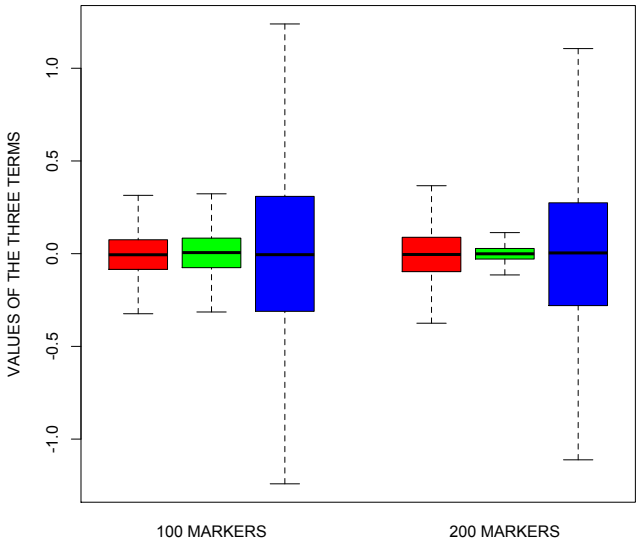

Fig. 2 Boxplots of the three terms in Eq. 7 for pairs of real genotyped individuals (but simulated **u**). Red corresponds to Term 1, green to Term 2, and blue to Term 3. The plots on the left display the results for 100 contiguous genotyped markers on chromosome 1, whereas those on the right display the results for 200 such markers elsewhere on the same chromosome. Values lying beyond 1.5 times the interquartile range past either the 25th or 75th percentile are omitted.

illustrate this fact, we simulated 1,000 individuals with genomes consisting of 100 markers in LE and MAFs drawn uniformly from (0.05, 0.50). The elements of **u** were drawn from a normal distribution and constrained to produce $\sigma^2_{A,\mathrm{SNP}} = 0.50$. The results of calculating all three terms in Eq. 7 for each pair show that Term 2 is often comparable in magnitude to Term 1 and that Term 3 is frequently larger (Fig. 1). Increasing the number of markers from 100 to 1,000 only reinforced this conclusion. It is also of some interest to calculate the three terms in Eq. 7 using real data so as to examine the impact of LD. We therefore used the Chabris et al (2013) data to run simulations using distinct sets of 100 and 200 contiguous SNPs on chromosome 1. Each real SNP typically showed moderate to strong LD with neighbors, and there were pairs of SNPs with values of r^2 exceeding 0.90. As can be seen in Fig. 2, the use of real data did not cause Terms 2 and 3 to vanish. Note that the average of each term was still close to zero.

Despite the failure of Terms 2 and 3 to vanish, GREML is often unbiased as a means of estimating $\sigma^2_{A,\mathrm{SNP}}$ when applied to real genetic data (Speed et al, 2012, 2013; Zhou et al, 2013; Browning and Browning, 2013; Lee et al, 2013), which implies that the vanishing of the additional terms is not a necessary

condition. Here we seek a more general characterization of those cases where GREML is accurate.

The variance components are estimated with GREML using REML (Lynch and Walsh, 1998; Yang et al, 2010, 2011a; Vattikuti et al, 2012). We now derive certain conditions that the maximum-likelihood (ML) estimates must satisfy. The GREML model in Eq. 4 is equivalent to treating $\mathbf{y}$ as drawn from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{V} = \mathbf{A}\sigma^2_{A,\mathrm{GREML}} + \mathbf{I}_n\sigma^2_{E,\mathrm{GREML}}$. In the absence of fixed effects, the log-likelihood is thus

$$L(\mathbf{V}\,|\,\mathbf{y}) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}|\mathbf{V}| - \frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}. \tag{8}$$

The ML estimates of the variance components are obtained by taking the partial derivatives of Eq. 8 with respect to $\sigma^2_{A,\mathrm{GREML}}$ and $\sigma^2_{E,\mathrm{GREML}}$ and setting the resulting equations to zero. Now recall that if $\mathbf{M}$ is a square matrix whose elements are multiples of a scalar x, then

$$\frac{\partial \ln|\mathbf{M}|}{\partial x} = \mathrm{Tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\right),$$
$$\frac{\partial \mathbf{M}^{-1}}{\partial x} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\mathbf{M}^{-1}.$$

Using these facts, we differentiate Eq. 8 and obtain

$$\frac{\partial L(\mathbf{V}\,|\,\mathbf{y})}{\partial \sigma^2_i} = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{V}^{-1}\mathbf{V}_i\right) + \frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}\mathbf{y}, \tag{9}$$

where

$$\frac{\partial \mathbf{V}}{\partial \sigma^2_i} \equiv \mathbf{V}_i = \begin{cases} \mathbf{I}_n & \text{if } \sigma^2_i = \sigma^2_{E,\mathrm{GREML}} \\ \mathbf{A} & \text{if } \sigma^2_i = \sigma^2_{A,\mathrm{GREML}} \end{cases}.$$

The ML conditions are thus

$$\mathrm{Tr}\left(\widehat{\mathbf{V}}^{-1}\right) = \mathbf{y}'\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{V}}^{-1}\mathbf{y},$$
$$\mathrm{Tr}\left(\widehat{\mathbf{V}}^{-1}\mathbf{A}\right) = \mathbf{y}'\widehat{\mathbf{V}}^{-1}\mathbf{A}\widehat{\mathbf{V}}^{-1}\mathbf{y}, \tag{10}$$

where $\widehat{\mathbf{V}} = \mathbf{A}\widehat{\sigma}^2_{A,\mathrm{GREML}} + \mathbf{I}_n\widehat{\sigma}^2_{E,\mathrm{GREML}}$ is the ML estimate of $\mathbf{V}$. Since ML estimates are consistent given mild regularity conditions, we set the estimate $\left(\widehat{\sigma}^2_{A,\mathrm{GREML}}, \widehat{\sigma}^2_{E,\mathrm{GREML}}\right)$ equal to $\left(\sigma^2_{A,\mathrm{GREML}}, \sigma^2_{E,\mathrm{GREML}}\right)$ for convenience.

