European Journal of Human Genetics (2016) 24, 1810–1816

www.nature.com/ejhg

ARTICLE

On the reconciliation of missing heritability for genome-wide association studies

Guo-Bo Chen*

The definition of heritability has been unique and clear, but its estimation and estimates vary across studies. Linear mixed model (LMM) and Haseman–Elston (HE) regression analyses are commonly used for estimating heritability from genome-wide association data. This study provides an analytical resolution that can be used to reconcile the differences between LMM and HE in the estimation of heritability given the genetic architecture, which is responsible for these differences. The genetic architecture was classified into three forms via thought experiments: (i) coupling genetic architecture that the quantitative trait loci (QTLs) in the linkage disequilibrium (LD) had a positive covariance; (ii) repulsion genetic architecture that the QTLs in the LD had a negative covariance; (iii) and neutral genetic architecture that the QTLs in the LD had a covariance with a summation of zero. The neutral genetic architecture is so far most embraced, whereas the coupling and the repulsion genetic architecture have not been well investigated. For a quantitative trait under the coupling genetic architecture, HE overestimated the heritability and LMM underestimated the heritability; under the repulsion genetic architecture, HE underestimated but LMM overestimated the heritability for a quantitative trait. These two methods gave identical results under the neutral genetic architecture. A general analytical result for the statistic estimated under HE is given regardless of genetic architecture. In contrast, the performance of LMM remained elusive, such as further depended on the ratio between the sample size and the number of markers, but LMM converged to HE with increased sample size.
European Journal of Human Genetics (2016) **24,** 1810–1816; doi:10.1038/ejhg.2016.89; published online 20 July 2016

INTRODUCTION

Heritability has been defined under the context of the multiple regression of an infinitesimal sample size.[1] Recently, various methods have been developed to search for missing heritability in genome-wide association study (GWAS) data, which is typical high-dimensional data ($M>>N$; ie, the number of markers is larger than the sample size).[2] Variance component methods, such as the modified Haseman–Elston (HE) regression and the linear mixed model (LMM), estimate heritability as being higher than do single-marker association studies.[3–5] For a quantitative trait such as height, the estimated heritability was about 0.5 using variance component methods. Compared with the empirical upper bound of the heritability of height,[6,7] which was thought to be approximately 0.8, the gap of 'missing heritability' has been much narrowed using variance component methods. However, 'true heritability' has not yet been attained. The majority endeavors focused on searching for 'missing heritability' aim to fill the gaps between state-of-the-art (GWAS data; single-marker GWAS provides the lower bound, and variance component GWAS often offers a higher estimate) and old-fashioned designs (such epidemiological data, providing the upper bound of heritability). However, epidemiological and family-based studies are often criticized for potential overestimation of heritability, if not fully justified, due to the shared environment. Similarly, it is unknown whether the variance component will overestimate or underestimate the heritability for GWAS data.

Some researchers showed that LMM provided valid estimate of heritability,[8] whereas under various genetic architecture the estimated heritability from GWAS data using maximum likelihood framework may be ambiguous.[9] Recently, Golan *et al.*[5] and Chen[10] independently discovered the discrepancy between the estimates from LMM and HE for the estimates of heritability for case-control studies (Chen also found a discrepancy in quantitative traits). Of note, a method by Golan *et al.*[5] is called phenotype correlation–genotype correlation (PCGC) regression, when without adjustment for covariates PCGC resembles HE. For convenience, we call both of them HE thereby. For more detailed discussion and controversies on the estimation of heritability please refer to Table 1.

Two common issues should be noted. First, the estimation of heritability is most often treated as a statistical procedure: a parameter is estimated and assumed to be heritability as granted. Second, the effect sizes of QTLs are assumed to be from a random distribution. Estimation of heritability can be influenced by the genetic architecture, such as the genomic locations of causal variants/QTLs or the ranges of their effect sizes.[11] As heritability is a genetic architecture parameter, it is reasonable to examine how certain forms of genetic architecture, implicitly or explicitly, will influence the estimation of variance component methods. Although little is known about genetic architecture, the estimation of heritability depends on the genetic architecture.[11]

This study closely scrutinized the genetic architecture without assuming that QTLs were random along the genome, and addressed this implication in the estimation of heritability. As demonstrated in this study, the estimation of heritability depended on the genetic architecture, which can be classified into three forms, underlying a complex trait. Given the various methods proposed for estimating heritability in GWAS data, LMM, which represents a method that is built on maximum likelihood, and HE, which is built on least squares, were studied in detail; they may differ dramatically in

Evergreen Landscape and Architecture Studio, Hangzhou, China
*Correspondence: Dr G-B Chen, Evergreen Landscape and Architecture Studio, Xixi Road 562, Hangzhou 310026, China. Tel: +86 15867181509; E-mail: chenguobo@gmail.com
Received 3 November 2015; revised 26 May 2016; accepted 14 June 2016; published online 20 July 2016

Table 1 A summary of various arguments regarding searching for missing heritability

		Conclusion	
Author(s)	*Methods*	*Quantitative traits*	*Case-control*
Yang *et al.*[3]	LMM	Unbiased	NA
	HE	Unbiased	NA
Lee *et al.*[4]	LMM	NA	Unbiased
	HE	NA	NA
Speed *et al.*[13]	LMM (weights)	Unbiased	Underestimated
	HE	NA	NA
Lee and Chow[8]	LMM	Unbiased	
		NA	NA
de los Campos *et al.*[9]	LMM (a Bayesian version)	Biased	NA
	HE	NA	NA
Golan *et al.*[5]	LMM	Unbiased	Underestimate
	HE	Unbiased	Unbiased
Chen.[10]	LMM	Biased (under- or overestimate)	Biased (under- or overestimate)
	HE	Biased (under- or overestimate)	Biased (under- or overestimate)

Abbreviations: HE, modified Haseman–Elston regression; LMM, linear mixed model; NA, if not mentioned in their reports.

their estimations of heritability and reflect the genetic architecture underlying a complex trait. The following conclusions were made: (1) an increased estimate of heritability via variance component methods can be the result of overestimation under the certain forms of genetic architecture; (2) the difference between the estimated heritability in LMM and that in HE may reveal the genetic architecture underlying a complex trait.

