Cuyabano *et al. Genet Sel Evol (2018) 50:41*
https://doi.org/10.1186/s12711-018-0411-0

GSE Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Understanding the potential bias of variance components estimators when using genomic models

Beatriz C. D. Cuyabano*, A. Christian Sørensen and Peter Sørensen

Abstract

Background: Genomic models that link phenotypes to dense genotype information are increasingly being used for infering variance parameters in genetics studies. The variance parameters of these models can be inferred using restricted maximum likelihood, which produces consistent, asymptotically normal estimates of variance components under the true model. These properties are not guaranteed to hold when the covariance structure of the data specified by the genomic model differs substantially from the covariance structure specified by the true model, and in this case, the likelihood of the model is said to be misspecified. If the covariance structure specified by the genomic model provides a poor description of that specified by the true model, the likelihood misspecification may lead to incorrect inferences.

Results: This work provides a theoretical analysis of the genomic models based on splitting the misspecified likelihood equations into components, which isolate those that contribute to incorrect inferences, providing an informative measure, defined as κ, to compare the covariance structure of the data specified by the genomic and the true models. This comparison of the covariance structures allows us to determine whether or not bias in the variance components estimates is expected to occur.

Conclusions: The theory presented can be used to provide an explanation for the success of a number of recently reported approaches that are suggested to remove sources of bias of heritability estimates. Furthermore, however complex is the quantification of this bias, we can determine that, in genomic models that consider a single genomic component to estimate heritability (assuming SNP effects are all *i.i.d.*), the bias of the estimator tends to be downward, when it exists.

Background

Genomic models that incorporate dense genotype information are increasingly being used and studied to infer variance parameters [1–4]. We define a genomic model as any linear mixed model (LMM) that links a phenotype to multiple genotypes without knowledge of those that are associated with the phenotype. We refer to a general set of genotypes as single nucleotide polymorphisms (SNPs) and to the set of genotypes associated with the phenotype as quantitative trait loci (QTL). The variance parameters of the LMM can be inferred using restricted maximum likelihood (REML) [5], which produces consistent, asymptotically normal estimators of variance components, even if normality does not hold and the number of QTL increases dramatically, tending to infinity [6]. These asymptotic properties of the REML estimators are not guaranteed to hold when the likelihood of the genomic model used for inference differs substantially from the likelihood of the true model that conceptually generated the data. In such a situation, the likelihood is said to be misspecified. In a Gaussian setup, given the fixed effects, this will be the case when the covariance structures of the data specified by the genomic and the true models differ.

*Correspondence: bia.cdc@gmail.com
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Blichers Allé 20, Postboks 50, 8830 Tjele, Denmark

The correct covariance structure (referred to in our work as $\mathbf{G}_Q$) requires knowledge of the QTL. Since these are typically unknown, in practice, the genomic model makes use of the available SNP genotypes instead in order to compute a covariance structure (referred to in our work as $\mathbf{G}$), leading to misspecification of the likelihood. The patterns of realized relationships at different sets of loci vary across the genome [7]. Because of this, $\mathbf{G}$ may provide a poor description of $\mathbf{G}_Q$, and the likelihood misspecification may lead to biased estimators of variance parameters.

REML was first implemented with a genomic model in [1], where the focus of inference was the proportion of the variance of a quantitative trait explained by the LMM, including all genotyped SNPs simultaneously. In more recent years, concerns have been raised about the quality of the inferred variance parameters when genomic models are used without directly addressing the problem of likelihood misspecification. Speed et al. [4] argued that uneven linkage disequilibrium (LD) between SNPs can lead to upward or downward bias of variance parameters estimators. The consequences of using $\mathbf{G}$ instead of $\mathbf{G}_Q$ on the likelihood were also investigated by [8]. These authors used the singular value decomposition of $\mathbf{G}$ and expressed the likelihood function of the genomic model as a function of these decomposition, concluding that the likelihood-based estimators are unreliable because they are sensitive to small perturbations on the eigen-values. This work generated back-and-forth discussions [9, 10].

The problem of misspecification of the likelihood of the genomic model was first raised by [11] and was studied using simulation by [12]. However, in [12] the authors addressed the problem by redefining the variance parameters according to the genomic models. In our work, we compare the variance parameters estimators to the parameters defined by the true model, as previously studied by [13]. The assumptions posed by [13] on the true model, however, differ substantially from those posed by our study, which can lead to different conclusions. Jiang et al. [13] assumed that the number of QTL associated with a phenotype are large enough to be considered infinite, an assumption which we consider rather unrealistic, and therefore our study assumes that the number of QTL is finite, although possibly very large.

This work provides a theoretical analysis based on splitting the likelihood equations into components, isolating those that contribute to incorrect inferences. We describe a true model that associates a phenotype with QTL, and we use its likelihood as a basis for comparison with the likelihood of the genomic models. This theory was used to understand the potential bias of REML estimators of variance components under different scenarios, each with different assumptions on the true model, that are of interest in quantitative genetics studies.