To express the GREML variance components in terms of the observables $\mathbf{y}$ and $\mathbf{A}$ (which in turn depend on the parameters of primary interest $\sigma^2_{A,\mathrm{SNP}}$ and $\sigma^2_{E,\mathrm{SNP}}$), we need an approximation for $\mathbf{V}^{-1}$. A standard linear algebra approximation of the matrix inverse is

$$\mathbf{V}^{-1} \approx \sigma^{-2}_{E,\mathrm{GREML}}\left(\mathbf{I}_n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathbf{A}\right). \tag{11}$$

Multiplying Eq. 11 by $\mathbf{V}$, we obtain

$$\mathbf{I}_n - \frac{\sigma^4_{A,\mathrm{GREML}}}{\sigma^4_{E,\mathrm{GREML}}}\mathbf{A}^2, \tag{12}$$

and thus the vanishing of the second term will render Eq. 11 a good approximation of $\mathbf{V}^{-1}$. Given p markers in LE, the average of the squared off-diagonal elements of $\mathbf{A}$ is close to $1/p$ (which we confirmed in the simulations generating Fig. 1). The typical off-diagonal element of $\mathbf{A}^2$, $\sum_{j=1}^{n}\mathbf{A}_{ij}\mathbf{A}_{jk}$, is thus likely to be smaller than n/p. It follows that $\sigma^2_{A,\mathrm{GREML}} \ll \sigma^2_{E,\mathrm{GREML}}$ and $n \ll p$ are sufficient conditions for Eq. 11 to serve as an approximation of $\mathbf{V}^{-1}$. When Eq. 11 is substituted into Eq. 10, we obtain

$$\sigma^{-2}_{E,\mathrm{GREML}}\left[n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathrm{Tr}(\mathbf{A})\right] = \mathbf{y}'\sigma^{-4}_{E,\mathrm{GREML}}\left(\mathbf{I}_n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathbf{A}\right)\left(\mathbf{I}_n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathbf{A}\right)\mathbf{y}, \tag{13a}$$

$$\sigma^{-2}_{E,\mathrm{GREML}}\left[\mathrm{Tr}(\mathbf{A}) - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathrm{Tr}\left(\mathbf{A}^2\right)\right] = \mathbf{y}'\sigma^{-4}_{E,\mathrm{GREML}}\left(\mathbf{I}_n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathbf{A}\right)\mathbf{A}\left(\mathbf{I}_n - \frac{\sigma^2_{A,\mathrm{GREML}}}{\sigma^2_{E,\mathrm{GREML}}}\mathbf{A}\right)\mathbf{y}. \tag{13b}$$

In principle, we need to solve this system of two equations with two unknowns. However, for large n and a standardized phenotype we can immediately impose the equality $1 = \sigma^2_{A,\mathrm{GREML}} + \sigma^2_{E,\mathrm{GREML}}$, and noting that $\mathrm{Tr}(\mathbf{A}) \approx n$, we obtain the single equation

$$n\left(1-\sigma^2_{A,\mathrm{GREML}}\right)\left(1-2\sigma^2_{A,\mathrm{GREML}}\right) = \left(1-\sigma^2_{A,\mathrm{GREML}}\right)\mathbf{y}'\mathbf{y} - 2\sigma^2_{A,\mathrm{GREML}}\mathbf{y}'\mathbf{A}\mathbf{y} \tag{14}$$

to first order, which in turn implies

$$n\sigma^2_{A,\mathrm{GREML}} = n - \mathbf{y}'\mathbf{A}\mathbf{y}. \tag{15}$$

Therefore, if $\sigma^2_{A,\mathrm{GREML}} = \sigma^2_{A,\mathrm{SNP}}$, then $\mathbf{y}'\mathbf{A}\mathbf{y} = n\sigma^2_{E,\mathrm{SNP}}$, which corresponds to a necessary condition for GREML to provide a consistent estimator of $\sigma^2_{A,\mathrm{SNP}}$ in the small-(n/p), small-$\left(\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}\right)$ regime.

We now examine conditions under which $\mathbf{y}'\mathbf{A}\mathbf{y} = n\sigma^2_{E,\mathrm{SNP}}$ holds. Let us use $\mathbf{S}_{ij}$ to denote the sum of Terms 2 and 3 in Eq. 7. Then the expectation over random residuals of the quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ can be written as

$$\mathrm{E}\left(\sum_{i,j}\mathbf{A}_{ij}y_iy_j\right) = \sum_{i,j}\mathbf{A}_{ij}\left[\mathbf{A}_{ij}\sigma^2_{A,\mathrm{SNP}} + \mathbf{S}_{ij} + \mathrm{E}(e_ie_j)\right]. \tag{16}$$

The last term is indeed $n\sigma^2_{E,\mathrm{SNP}}$. The first term, $\sum \mathbf{A}^2_{ij}\sigma^2_{A,\mathrm{SNP}}$, has a diagonal contribution converging to $n\sigma^2_{A,\mathrm{SNP}}$ and an off-diagonal contribution converging to $2\binom{n}{2}(1/p)\sigma^2_{A,\mathrm{SNP}}$. Therefore the first term is approximately $(n + n^2/p)\sigma^2_{A,\mathrm{SNP}}$. The ratio of the first and last terms is approximately $(1 + n/p)(\sigma^2_{A,\mathrm{SNP}}/\sigma^2_{E,\mathrm{SNP}})$, and thus the sufficient conditions for Eq. 11 to give $\mathbf{V}^{-1}$ also ensure that the contribution of the last term to Eq. 16 dominates that of the first.