MATERIALS AND METHODS

The linear model of a complex trait

For a quantitative trait, under the Hardy–Weinberg equilibrium the additive genetic variance (σ_A^2) is

$$\sigma_A^2 = \Sigma_{l=1}^{L} 2p_l q_l \beta_l^2 + \Sigma_{l_1=1}^{L} \Sigma_{l_2 \neq l_1} \rho_{l_1,l_2} \sqrt{2p_{l_1} q_{l_1} 2p_{l_2} q_{l_2}} \beta_{l_1} \beta_{l_2} \quad (1)$$

in which $p_l (\leq 0.5)$ is the allele frequency for the reference allele at the l^{th} QTL, $q_l = 1 - p_l$ the frequency for the alternative allele, ρ_{l_1,l_2} is the correlation measure between the l_1^{th} and the l_2^{th} QTLs, and β_{l_1} is the additive effect for the l_1^{th} locus. Equation 1 is the classic definition of additive genetic variance, referring to page 102 in Lynch and Walsh.[12] For ease of discussion: (i) the phenotype and the genotypes are standardized, and consequently, $h^2 = \sigma_A^2$; (ii) only narrow-sense heritability is discussed here; and (iii) every QTL is perfectly tagged. Due to the context, σ_A^2 and h^2 will be used interchangeably.

The decomposition of genetic architecture

The additive variance component expressed in Equation 1 can be decomposed as

$$\sigma_A^2 = \sigma_{A.w}^2 + \sigma_{A.\overline{w}}^2 = \Sigma_{l=1}^{L} \beta_l^2 + \Sigma_{l_1=1}^{L} \Sigma_{l_2 \neq l_1} \rho_{l_1,l_2} \beta_{l_1} \beta_{l_2} \quad (2)$$

$\sum_{l=1}^{L} \beta_l^2$ is the within-locus variance, denoted as $\sigma_{A.w}^2$, and $\Sigma_{l_1=1}^{L} \Sigma_{l_2 \neq l_1} 2\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2}$ is the between-locus variance/covariance, denoted as $\sigma_{A.\overline{w}}^2$. Analogously, $h^2 = h_{A.w}^2 + h_{A.\overline{w}}^2$.

The unit of $\sigma_{A.\overline{w}}^2$ is $\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2}$, a three-element product characterizing a pair of QTLs. Given the two possible signs for ρ_{l_1,l_2}, β_{l_1}, and β_{l_2}, it generates eight combinations. Thus, for the ease of discussion of this two-QTL scenario, it is presumed that the reference alleles of two QTLs have been aligned such that $\rho_{l_1,l_2} > 0$ (changing the reference alleles will not change the sign of $\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2}$, Supplementary Note I). If $\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2} > 0$, the genetic unit is called the coupling phase, where QTLs with the same effect sign are clustered together. It is equivalent to argue whether a detected QTL is actually the aggregation of a pair of small-effect QTLs with the same sign. If $\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2} < 0$, it is called the repulsion phase, where QTLs with opposite effect signs are clustered together. It is analogous to argue whether a detected QTL is actually the aggregation of two QTLs with opposite signs. If $\rho_{l_1,l_2} \beta_{l_1} \beta_{l_2} = 0$, this is the neutral phase, where QTLs are in linkage equilibrium for the two-QTL scenario.

Thus, the smallest genetic architecture in this definition has at least two QTLs, and the total between-locus variance $\sigma_{A.\overline{w}}^2 = \Sigma_{l_1=1}^{L} \Sigma_{l_2 \neq l_1}^{L} \rho_{l_1,l_2} \beta_{l_1} \beta_{l_2}$ now can be written as the aggregation of the repulsion and coupling phases along the genome, $\sigma_{A.\overline{w}}^2 = \sum \text{repulsion phase} + \sum \text{coupling phase}$. Depending on if $\sigma_{A.\overline{w}}^2$ is 0, or greater/smaller than 0, the genetic architecture is split into three forms:

(1) the coupling genetic architecture where $\sigma_{A.\overline{w}}^2 > 0$;
(2) the repulsion genetic architecture where $\sigma_{A.\overline{w}}^2 < 0$;
(3) and the neutral genetic architecture where $\sigma_{A.\overline{w}}^2 = 0$.

When the neutral genetic architecture is assumed, the heritability can be simplified as $h_R^2 = h_{A.w}^2 = \Sigma_{l=1}^{L} \beta_l^2$. For almost all recent variance component publications,[3–5,8,9,13] a random distribution of effects, leading to a neutral genetic architecture, along the genome is assumed. Therefore, it is subsequently demonstrated that the coupling/repulsion genetic architecture helps to reconcile the estimation of heritability for GWAS data.