Methods

True and genomic models

We start this section by describing a general model that links phenotypes of a complex trait to genotype data:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{W}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where μ is the overall mean, $\mathbf{W}$ is a $n \times s$ standardized SNP genotypes matrix (with $W_{ij} = (Z_{ij} - 2\theta_j)/\sqrt{2\theta_j(1-\theta_j)}$, $\mathbb{E}(W_{ij}) = 0$, and $\mathbb{V}\mathrm{ar}(W_{ij}) = 1$; $Z_{ij} \in \{0, 1, 2\}$ is the count of the minor allele at the j-th SNP with minor allele frequency (MAF) θ_j, of the i-th individual, for all $i = 1, \ldots, n$ and $j = 1, \ldots, s$), $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}_s\sigma^2_b)$ is an $s \times 1$ vector of random SNP effects and $\boldsymbol{\varepsilon} \sim N(\mathbf{I}_n\sigma^2_{\varepsilon_Q})$ is an $n \times 1$ vector of the model's residuals. In the true model, $\mathbf{W} \doteq \mathbf{W}_Q$ is a $n \times q$ ($q \leq s$) matrix containing only the QTL, and all the elements and parameters that describe the model are sub-indexed with Q. In the genomic model, $\mathbf{W}$ contains $s = m + q$ SNPs, with $m = 0, \ldots, s$ the number of markers (non causative mutations). When $m = 0$, we are in fact in the case of the true model, and when $m = s$ the genomic model contains no QTL in the SNP data.

Variance components and REML estimation

The covariance structure specified by the true and genomic models are $\mathbf{G}_Q = q^{-1}\mathbf{W}_Q\mathbf{W}'_Q$ and $\mathbf{G} = s^{-1}\mathbf{W}\mathbf{W}'$, respectively, which also define the relationships between individuals at the genotype level [14]. Let $\sigma^2_T = s\sigma^2_b$ (total variance due to the genotypes), and $\gamma = \sigma^2_T/\sigma^2_\varepsilon$, and define the matrix $\mathbf{V}(\gamma) \doteq \mathbf{V} = \gamma\mathbf{G} + \mathbf{I}_n$, we then have that $\mathbb{V}\mathrm{ar}(\mathbf{y} \mid \mathbf{W}) = \mathbb{V}\mathrm{ar}(\mathbf{W}\mathbf{b}) + \mathbb{V}\mathrm{ar}(\boldsymbol{\varepsilon}) = \sigma^2_b\mathbf{W}\mathbf{W}' + \sigma^2_\varepsilon\mathbf{I}_n = \sigma^2_T\mathbf{G} + \sigma^2_\varepsilon\mathbf{I}_n = \sigma^2_\varepsilon\mathbf{V}$.

In genetics studies, the interest lies in estimating the narrow-sense heritability, *i. e.* the proportion of phenotypic variance explained by the genotypes. Under the true model, the heritability is defined as $h^2 = \sigma^2_{T_Q}/(\sigma^2_{T_Q} + \sigma^2_{\varepsilon_Q}) = \gamma_Q/(\gamma_Q + 1)$. Analogously, under the genomic model we have $h^2_{gen} = \gamma/(\gamma + 1)$. When we fit the true model (QTL only), $\lim_{n\to\infty}\hat{\gamma}_Q = \gamma_Q$ and consequently $\lim_{n\to\infty}\hat{h}^2 = \lim_{n\to\infty}\hat{\gamma}_Q/(\hat{\gamma}_Q + 1) = h^2$, for any number of QTL [5], even if normality does not hold, as demonstrated by [6] using large-sample theory. We also used large-sample theory to evaluate the likelihood of the genomic model, which is misspecified to the likelihood of the model that conceptually generated the phenotypes. Intuitively, if when we fit the genomic model, we obtain $\lim_{n\to\infty}\hat{\gamma} = \gamma_Q$, then $\lim_{n\to\infty}\hat{h}^2_{gen} = \lim_{n\to\infty}\hat{\gamma}/(\hat{\gamma} + 1) = h^2$. This means that if $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = \gamma_Q$, then $\lim_{n\to\infty}\mathbb{E}(\hat{h}^2_{gen}) = h^2$.

Therefore, because REML will yield the estimator $\hat{\gamma}$, we focus our analysis on $\mathbb{E}(\hat{\gamma})$.