The second term in Eq. 16, $\sum_{i,j} \mathbf{A}_{ij}\mathbf{S}_{ij} = \mathrm{Tr}(\mathbf{AS})$, must also be close to zero for $\sigma^2_{A,\mathrm{GREML}} \approx \sigma^2_{A,\mathrm{SNP}}$. The diagonal contribution to this sum, $\sum_i \mathbf{A}_{ii}\mathbf{S}_{ii}$, converges to $\sum_i \mathbf{S}_{ii}$ as p becomes large. Since $\mathbf{S}_{ii}$ is the deviation of individual i's squared breeding value from $\mathbf{A}_{ii}\sigma^2_{A,\mathrm{SNP}} \approx \sigma^2_{A,\mathrm{SNP}}$, the sum of these deviations over all individuals becomes zero. The off-diagonal contribution to $\mathrm{Tr}(\mathbf{AS})$ can be interpreted as (proportional to) a covariance between relatedness $\mathbf{A}_{ij}$ and the sum of Terms 2 and 3, and this covariance must also be zero.

The sign of $\mathrm{Tr}(\mathbf{AS})$ when it deviates from zero cannot in general be used to predict the sign of $\sigma^2_{A,\mathrm{GREML}} - \sigma^2_{A,\mathrm{SNP}}$ from Eq. 15, which is derived from an uncontrolled expansion with an unknown range of validity. The following argument for the direction of the bias induced by $\mathrm{Tr}(\mathbf{AS}) \neq 0$ appears to be valid for typical values of n, p, and $\sigma^2_{A,\mathrm{SNP}}/\sigma^2_{E,\mathrm{SNP}}$. If there is a positive covariance between $\mathbf{A}_{ij}$ and $\mathbf{S}_{ij}$—meaning that pairs with above-average (below-average) relatedness also tend to have above-average (below-average) values of Terms 2 or 3—the phenotypic products $y_i y_j$ are systematically too far from zero and thus lead GREML to infer an excessive SNP-based heritability ($\sigma^2_{A,\mathrm{GREML}} > \sigma^2_{A,\mathrm{SNP}}$). Conversely, if there is a negative covariance between Term 1 and the sum of Terms 2 and 3, the shrinking of phenotypic products toward zero leads GREML to underestimate SNP-based heritability ($\sigma^2_{A,\mathrm{GREML}} < \sigma^2_{A,\mathrm{SNP}}$). There is a close analogy here to the requirement of a zero correlation between the causal variable and the residual disturbance for least-squares regression to provide an unbiased estimate of a linear causal effect.

Note again that we have made no assumption with respect to whether the partial regression coefficients in $\mathbf{u}$ follow a normal distribution or indeed any probability distribution. In fact, as we will shortly demonstrate, GREML can serve as an accurate means of estimating h^2_{SNP} in the case of a single nonzero, and obviously a probability distribution prescribing one nonzero and $p - 1$ zeros is not normal. Therefore this feature of the genetic architecture *per se* should not affect the accuracy of GREML as a method for estimating h^2_{SNP}.

We now show that our account provides quantitative explanations of recent simulation results. Both Speed et al (2012) and Zhou et al (2013) remarked upon the fact that GREML remains approximately unbiased even as the number of nonzeros becomes very small. This is perhaps surprising because the majority of the Δ_k in this case are equal to $-u^2 = -\sigma^2_{A,\mathrm{SNP}}/p$. But suppose that there are s nonzeros, where s is an arbitrary positive integer smaller than or equal to p. As long as the markers are in LE, then the contribution to

$\mathrm{Tr}(\mathbf{AS})$ from the products of relatedness and Term 2 is zero since

$$\begin{aligned}
&\sum_{i,j}\sum_{k'} z_{ik'}z_{jk'}\left(\sum_{k=1}^{s} z_{ik}z_{jk}\Delta_k - \frac{\sigma^2_{A,\mathrm{SNP}}}{p}\sum_{k=s+1}^{p} z_{ik}z_{jk}\right)\\
&= n\sum_{k=1}^{s}\Delta_k - n\frac{p-s}{p}\sigma^2_{A,\mathrm{SNP}}\\
&= n\left[\sum_{k=1}^{s}\left(u_k^2 - \frac{\sigma^2_{A,\mathrm{SNP}}}{p}\right) - \frac{p-s}{p}\sigma^2_{A,\mathrm{SNP}}\right]\\
&= 0, \qquad (17)
\end{aligned}$$

where we used the property that LE and large n imply $\sum_i z_{ik}z_{ik'} = n\delta_{k,k'}$ ($\delta_{k,k'}$ is the Kronecker delta). Note that each of the s nonzeros may have an arbitrary MAF and u_k. Furthermore, the typical term in the expansion of the covariance between relatedness and Term 3 is

$$z_{ik}z_{jk}z_{i\ell}z_{jm}u_\ell u_m, \qquad (18)$$

and LE also ensures that the average of this product vanishes. Whenever ℓ or m indexes a marker outside of the nonzeros, the product vanishes regardless, and a single nonzero thus trivially guarantees a zero covariance. Therefore, since LE guarantees that $\mathrm{Tr}(\mathbf{AS}) = 0$ in the small-(n/p), small-$\left(\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}\right)$ regime, GREML will accurately estimate the heritability captured by a set of independent markers even as the number of nonzeros decreases down to one.