The conventional definition of heritability under the context of multiple regression (h_R^2)

As argued above, the heritability can be split into two components under the multiple regression of infinitesimal sample size

$$h_R^2 = h_{A.w}^2 + h_{A.\overline{w}}^2 \quad (3)$$

Before the availability of GWAS data, h_R^2 is often estimated from epidemiological data via structural equation or linkage analysis.[6,7] Those estimates are often served as the upper bound in searching 'missing heritability'.

Heritability estimated under LMM (h_{LMM}^2 and $h_{\text{LMM}.e}^2$)

In LMM, the variance component of a trait is modeled as

$$\text{var}(y) = A\sigma_A^2 + I\sigma_e^2 = A\left(\sigma_{A.w}^2 + \sigma_{A.\overline{w}}^2\right) + I\sigma_e^2$$

Where A is a realized genetic relatedness matrix for samples. Between a pair of individuals i and j, $A_{ij} = \frac{1}{M} \Sigma_k^M \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k q_k}$, in which M is the number of markers and x_k counts the reference alleles for the k^{th} marker. σ_A^2 can be estimated via the restricted maximum likelihood estimator.[14,15] It should be noted, as will be shown below, that when $\sigma_{a.\overline{w}}^2 \neq 0$, LMM will give a biased estimate. As discussed by de los Campos *et al*,[9] the actual estimated statistic, which is often taken as heritability, remains a fundamental question for LMM. h_{LMM}^2 denotes the heritability estimated by LMM, and $h_{\text{LMM}}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$, in which

$\sigma_A^2 + \sigma_e^2$ is considered a proxy for σ_y^2. Alternatively, in this study, an *ad hoc* estimate of heritability was also defined as

$$h^2_{\mathrm{LMM}.e} = 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{4}$$

h^2_{LMM} and $h^2_{\mathrm{LMM}.e}$ might differ across genetic architectures. Other *ad hoc* estimates include introducing weights to A matrix as proposed by Speed *et al.*[13]

Heritability as defined under the HE regression (h^2_{HE})

Using a modified HE (or PCGC as proposed by Golan *et al.*),[5,10] the variance component can be modeled as $Y_{ij} = u + bA_{ij} + \varepsilon_{ij}$, in which μ is the mean of the model, b is the regression coefficient, $Y_{ij} = (y_i - y_j)^2$ is the squared difference between a pair of individuals, and ε_{ij} is the residual. A is the realized genetic relatedness matrix for samples, as used for LMM above. Chen[10] and Golan *et al.*[5] both adopted this framework. Golan *et al.* derive the regression coefficient under the assumption that $\sigma^2_{A.\overline{w}} = 0$; when $\sigma^2_{A.\overline{w}} \neq 0$, Golan *et al.* did not provide a solution. In Chen's work, the mathematical expectation of the regression coefficient was derived regardless of whether $\sigma^2_{A.\overline{w}}$ was zero or not.[10] After all, Golan *et al.*[5] took the estimate as heritability directly, whereas Chen found a much broad variation of the estimate and discussed the condition it was, or not was, equal to heritability (h^2_R).

Chen's original work is too long to present here; therefore, only partial results are shown.[10] Without losing generality, the regression coefficient of HE is as follows (see Equation 1 and Table 1 in Chen's original work)

$$b = -2\frac{\Sigma_{k=1}^{M}\Sigma_{l_1=1}^{L}\Sigma_{l_2=1}^{L}\rho_{kl_1}\rho_{kl_2}\beta_{l_1}\beta_{l_2}/M}{\Sigma_{k_1=1}^{M}\Sigma_{k_2=1}^{M}\rho^2_{k_1k_2}/M^2} \tag{5}$$

The denominator $\Sigma_{k_1=1}^{M}\Sigma_{k_2=1}^{M}\rho^2_{k_1k_2}/M^2 = \overline{\rho}^2_{\mathcal{M}.\mathcal{M}}$ is the averaged linkage disequilibrium (LD) between the markers. The heritability is estimated as $h^2_{\mathrm{HE}} = -\frac{b}{2}$. When there is only one marker in Equation 5, $b = -2\rho^2_{\mathcal{M}.\mathcal{Q}}h^2_{\mathcal{Q}}$, in which $\rho_{\mathcal{M}.\mathcal{Q}}$ is the correlation between the marker and the QTL; $h^2_{\mathcal{Q}}$ is the heritability of the QTL. An alternative expression for HE regression coefficient is

$$b = -2\left\{\frac{\frac{1}{M}\left\{\Sigma_k^M\left[\sigma^2_{A,k.w}\right] + \Sigma_k^M\left[\sigma^2_{A,k.\overline{w}}\right]\right\}}{\overline{\rho}^2_{\mathcal{M}.\mathcal{M}}}\right\} \tag{6}$$

which decomposes the numerator to the within-locus variance, $\sigma^2_{A,k.w}$, and the between-locus variance, $\sigma^2_{A,k.\overline{w}}$ (Supplementary Note II). Equation 6, which resembles Equation 2, indicates how $\sigma^2_{A,k.\overline{w}}$, the covariance, will influence the estimate. The implications of these two components are not trivial in the inference of genetic architecture for complex traits.

Scenario 1**:** When it is the neutral genetic architecture, $\Sigma_k^M\left[\sigma^2_{A,k.\overline{w}}\right] = 0$. Equation 5 can be simplified as

$$b = -2\Lambda h^2_{A.w} \tag{7}$$

in which $\Lambda = \frac{\overline{\rho}^2_{\mathcal{Q}.\mathcal{M}}}{\overline{\rho}^2_{\mathcal{M}.\mathcal{M}}}$. $\overline{\rho}^2_{\mathcal{Q}.\mathcal{M}}$, an unknown parameter, indicates the mean of the LD between a marker and a QTL. When there is no $h^2_{A.\overline{w}}$, $h^2_{\mathrm{HE}} = \frac{b}{-2} = \Lambda h^2_R$. Λ can be 1 when every marker is a QTL, and $h^2_{\mathrm{HE}} = h^2_R$ directly leads to an unbiased estimate of heritability.