It has been reported that the MAF spectrum of the nonzeros can affect the accuracy of GREML (Speed et al, 2012, 2013; Lee et al, 2013). Because the calculations related to Eqs. 17 and 18 rely on LE rather than any assumption regarding the MAF spectrum, this sensitivity must arise from LD and the tendency of higher-MAF variants to be better tagged by neighboring markers. One way in which LD affects GREML can be easily explained upon making the convenient assumption that the panel of markers is partitioned into two subsets, one of which is characterized by complete LE and the other by complete LD. Then Term 2 can be rewritten as

$$\sum_{\text{LE set}} z_{ik}z_{jk}\Delta_k + \sum_{\text{LD set}} z_{ik}z_{jk}\Delta_k = \sum_{\text{LE set}} z_{ik}z_{jk}\Delta_k + |\text{LD set}|\,\mathbf{A}_{ij,\text{LD set}}\sum_{\text{LD set}}\Delta_k,$$

where $|\cdot|$ denotes set cardinality and $\mathbf{A}_{ij,\text{LD set}}$ is the relatedness over just the markers in the LD set. The factor can be pulled out from the second sum because perfect LD implies that $z_{ik}z_{jk}$ equals the constant $|\text{LD set}|\,\mathbf{A}_{ij,\text{LD set}}$ for each k in the set. Notice that $\mathbf{A}_{ij,\text{LD set}}$ makes a disproportionate contribution to $\mathbf{A}_{ij}$. For example, if there are 10,000 markers and $|\text{LD set}| = 5{,}000$, then $\mathbf{A}_{ij}$ is a weighted sum of 5,001 contributions where the weight of the LD set is 5,000 times as large as any other. If the nonzeros (positive Δ_k) are all in the LD set, then a positive correlation may be induced between Terms 1 and 2. Conversely, if the nonzeros are all in the LE set, then a negative correlation

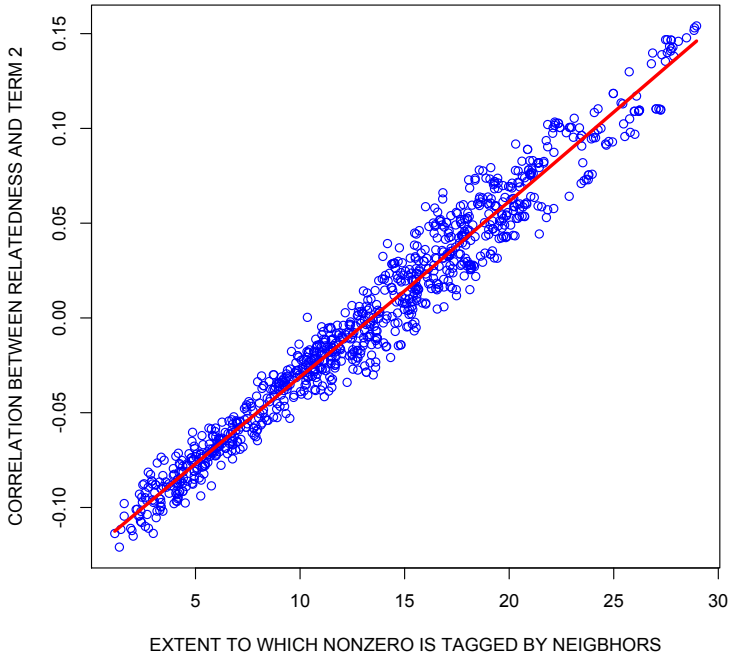

Fig. 3 The strong relationship between the extent to which the simulated nonzero SNP is tagged by neighboring markers and the resulting correlation between relatedness (Eq. 5) and Term 2 (Eq. 7). Each point represents one of 1,000 contiguous genotyped SNPs on chromosome 1 in the dataset of Chabris et al (2013). The x-axis represents the sum of the elements in the row of $\mathbf{C}$ (Eq. 1) corresponding to the nonzero SNP, and the y-axis represents the correlation between relatedness and Term 2. The LOESS curve is displayed in red.

may be induced because $\Delta_k = -u^2$ for each k in the LD set. A small number of nonzeros can therefore lead to upward (downward) bias because by chance the nonzeros may be strongly (poorly) tagged by neighboring SNPs. On the other hand, suppose that there are an equal number of nonzeros in the LE and LD sets. So long as there is no tendency for the Δ_k of the nonzeros to be larger in one of the sets, the magnification of the nonzeros in the LD set should be balanced by the diminution of those in the LE set, leading to an overall estimate of h^2_{SNP} that may be nearly unbiased.

To illustrate the impact of LD on the covariance between relatedness and Term 2, we performed simulations where each of 1,000 contiguous SNPs on chromosome 1 in the Chabris et al (2013) data was stipulated to be the single nonzero. Fig. 3 shows that there was a strong correlation, approaching unity, between the chosen nonzero's redundancy with neighboring SNPs (Eq. 1) and the resulting covariance between relatedness and Term 2. Because we used only a single nonzero in each simulation, Term 3 was trivially zero for all pairs and

therefore did not need to be computed. The results displayed in Fig. 3 clearly bear out the fact that a clustering of nonzeros among the most (least) well-tagged markers leads to a positive (negative) covariance between relatedness and the additional terms in Eq. 7.

A shortcoming of our mathematical expressions is that they make no predictions when either n/p or $\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}$ is relatively large. However, because the requirements of small n/p and $\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}$ only arise from our need to approximate $\mathbf{V}^{-1}$ for the purpose of obtaining a closed-form solution of the ML equations, it is quite plausible that our condition for the unbiasedness of GREML as a method for estimating h^2_{SNP} continues to be necessary outside of the small-(n/p), small-$\left(\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}\right)$ regime. In particular, the condition of a zero correlation between relatedness and the additional terms in Eq. 7 should continue to hold for the following intuitive reason. If we reduced the dataset such that the phenotypic products of pairs within a given bin of relatedness (plus/minus a small quantity) were averaged together, then the conditional average of the additional terms equaling zero given any relatedness would ensure that the average phenotypic product of pairs within a fixed bin of relatedness is equal to that relatedness times $\sigma^2_{A,\mathrm{SNP}}$.