Scenario 2**:** When either the coupling or repulsion genetic architecture is present, the between-locus component contributes to the estimate in Equation 6. $\Sigma_k^M\left[\sigma^2_{A,k.\overline{w}}\right]$ can be positive or negative depending on the underlying genetic architecture. Then there is no simple way to find the heritability (h^2_R) for the trait. As demonstrated in the simulation below, a discrepancy was observed between the respective estimates of LMM and HE.

In addition, weights can be introduced into the HE regression for A_{ij}, as proposed by Speed *et al.*[13] In general, as long as the weights follow a normal distribution, the estimated heritability will be nearly identical to that without weights (Supplementary Note III).

RESULTS

Simulation I: the genetic unit (two QTLs) of the genetic architecture

In order to demonstrate how genetic architecture affects the estimation of heritability, the smallest genetic architecture, which only has two QTLs, was considered first. We simulated 1000 unrelated individuals, and two equally frequent QTLs, which had identical effect sizes, were tagged perfectly on the genome. The heritability was $h^2_R = 2p_1q_1\beta_1^2 + 2p_2q_2\beta_2^2 + 2\rho_{1,2}\sqrt{2p_1q_1}\sqrt{2p_2q_2}\beta_1\beta_2 = h^2_{A.w} + h^2_{A.\overline{w}} = 0.5$. The LD between the pair of two consecutive single-nucleotide polymorphism markers was set to $\rho_{j,j+1} = \{-0.9, -0.8, -0.7, \ldots 0.7, 0.8, 0.9\}$. When $\rho_{j,j+1}$ was positive, it led to the coupling genetic architecture; when $\rho_{j,j+1}$ was negative, it led to the repulsion genetic. The number of genetic markers M was set to 2, 100, 200, 500, 750, 1000, 2000, and 5000, and the allele frequency was 0.5 for each marker. These two QTLs were always located on the $\left(\frac{M}{2}\right)^{\mathrm{th}}$ and $\left(\frac{M}{2}+1\right)^{\mathrm{th}}$ markers. The genetic relatedness matrix A between the individuals was calculated using all M markers. Heritability was estimated using LMM and HE, respectively. For HE, based on Equation 5, it could be predicted that $E\left(h^2_{\mathrm{HE}}\right) = \frac{1+\rho_{1,2}}{\left(1+\rho^2_{1,2}\right)}h^2_R$ (Appendix). The analytical results for h^2_{LMM} was hardly known due to the likelihood, which maximizes to unpredictable maximization if it is not specified correctly.

Figure 1 illustrates the influence of the coupling/repulsion genetic architecture on the estimation of heritability. Under either the coupling or repulsion genetic architecture, neither HE nor LMM gave unbiased estimates of heritability; unbiased estimates were only generated under the neutral genetic architecture (LD = 0), regardless of the number of QTLs and the markers. For HE, the influence of genetic architecture was predictable (Appendix), and the estimated heritability agreed well with $E(h^2_{\mathrm{HE}})$. HE underestimated the true heritability, $\hat{h}^2_{\mathrm{HE}} < 0.5$, under the repulsion genetic architecture; HE overestimated the true heritability, $\hat{h}^2_{\mathrm{HE}} > 0.5$, under the coupling genetic architecture. The whole pattern was consistent for h^2_{HE} under different M/N ratios. So, the performance of HE should be predictable for the estimation of heritability, at least in the simulated scenarios.

For LMM, the findings were more complicated. The bias of the estimate was not only due to the genetic architecture but also to the ratio of M/N. As observed, when $M/N < 0.5$, $\hat{h}^2_{\mathrm{LMM}}$ overestimated the true heritability under the repulsion genetic architecture, and underestimated the true heritability under the coupling genetic architecture. It seemed that $\hat{h}^2_{\mathrm{LMM}}$ was not influenced by the genetic architecture when $M/N = 0.5$. Nevertheless, $\hat{h}^2_{\mathrm{LMM}}$ changed its response to the genetic architecture when $M/N > 0.5$. However, the number of markers increased and the performance of h^2_{LMM} converged with $E(h^2_{\mathrm{HE}})$. No known theory can explain the performance of h^2_{LMM}. The heritability estimated by $h^2_{\mathrm{LMM}.e}$ was more precise than that of both h^2_{LMM} and $E(h^2_{\mathrm{HE}})$ when $M < 500$. When the number of markers was greater than 500, its performance also converged with h^2_{HE}.

In addition, weights were introduced to generate genetic relatedness,[13] but a nearly identical patterns for both $\hat{h}^2_{\mathrm{LMM}}$ and $\hat{h}^2_{\mathrm{HE}}$ were observed with weights as without weights (Supplementary Figure S1).

The scenarios for more QTLs were also considered, but the general pattern remained the same for HE and LMM as observed for the two-QTL scenarios (Supplementary Figure S2).