We found that the small values of n, p, and s used in our simulations based on the Chabris et al (2013) dataset prevented the diagonal contribution to Tr(**AS**) from closely approaching zero, thereby rendering this dataset unsuitable for simulations extrapolating our deductions. For this reason we turned to our second dataset, where n and p are more typical of GREML applications. In this series of simulations, each of the markers in Table 1 was specified in turn to be the single nonzero. The true h^2_{SNP} was set to 0.50, and we recorded the $\widehat{h}^2_{\mathrm{GREML}}$ produced by each replicate.

The results displayed in Fig. 4 bear out our predictions.

1. It is possible for $h^2_{\mathrm{GREML}} \approx h^2_{\mathrm{SNP}}$ even if there is only one nonzero, as long as its level of LD with neighbors is typical of the SNP panel. In the case of a moderately tagged nonzero, no average $\widehat{h}^2_{\mathrm{GREML}}$ missed the true h^2_{SNP} by more than 0.014.
2. Strong (weak) tagging leads to upward (downward) bias in h^2_{GREML} as an estimate of h^2_{SNP}, and this bias increases with the nonzero's deviation from the typical level of tagging. This dose-response relationship is consistent with the increasing magnitude of the correlation between relatedness and Term 2 in Eq. 7 as the nonzero deviates from the typical level of tagging (Fig. 3). There was one anomalous result: our randomly chosen marker satisfying our criteria for low MAF and poor tagging (rs11039838) showed an average $\widehat{h}^2_{\mathrm{GREML}}$ of 0.487, not far from the true h^2_{SNP}. To determine whether this was an unusual deviation from the overall trend, we reran the simulation of this scenario, this time specifying all 125 markers satisfying our criteria for low MAF and poor tagging as nonzeros of equal coefficient magnitude (u_k). The resulting average $\widehat{h}^2_{\mathrm{GREML}}$ was 0.347 [95% CI = (0.340,

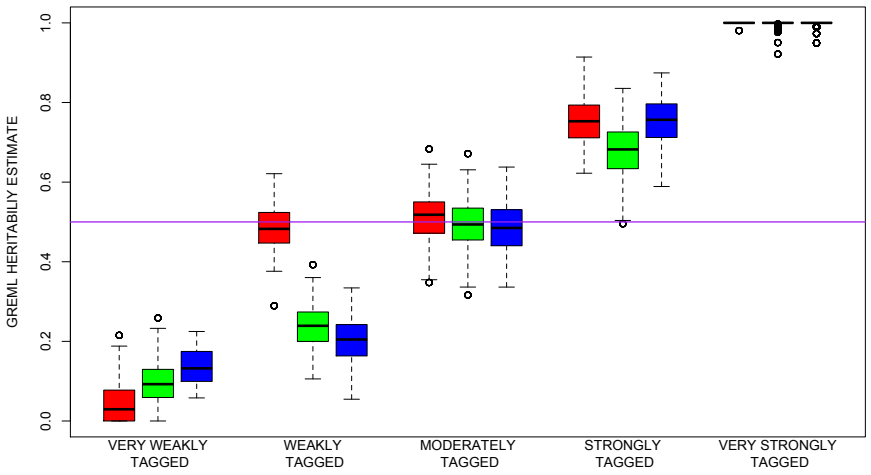

Fig. 4 Simulations of GREML performance in the case of a single nonzero. Each group of three markers was characterized by very similar tagging levels within the group. Red corresponds to markers with MAF ∼0.01, green to MAF ∼0.25, and blue to MAF ∼0.50. Each scenario was tested with 200 replicates, which led to very precise estimates of the central tendencies. The purple horizontal line corresponds to the true h^2_{SNP} of 0.50.

0.353)], further from the true h^2_{SNP} and closer to those observed at the other MAFs within our group of poorly tagged markers.

3. Once the level of tagging is controlled, the MAF of the nonzero has no discernible systematic impact on h^2_{GREML}. Even the anomalous result produced by our initial choice of a poorly tagged low-MAF marker (rs11039838) deviated in the opposite direction from the prediction of an account positing an association between low MAF of a nonzero *per se* and underestimation of h^2_{SNP}.
4. The average of the heritability estimates across all 15 markers (0.525) was close to the true h^2_{SNP}.

To confirm that $h^2_{\mathrm{GREML}} \approx h^2_{\mathrm{SNP}}$ if the nonzeros are representative of the entire genotyping chip with respect to tagging, we ran another simulation specifying 10,000 randomly chosen markers as the nonzeros. The distribution of tagging (Eq. 1) in this subset was nearly identical to the distribution among all 697,709 genotyped markers. The mean of the h^2_{GREML} estimates was 0.499 [95% CI = (0.487, 0.511)]. It is worth pointing out that we drew the magnitudes of the nonzero coefficients from a normal distribution. If the distribution is such that a few coefficients are much larger than others, than it is possible that chance unrepresentative tagging of the dominating nonzeros will lead to some bias. However, this potential problem does not seem too threatening in practice,

Table 2 Comparison of empirical and GCTA standard errors

SNP i	standard deviation of $\widehat{h}^2_{\mathrm{GREML}}$	mean of GCTA standard errors
very weakly tagged nonzeros		
rs4674229	0.0456	0.0637
rs4716447	0.0539	0.0661
rs4968679	0.0479	0.0672
weakly tagged nonzeros		
rs11039838	0.0593	0.0679
rs4912830	0.0577	0.0681
rs2654534	0.0577	0.0680
moderately tagged nonzeros		
rs7620645	0.0613	0.0691
rs12692474	0.0588	0.0684
rs4870308	0.0624	0.0691
strongly tagged nonzeros		
rs16841231	0.0596	0.0667
rs6424728	0.0657	0.0678
rs328890	0.0572	0.0660
very strongly tagged nonzeros		
rs10138824	0.0014	0.0596
rs2718306	0.0073	0.0606
rs8006587	0.0075	0.0603

because loci of large effect are relatively easy to detect and can be removed from the analysis.