Simulation II: scenarios for case–control data

In previous works by Chen[10] and Golan *et al*,[5] it was demonstrated that HE (or PCGC) was unbiased in estimating heritability for case-control data. However, that conclusion was incomplete. When the

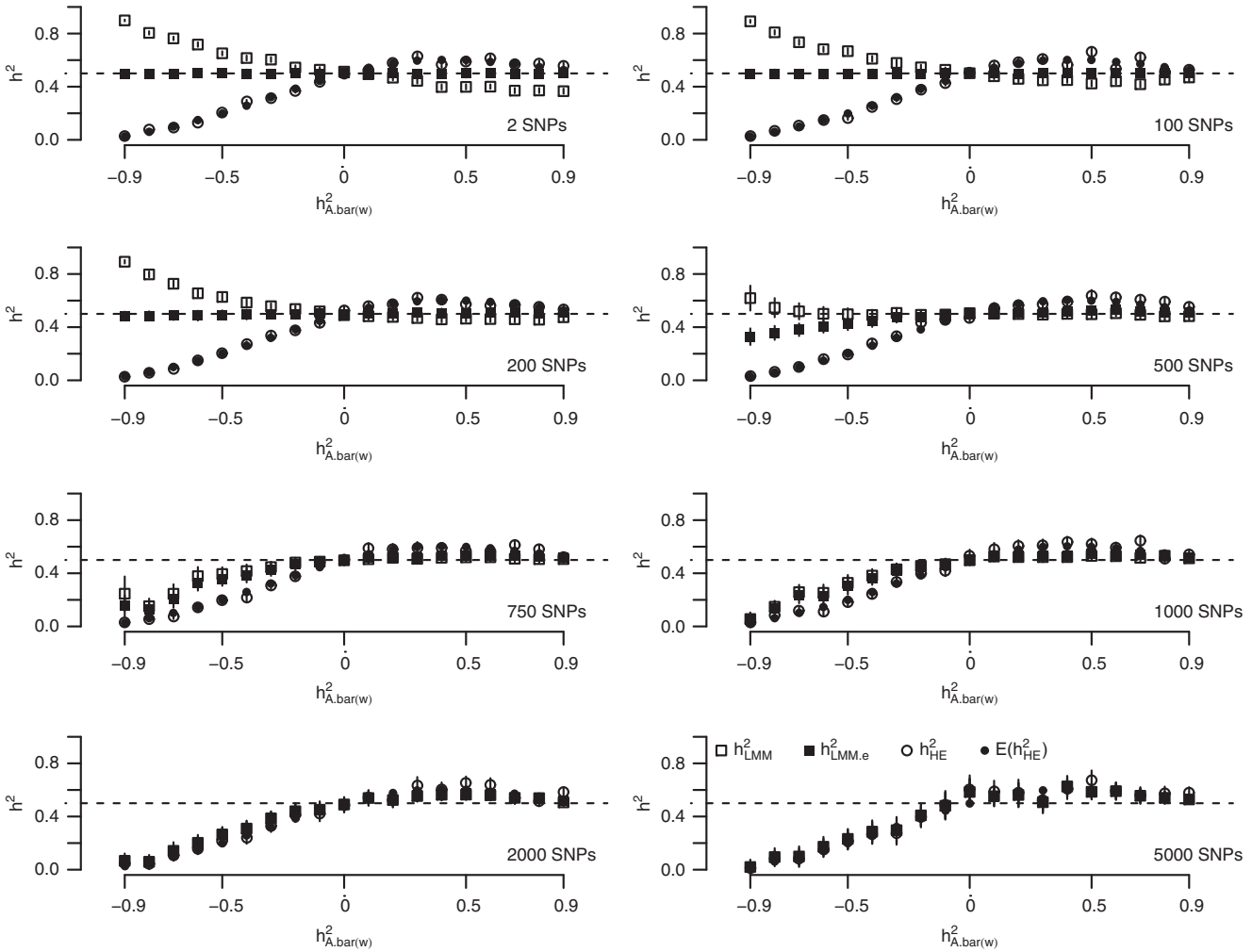

Figure 1 The influence of the genetic architecture on the estimation of heritability from the Haseman–Elston Regression (HE) and the linear mixed model (LMM) – 2 QTLs. The *x* axis indicates the genetic architecture as quantified by $h^2_{A.\overline{w}}$. $h^2_{A.\overline{w}} < 0$ refers to the repulsion genetic architecture, $h^2_{A.\overline{w}} > 0$ refers to the coupling genetic architecture; and $h^2_{A.\overline{w}} = 0$ refers to the neutral genetic architecture. $E(h^2_{A.\overline{w}})$ can be derived by Equation 5. The SD of each estimate was calculated from 100 replications of the simulations.

base population from which the cases and controls were sampled was characterized by either the coupling or repulsion genetic architecture, HE could also be biased. To demonstrate this phenomenon, 1000 cases and 1000 controls were simulated, $M = \{100, 500, 750, 1000\}$, and equally frequent biallelic QTLs were simulated. To introduce the repulsion and coupling genetic architectures, the effects of the QTLs were sampled from the standard normal distribution, and furthermore, from the first QTL to the last QTL, the effect assumed a quantity of $\Phi^{-1}(P)$. $\Phi^{-1}(P)$ generated a quantity from the normal distribution, given a *p*-value of *P*. Here $P = \frac{j}{M}$ for the j^{th} QTL. The LD between two consecutive QTLs was $\rho_{j,j+1} = \{-0.9, -0.8, -0.7, \ldots 0.7, 0.8, 0.9\}$. The total heritability on the liability scale was constrained to 0.5.

As illustrated in Figure 2, after transformation to the liability scale,[4] a pattern was observed that was similar for quantitative traits: when the base population was under the repulsion genetic architecture, $\hat{h}^2_{\mathrm{HE}}$ underestimated the heritability, and when it was under the coupling genetic architecture, $\hat{h}^2_{\mathrm{HE}}$ overestimated the heritability. When the base population was under the neutral genetic architecture, $\hat{h}^2_{\mathrm{HE}}$ produced an unbiased estimate of the heritability. In contrast, $\hat{h}^2_{\mathrm{LMM}}$ depended on both the number of markers and the genetic architecture. As observed, when $M = 100$ and $K = 0.1$, $\hat{h}^2_{\mathrm{LMM}}$ overestimated the heritability when it was under the repulsion genetic architecture, and underestimated the heritability when it was under the coupling genetic architecture. However, the pattern depended upon the number of markers: when the number of markers increased to 1000, $\hat{h}^2_{\mathrm{LMM}}$ always underestimated the heritability. $\hat{h}^2_{\mathrm{LMM}.e}$ was not as precise in this situation as it was for quantitative traits.

Of note, the crossover between $\hat{h}^2_{\mathrm{HE}}$ and $\hat{h}^2_{\mathrm{LMM}}$ did not occur at the point where there was neutral genetic architecture, but slightly under the repulsion genetic architecture. This was likely because ascertainment would introduce genetic architecture that resembled the coupling genetic architecture, which is known as Bulmer's effect in selection studies.[16]

Weights were also introduced to the genetic relatedness between individuals,[13] and the results were nearly identical to those without weighting (Supplementary Figure S3).

Summary: reconciliation of missing heritability in both theory and practice

Table 2 summarizes the theoretical and simulation results presented. The genetic architecture can be classified into coupling, repulsion, and

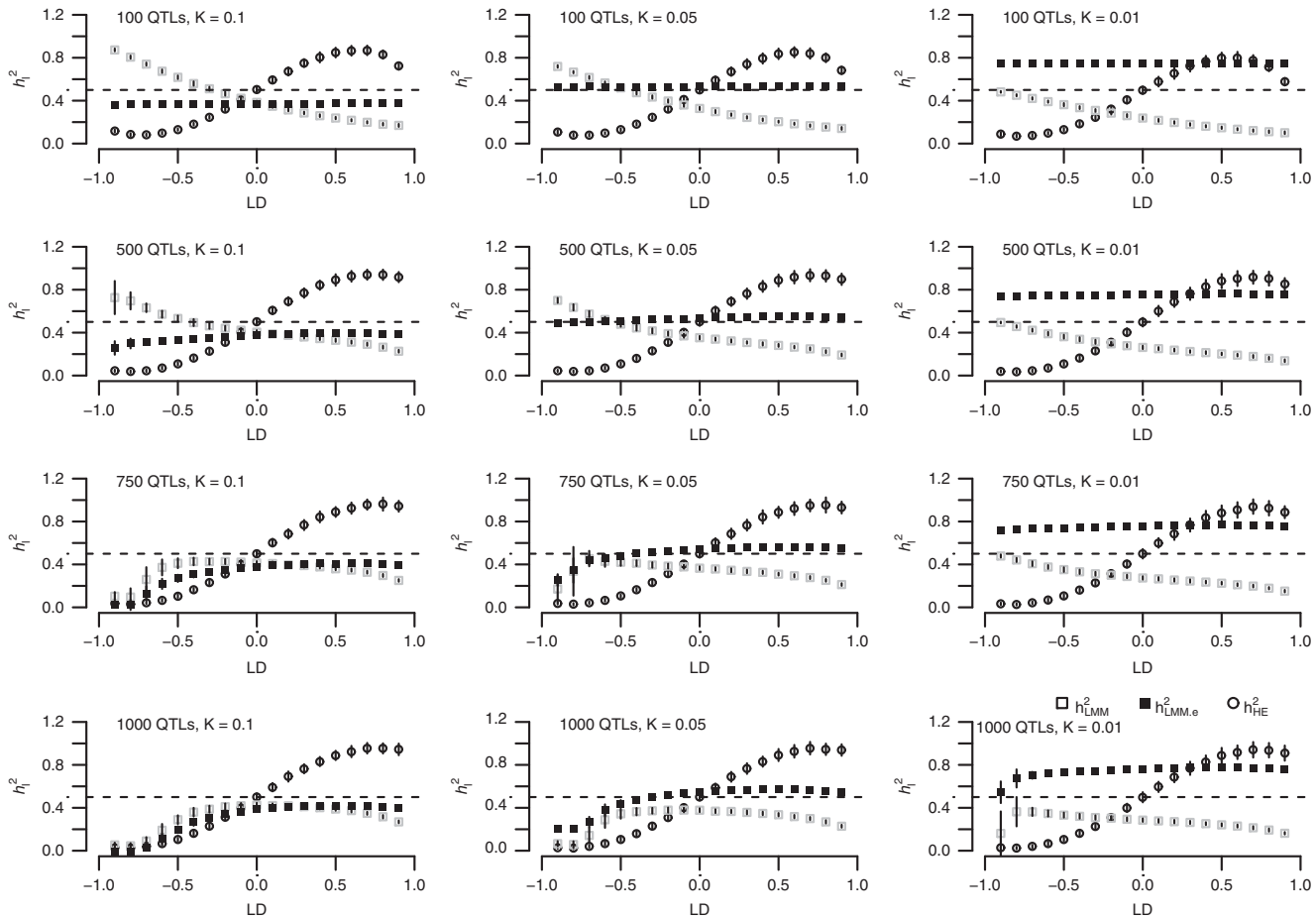

Figure 2 The simulation results for case-control data with varying prevalence (K) and number of QTLs. The x axis quantifies the genetic architecture for the base population, in which the cases and the controls are sampled. The y axis represents the heritability on the liability scale. The x axis reflects the genetic architecture: negative/positive LD indicates the repulsion/coupling genetic architecture; LD = 0 indicates the neutral genetic architecture. Of note, when $\hat{h}^2_{LMM} = \hat{h}^2_{HE}$, it was under the repulsion genetic architecture.