Interestingly, although we have not mathematically analyzed the standard errors produced by GREML software, we found in our simulations that GCTA's standard errors are reasonably accurate (so long as h^2_{GREML} is not near either boundary) (Table 2). The fact that the GCTA standard errors are slightly larger than the corresponding empirical standard deviations of $\widehat{h}^2_{\mathrm{GREML}}$ is not necessarily a drawback of GCTA because the empirical standard deviations do not reflect the variation in realized $\sigma^2_{A,\mathrm{SNP}}$ (attributable to variation in $\mathbf{Z}$) from sample to sample. The figures of Speed et al (2012) and Zhou et al (2013) show very large standard deviations of heritability estimates in the case of few nonzeros because they varied the identities of the nonzeros across replicates. Across repeated studies of the same phenotype, where of course the identities of the nonzeros do not vary, it appears that GREML procedures produce robust standard errors after all.

Discussion

In the present work, we have deduced a necessary condition for equality between the parameter h^2_{GREML} (which is estimated by software packages such as GCTA) and h^2_{SNP} (the proportion of the phenotypic variance attributable to SNPs assayed by the given genotyping chip), in a regime allowing the inverse of the matrix $\mathbf{V}$ to be approximated by an explicit expression. In short,

the condition is that $\sum_{i,j} \mathbf{A}_{ij}\mathbf{S}_{ij} = 0$, where $\mathbf{A}_{ij}$ is the realized relatedness of individuals i and j and $\mathbf{S}_{ij}$ is the deviation of the expected phenotypic product of individuals i and j from $\mathbf{A}_{ij}\sigma^2_{A,\mathrm{SNP}}$. This condition in turn requires a zero correlation between relatedness and $\mathbf{S}_{ij}$.

It is extremely plausible that this condition continues to be necessary for $h^2_{\mathrm{GREML}} = h^2_{\mathrm{SNP}}$ outside of the small-(n/p), small-$\left(\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}\right)$ regime in which we derived it. For suppose that the condition fails, perhaps because $\mathbf{A}_{ij}$ and $\mathbf{S}_{ij}$ are positively correlated. Inspection of Eq. 7 shows that a positive correlation and consequent non-vanishing of the additional two terms causes the average phenotypic product of pairs exhibiting a given positive relatedness $\mathbf{A}_{ij}$ to exceed $\mathbf{A}_{ij}\sigma^2_{A,\mathrm{SNP}}$. This excess phenotypic similarity between positively related individuals should "trick" GREML into overestimating heritability ($h^2_{\mathrm{GREML}} > h^2_{\mathrm{SNP}}$). A positive (negative) correlation between relatedness and at least Term 2 in Eq. 7 can be induced by an overrepresentation of the nonzeros among the most (least) strongly tagged markers, and our simulations using small n/p but large $\sigma^2_{A,\mathrm{SNP}}/\sigma^2_{E,\mathrm{SNP}}$ confirmed that such overrepresentation leads to a biasing of estimates in the expected direction.

This account appears to be consistent with all simulation studies of GREML performance that have appeared thus far. Speed et al (2012, 2013) also found that the extent to which nonzeros are in strong (weak) LD with neighbors is associated with the degree to which GREML produces upwardly (downwardly) biased estimates of h^2_{SNP}. In simulations with independent markers, where of course each nonzero is as well tagged as any other marker, Zaitlen and Kraft (2012) found that GREML produces unbiased estimates of h^2_{SNP}. Speed et al (2012, 2013), Zhou et al (2013), Browning and Browning (2013), and Lee et al (2013) used real genetic data characterized by LD in their simulations, and we replicated their findings that choosing a large and random sample of markers to serve as the nonzeros leads to an absence of substantial bias. We note again that many of the simulations by ourselves and others finding that GREML is unbiased when $\sum_{i,j} \mathbf{A}_{ij}\mathbf{S}_{ij} = 0$ have not adhered to small n/p and $\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}$. For example, Zaitlen and Kraft (2012) found that GREML can be unbiased even in the case that $n > p$. The restrictive assumptions of small n/p and $\sigma^2_{A,\mathrm{GREML}}/\sigma^2_{E,\mathrm{GREML}}$ that we employed to derive the condition $\sum_{i,j} \mathbf{A}_{ij}\mathbf{S}_{ij} = 0$ thus appear to be matters of mathematical convenience only. Therefore studies that partition heritability among different parts of the genome or that analyze highly heritable phenotypes should be sound, so long as the condition $\sum_{i,j} \mathbf{A}_{ij}\mathbf{S}_{ij} = 0$ is satisfied.

We have not discussed the potential for a correlation between Terms 1 and 3 (Eq. 7). Presumably such a correlation might arise if the phenotype is subject to assortative mating, which tends to induce positive LD between causal variants (Fisher, 1918; Crow and Kimura, 1970; Lynch and Walsh, 1998). In this case individual i's genotype at marker k is a weak proxy for i's genotype at ℓ, and the fact that i and j show a positive realized similarity at k (Term 1) may also mean that they tend to show similarity at markers k and ℓ (Term 3). However, because phenotypes subject to assortative mating are probably also subject to natural selection, which tends to induce negative

LD (Lande, 1977; Bulmer, 1980), it may be that even assortative mating does not suffice to induce a correlation between Terms 1 and 3 large enough to invalidate GREML heritability estimates. This is perhaps an important issue for further research.