neutral genetic architectures. They reflect the physical features of QTLs along the genome, and heritability is a commonly used statistic to summarize this. Depending upon the genetic architecture, multiple regression (h^2_R) – the standard definition for heritability, LMM (h^2_{LMM}), and HE (h^2_{HE}) estimate heritability differently. As heritability is defined under the context of multiple regression, which leads to h^2_R, it may or may not agree with an alternative heritability estimation, such as h^2_{LMM} and h^2_{HE}.

Under the neutral genetic architecture, these three statistics may be closely related. In particular, h^2_R and h^2_{HE} are identical, but via different statistical mechanisms, as described in Equation 6. However, under the coupling or repulsion genetic architecture, h^2_{HE} may under- or overestimate h^2_R. However, it is not easy to predict the performance of h^2_{LMM}; under the neutral genetic architecture, its performance should resemble HE, and very likely converges with the performance of h^2_{HE} under a wide range of genetic architectures.

In application, the difference between h^2_{HE} and h^2_{LMM}, if observed for a trait, may reflect the underlying genetic architecture of that trait. As demonstrated in Simulation I, given the increasing ratio between M/N, h^2_{LMM} converges with h^2_{HE}. When these values are close, it does not mean that the estimated heritability is correct; however, this may reflect a condition in which one can presume that the estimated heritability was likely unbiased. It is unclear whether the convergence is also the case in real data analyses. Further investigation is required to examine how often h^2_{LMM} converges with h^2_{HE}, otherwise many reported heritability from LMM remains to be *ad hoc* because of its unwarranted outcomes.

DISCUSSION

As acknowledging the genetic architecture is important for the estimation of heritability, three possible forms of genetic architecture were introduced. Under these three forms, the performance of LMM and HE could be classified. In previous studies, it was suspected that LMM may underestimate the heritability in case-control data,[5,10] and this study showed that the bias could even occur for quantitative traits. Furthermore, under the coupling genetic architecture, HE overestimated the heritability; under the repulsion genetic architecture, HE underestimated heritability. Under the neutral genetic architecture, HE gave an unbiased estimate. LMM depended on factors other than the genetic architecture, such as the ratio between M and N. Although there was uncertainty in h^2_{LMM}, an approximation of h^2_{LMM} could be archived under the neutral genetic architecture. However, as the density of markers can fluctuate the estimation of h^2_{LMM}, an increased

Table 2 Summary for LMM and HE in the estimation of heritability for genome-wide association data

Genetic architecture[a]	*Estimator*		
	h^2_{R}	h^2_{HE}[b]	h^2_{LMM}[c]
Neutral	$h^2_{A.w}$	$\Lambda h^2_{A.w}$	$h^2_{\mathrm{LMM}} \approx h^2_{\mathrm{HE}} = \Lambda h^2_{A.w}$
Positive coupling	$h^2_{A.w} + h^2_{A.\overline{w}}$	$\frac{\sum_{k=1}^{M}\sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\rho_{kl_1}\rho_{kl_2}\beta_{l_1}\beta_{l_2}/M}{\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\rho^2_{k_1k_2}/M^2}$	$h^2_{\mathrm{LMM}} \approx h^2_{\mathrm{HE}}$
Negative coupling			

Abbreviations: HE, modified Haseman–Elston regresson; LMM, linear mixed model.
The conclusion was largely based on quantitative traits.
[a]Neutral/positive/negative coupling genetic architecture lead to $h^2_{A.\overline{w}}$ equal/greater/smaller than zero.
[b]The HE regression always has an analytical result regardless of genetic architecture. Under the neutral genetic architecture, it is easy to interpret that $\Lambda = \frac{\bar{\rho}^2_{\mathscr{Q}.\mathscr{M}}}{\bar{\rho}^2_{\mathscr{M}.\mathscr{M}}}$ (Equation 7), a value between 0 and 1, indicates how well causal variant has been tagged. Under the either positive or negative coupling genetic architecture, the interpretation of the estimate is available in a high-dimension space (Supplementary Note II).
[c]LMM does not have close-form result for the estimate. However, when the number of markers was getting larger, its performance was approaching HE.

estimate of heritability may reflect better tagging of QTLs, which may be a good thing, or an overestimation, which is not expected.

These three classes of genetic architecture can be naturally translated into a biological question: how do QTLs emerge on a regional scale and do those nearby QTLs resemble each other or not? A GWAS hit is often an aggregation of much smaller signals, such as those observed in the GIANT height study.[17] In the future, it should be possible to determine whether a region that harbors a GWAS hit actually has more than one signal. The local clustering of QTLs will lead to the repulsion or coupling genetic architecture on the whole-genome scale, but large sample size is required to observe it.

As argued by Bulmer,[16] selection could drive the departure of $\sigma^2_{A.\overline{w}}$ from zero, and consequently lead to the coupling or repulsion genetic architecture. This raises the question of how likely the repulsion or coupling genetic architecture in real data. A departure from the neutral genetic architecture is indicated when a trait's $\hat{h}^2_{\mathrm{LMM}}$ may differ from its $\hat{h}^2_{\mathrm{HE}}$. As HE has been under reported in the literature, assessing which genetic architecture is more likely among the three proposed genetic architectures is not possible now. As h^2_{HE} is easy to implement, testing for genetic architecture forms in various species should be possible,[5,10,18] particularly among beef cattle or chickens, which are often under strong directional selection and whose traits are also likely under strong selection.

This study used very simple scenarios to demonstrate the genetic architecture and its impact on estimating heritability. Other factors, such as quality control and population structure, may lead to different estimates of heritability using LMM and HE. After all, the current paradigm used to search for missing heritability favors a higher estimate of heritability; however, one should be careful because a much higher estimate may be an overestimation due to methodological limitations rather than approach the missing heritability.[19]

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

This article benefited greatly from detailed feedback by the reviewers. I thank Wouter Peyrot and Liyuan Zhou for their helpful discussion. No specific funding was received for this work.

1 Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*. 4th edn. Pearson Education Limited: Harlow, UK, 1996.
2 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
3 Yang J, Benyamin B, McEvoy BP *et al*: Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
4 Lee SH, Wray NR, Goddard ME, Visscher PM: Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88**: 294–305.
5 Golan D, Lander ES, Rosset S: Measuring missing heritability: Inferring the contribution of common variants. *Proc Natl Acad Sci USA* 2014; **111**: E5272–E5281.
6 Visscher PM, Medland SE, Ferreira MAR *et al*: Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2006; **2**: e41.
7 Chen X, Kuja-Halkola R, Rahman I *et al*: Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *Am J Hum Genet* 2015; **97**: 708–714.
8 Lee JJ, Chow CC: Conditions for the validity of SNP-based heritability estimation. *Hum Genet* 2014; **133**: 1011–1022.
9 de los Campos G, Sorensen D, Gianola D: Genomic Heritability: What Is It? *PLoS Genet* 2015; **11**: e1005048.
10 Chen G-B: Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front Genet* 2014; **5**: 107.
11 Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM: Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genet* 2015; **11**: e1004969.
12 Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc.: Sunderland, MA, USA, 1998.
13 Speed D, Hemani G, Johnson MR, Balding DJ: Improved Heritability Estimation from Genome-wide SNPs. *Am J Hum Genet* 2012; **91**: 1011–1021.
14 Patterson HD, Thompson R: Recovery of inter-block information when block sizes are unequal. *Biometrika* 1971; **58**: 545–554.
15 Gilmour AR, Thompson R, Cullis BR: Average information REML: an efficient in linear mixed models variance parameter estimation in linear mixed models. *Biometrics* 1995; **51**: 1440–1450.
16 Bulmer MG: The effect of selection on genetic variability. *Am Nat* 1971; **105**: 201–211.
17 Yang J, Ferreira T, Morris AP *et al*: Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44**: 369–375.
18 Hu Z, Yang R-C: Marker-based estimation of genetic parameters in genomics. *PLoS One* 2014; **9**: e102715.
19 Yang J, Bakshi A, Zhu Z *et al*: Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 2015; **47**: 1114–1120.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

APPENDIX

Numerical example of a two-QTL scenario

Given $p_1 = p_2 = 0.5$ for both QTLs, which had the same effect size, $\rho_{1,2} \in [-1, 1]$, and the real heritability is $h_R^2 = 2p_1q_1\beta_1^2 + 2p_2q_2\beta_2^2 + 2\rho_{1,2}\sqrt{2p_1q_1 2p_2q_2}\beta_1\beta_2$. For this case, we set $\beta_1 = \beta_2 = \sqrt{\frac{h_R^2}{(1+\rho_{1,2})}}$.

The numerator of the HE regression coefficient was calculated as $\frac{\Sigma_{k=1}^2\Sigma_{l_1=1}^2\Sigma_{l_2=1}^2\rho_{u_{l_1}}\rho_{u_{l_2}}\beta_{l_1}\beta_{l_2}}{2} = \frac{1}{2}\left\{[\beta_1\beta_2]\begin{bmatrix}1 & \rho_{1,2}\\ \rho_{1,2} & \rho_{1,2}^2\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + [\beta_1\beta_2]\begin{bmatrix}\rho_{1,2}^2 & \rho_{1,2}\\ \rho_{1,2} & 1\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}\right\} = \beta^2(1+\rho_{1,2})^2$.

The denominator was calculated as $\overline{\rho}^2_{\mathcal{M}.\mathcal{M}} = \frac{\Sigma_{k_1=1}^2\Sigma_{k_2=1}^2\rho_{k_1k_2}^2}{4} = \frac{1}{4}[1,1]\begin{bmatrix}1 & \rho_{1,2}^2\\ \rho_{1,2}^2 & 1\end{bmatrix}[1,1]^T = \frac{1}{2}(1+\rho_{1,2}^2)$.

Therefore, $b = -\frac{(1+\rho_{1,2})}{\frac{1}{2}(1+\rho_{1,2}^2)}h_R^2 = -2\frac{1+\rho_{1,2}}{1+\rho_{1,2}^2}h_R^2$. If taking the heritability as the negative half of the regression coefficient, consequently, $E(h_{\mathrm{HE}}^2) = \frac{b}{-2} = \frac{1+\rho_{1,2}}{(1+\rho_{1,2}^2)}h_R^2$, which is represented in Figure 1. It means that the statistic estimated from HE may or may not be the heritability. When $\rho_{1,2} = 0$, indicating neutral genetic architecture, HE can provide unbiased estimate of heritability.