Here we provide a sketch of how our account generalizes to the SNP-based genetic correlation between two traits. If we use $\mathbf{v} \in \mathbb{R}^p$ to denote the vector of partial regression coefficients in the regression of the *second* phenotype on standardized genotypes, then we can write

$$\left(\sum_{k=1}^{p} z_{ik}u_k\right)\left(\sum_{k=1}^{p} z_{jk}v_k\right) = \sum_{k=1}^{p} z_{ik}z_{jk}u_kv_k + \sum_{k\neq\ell} z_{ik}u_kz_{jk}v_\ell$$
$$= \mathbf{A}_{ij}\sigma_{A,\mathrm{SNP}}(\text{trait 1}, \text{trait 2}) + \mathbf{S}_{ij},$$

where the definition of $\mathbf{S}_{ij}$ is more or less retained from the univariate case. One complication is that the part of $\mathbf{S}_{ij}$ corresponding to LD among distinct causal loci (Term 3) is defined to be part of the genetic correlation by some authors (Lynch and Walsh, 1998). If we ignore this complication and assume LE among causal loci, then the SNP-based genetic covariance will be estimated without bias by GREML if the markers that are nonzeros with respect to both traits are an effectively random sample of all genotyped markers. Furthermore, the GREML-estimated genetic correlation (the ratio of genetic covariance to the square roots of the genetic variances) may be close to unbiased as an estimate of the true genetic correlation under fairly general conditions, since biases attributable to missing causal variants and unrepresentative tagging of nonzeros may cancel from both the numerator and denominator (Trzaskowski et al, 2013). These issues may also be a worthwhile focus of future research.

Our explication shows that GREML estimates of SNP-based heritability will be reasonably accurate under much wider circumstances than those under this approach has been previously derived. Such estimates are insensitive to the number, MAF spectrum, and coefficient magnitudes *per se* of the markers with nonzero regression coefficients. The sensitivity to the LD properties of the genomic regions containing the nonzeros, however, does raise some concern. This is the crucial question: how likely is it that the nonzeros of a given phenotype are effectively like a large and random sample of all genotyped (or imputed) markers with respect to tagging?

So far we have two sources of guidance. First, it has been empirically found that the frequency spectrum of nonzeros tends to be skewed toward low MAF (Park et al, 2011). Such a skew is also plausible for evolutionary reasons. A trait-affecting mutation is likely to face a slight selection pressure that disfavors its frequency increase (Eyre-Walker, 2010), and because a causal variant can only be in strong LD with a marker if the two sites have similar MAFs (Wray et al, 2011), the spectrum of markers tagging causal variants should also be skewed toward low MAF. Since low-MAF variants tend to be less strongly tagged on the whole, we might then expect $h^2_{\mathrm{GREML}} < h^2_{\mathrm{SNP}}$ to be typical. On the other hand, because the correlation between MAF and tagging is less than perfect ($\sim .30$ in our larger dataset), it might be reasonable to expect

that such a bias will usually be mild if the number of nonzeros is large. In the simulations of Speed et al (2012, 2013) and Lee et al (2013), even a strong correlation between MAF and coefficient magnitude introduced biases of only about .05, and perhaps underestimates of h^2_{SNP} to this extent have no practical bearing on the discussion of missing heritability. Second, Speed et al (2012) have implemented an ingenious method in the LDAK package that weights markers by the extent to which they are tagged by neighbors when calculating realized relatedness. It appears from their simulations that using the resulting LD-adjusted $\mathbf{A}$ matrix to estimate h^2_{SNP} is usually successful in removing most of any bias affecting an estimate based on the unadjusted $\mathbf{A}$ matrix (Eq. 5). When they applied the LDAK method to several real phenotypes, they found a tendency for the LDAK-corrected estimates to be larger than the standard GREML estimates (supporting the notion that causal loci tend to reside at low MAF), but these increases were modest and perhaps of little practical relevance to the issue of missing heritability.

In future studies we recommend employing both the LDAK and standard GREML methods and considering their results together. Although LDAK can markedly attenuate substantial biases affecting standard GREML estimates, in some cases LDAK introduces a small bias that is otherwise absent (Speed et al, 2012, 2013; Lee et al, 2013). Perhaps surprisingly, given the mathematical nonequivalence of what we have called h^2_{SNP} and h^2_{GREML}, the GREML method is quite robust. It should remain a valuable tool in quantitative genetics and gene-trait mapping research for some time to come.

Web Resources

Genome-wide Complex Trait Analysis (GCTA), `http://www.complextraitgenomics.com/software/gcta`

Linkage-Disequilibrium Adjusted Kinships (LDAK), `http://dougspeed.com/ldak`

PLINK, `http://pngu.mgh.harvard.edu/~purcell/plink`, `https://www.cog-genomics.org/plink2`

Acknowledgements We thank Doug Speed and Xiang Zhou for answering our queries. This work was supported by the Intramural Program of the NIH, The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information.

References

Bell AE (1977) Heritability in retrospect. Journal of Heredity 68:297–300

Benjamin DJ, Cesarini D, Chabris CF, Glaeser EL, Laibson DI, Guonason V, Harris TB, Launer LJ, Purcell SM, Smith AV, Johannesson M, Magnusson PKE, Beauchamp JP, Christakis NA, Atwood CS, Hebert B, Freese J, Hauser RM, Hauser TS, Grankvist A, Hultman CM, Lichtenstein P (2012) The promises and pitfalls of genoeconomics. Annual Review of Economics 4:627–662, DOI 10.1146/annurev-economics-080511-110939

Browning SR, Browning BL (2011) Population structure can inflate SNP-based heritability estimates. American Journal of Human Genetics 89:191–193, DOI 10.1016/j.ajhg.2011.05.025

Browning SR, Browning BL (2013) Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. Human Genetics 132:129–138, DOI 10.1007/s00439-012-1230-y

Bulmer MG (1980) The Mathematical Theory of Quantitative Genetics. Oxford University Press, New York

Chabris CF, Lee JJ, Benjamin DJ, Beauchamp J, Glaeser EL, Borst G, Pinker S, Laibson D (2013) Why is it hard to find genes that are associated with social science traits? Theoretical and empirical considerations. American Journal of Public Health 103:S152–S166, DOI 10.2105/AJPH.2013.301327

Crow JF, Kimura M (1970) An Introduction to Population Genetics Theory. Harper and Row, New York

Dickson SP, Wang K, Krantz I, Hakonarson HH, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biology 8:e1000,294, DOI 10.1371/journal.pbio.1000294

Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences USA 107:1752–1756, DOI 10.1073/pnas.0906182107

Falconer DS (1960) Introduction to Quantitative Genetics, 1st edn. Oliver and Boyd, Edinburgh, UK

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399–433

Fisher RA (1941) Average excess and average effect of a gene substitution. Annals of Eugenics 11:53–63

Fisher RA (1973) Statistical Methods and Scientific Inference, 3rd edn. Hafner, New York

Gibson G (2012) Rare and common variants: Twenty arguments. Nature Reviews Genetics 13:135–145, DOI 10.1038/nrg3118

Goddard ME, Lee SH, Yang J, Wray NR, Visscher PM (2011) Response to Browning and Browning. American Journal of Human Genetics 89:193–195, DOI 10.1016/j.ajhg.2011.05.022

Hemani G, Knott S, Haley C (2013) An evolutionary perspective on epistasis and the missing heritability. PLoS Genetics 9:e1003,295, DOI 10.1371/journal.pgen.1003295

Hill WG, Mackay TFC (2004) D. S. Falconer and *Introduction to Quantitative Genetics*. Genetics 167:1529–1536

Janss L, de los Campos G, Sheehan N, Sorensen D (2012) Inferences from genomic models in stratified populations. Genetics 192:693–704, DOI 10.1534/genetics.112.141143

Lande R (1977) The influence of the mating system on the maintenance of genetic variability in polygenic characters. Genetics 86:485–496

Lee JJ, Chow CC (2013) The causal meaning of Fisher's average effect. Genetics Research 95:89–109, DOI 10.1017/S0016672313000074

Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. American Journal of Human Genetics 88:294–305, DOI 10.1016/j.ajhg.2011.02.002

Lee SH, Yang J, Chen GB, Ripke S, Stahl EA, Hultman CM, Sklar P, Visscher PM, Sullivan PF, Goddard ME, Wray NR (2013) Estimation of SNP heritability from dense genotype data. American Journal of Human Genetics 93:1151–1155, DOI 10.1016/j.ajhg.2013.10.015

Lynch M, Walsh B (1998) Genetics and the Analysis of Quantitative Traits. Sinauer, Sunderland, MA

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753, DOI 10.1038/nature08494

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Chatterjee N (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences USA 108:18,026–18,031, DOI 10.1073/pnas.1114759108

Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11:800–805, DOI 10.1038/nrg2865

Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics 91:1011–1021, DOI 10.1016/j.ajhg.2012.10.010

Speed D, Hemani G, Johnson MR, Balding DJ (2013) Response to Lee et al.: SNP-based heritability analysis with dense data. American Journal of Human Genetics 93:1155–1157, DOI 10.1016/j.ajhg.2013.10.016

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FAS, Diabetes Genetics Replication and Meta-Analysis Consortium, Myocardial Infarction Genetics Consortium, Kathiresan S, Wijmenga C, Gregersen PK, Alfredsson L, Siminovitch KA, Worthington J, de Bakker PIW, Raychaudhuri S, Plenge RM (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genetics 44:483–489, DOI 10.1038/ng.2232

Trzaskowski M, Davis OSP, DeFries JC, Yang J, Visscher PM, Plomin R (2013) DNA evidence for strong genome-wide pleiotropy of cognitive and learning abilities. Behavior Genetics 43:267–273, DOI 10.1007/s10519-013-9594-x

Vattikuti S, Guo J, Chow CC (2012) Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genetics 8:e1002,637, DOI 10.1371/journal.pgen.1002637

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era—concepts and misconceptions. Nature Reviews Genetics 9:255–266, DOI 10.1038/nrg2322

Visscher PM, Yang J, Goddard ME (2010) A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Research and Human Genetics 13:517–524, DOI 10.1375/twin.13.6.517

Wray NR, Purcell SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biology 9:e1000,579, DOI 10.1371/journal.pbio.1000579

Wright S (1921) Systems of mating. Genetics 144:111–178

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42:565–569, DOI 10.1038/ng.608

Yang J, Lee SH, Goddard ME, Visscher PM (2011a) GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics 88:76–82, DOI 10.1016/j.ajhg.2010.11.011

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 43:519–525, DOI 10.1038/ng.823

Zaitlen NA, Kraft P (2012) Heritability in the genome-wide association era. Human Genetics 131:1655–1664, DOI 10.1007/s00439-012-1199-6

Zaitlen NA, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, Price AL (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genetics 9:e1003,520, DOI 10.1371/journal.pgen.1003520

Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics 9:e1003,264, DOI 10.1371/journal.pgen.1003264

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences USA 109:1193–1198, DOI 10.1073/pnas.1119675109