Define $\mathbf{P}(\gamma) \doteq \mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{1}_n\left(\mathbf{1}_n'\mathbf{V}^{-1}\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\mathbf{V}^{-1}$. The REML log-likelihood of the genomic model is [15]:

$$\ell\left(\gamma, \sigma_\varepsilon^2 \mid \mathbf{W}\right) \propto -\log\left(\sigma_\varepsilon^{2n} \mid \mathbf{V} \mid\right) - \log\left(\sigma_\varepsilon^{-2} \mid \mathbf{1}_n'\mathbf{V}^{-1}\mathbf{1}_n \mid\right) - \sigma_\varepsilon^{-2}\mathbf{y}'\mathbf{P}\mathbf{y}, \quad (2)$$

and by equating its gradient to zero at the point of maximum, $\hat{\gamma}$ is the root of the REML equation [16]:

$$\frac{\mathbf{y}'\mathbf{P}\mathbf{G}\mathbf{P}\mathbf{y}}{\mathrm{tr}(\mathbf{P}\mathbf{G})} - \frac{\mathbf{y}'\mathbf{P}^2\mathbf{y}}{\mathrm{tr}(\mathbf{P})} = 0. \quad (3)$$

Using the eigen-decomposition $\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$, the REML Eq. (3) can be written as a function of γ (see "Appendix 1"):

$$g(\gamma) = \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{(\mathbf{y}'\mathbf{U}_i)^2}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i - \lambda_j}{1+\gamma\lambda_j}\right) + n\bar{y}^2\sum_{j=1}^{n-1}\frac{\lambda_j}{1+\gamma\lambda_j}, \quad (4)$$

such that $\hat{\gamma}$ is the root of $g(\gamma)$. Now, since $(\mathbf{y}'\mathbf{U}_i)^2 \longrightarrow \mu^2(\mathbf{1}_n'\mathbf{U}_i)^2 + (\mathbf{b}_Q'\mathbf{W}_Q'\mathbf{U}_i)^2$ for n sufficiently large (see "Appendix 2"), with $(\mathbf{b}_Q'\mathbf{W}_Q'\mathbf{U}_i)^2 \propto \sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}$, we can write the non-observable REML function at the root as the following (see "Appendix 3"):

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i - \lambda_j}{1+\hat{\gamma}\lambda_j}\right) = 0. \quad (5)$$

We refer to Eq. (5) as non-observable because it is written as a function of $\mathbf{U}_Q$ and $\boldsymbol{\Lambda}_Q$, which cannot be observed directly when only phenotype and genomic data are available, and we have no knowledge about the QTL. The use of such function is purely theoretical as an aid to obtaining deeper understanding of REML mechanisms. In practical implementations, we find the root of Eq. (4) to obtain $\hat{\gamma}$.

Genomic models for scenarios of interest in quantitative genetics

We evaluated genomic models for SNP data consisting of QTL and markers that can be either uncorrelated or correlated, considering two configurations: (i) QTL plus markers and (ii) markers only. The covariance structure specified by these configurations, as well as their eigen-decomposition are denoted respectively by $\mathbf{G}_{QM} = \mathbf{U}_{QM}\boldsymbol{\Lambda}_{QM}\mathbf{U}_{QM}'$ and $\mathbf{G}_M = \mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}_M'$.

Before we proceed to define our scenarios of interest and evaluate κ_i for each of them, we provide a brief discussion about our assumptions for the true model. In the study by Jiang et al. [13], the authors assumed that both the number of QTL (q) and the number of markers (m) increase simultaneously with increasing SNP data density ($q, m \to \infty$). Define $\mathbf{A}$ as the matrix of expected relationships between individuals, such that $\mathbb{E}(\mathbf{G}_Q) = \mathbb{E}(\mathbf{G}) = \mathbf{A}$ [17]. Then $\lim_{q,m\to\infty}\mathbf{G} = \lim_{q\to\infty}\mathbf{G}_Q = \mathbf{A}$, and using the eigen-decompositions $\mathbf{A} = \mathbf{U}_A\boldsymbol{\Lambda}_A\mathbf{U}_A'$, $\mathbf{G}_Q = \mathbf{U}_Q\boldsymbol{\Lambda}_Q\mathbf{U}_Q'$ and $\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$, we can state that $\lim_{q,m\to\infty}\mathbf{U} = \lim_{q\to\infty}\mathbf{U}_Q = \mathbf{U}_A$ and $\lim_{q,m\to\infty}\boldsymbol{\Lambda} = \lim_{q\to\infty}\boldsymbol{\Lambda}_Q = \boldsymbol{\Lambda}_A$. Therefore, $\lim_{q,m\to\infty} g(\gamma) = \lim_{q\to\infty} g_Q(\gamma) = g_A(\gamma) = \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}(\mathbf{y}'\mathbf{U}_{Ai})^2(\lambda_{Ai} - \lambda_{Aj})/[(1+\gamma\lambda_{Ai})^2(1+\gamma\lambda_{Aj})] + n\bar{y}^2\sum_{j=1}^{n-1}\lambda_{Aj}/(1+\gamma\lambda_{Aj})$, meaning that the REML functions of the true and genomic models become equal with increasing SNP data density. Because $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}_Q) = \gamma_Q$ [6], if $\lim_{q,m\to\infty} g(\gamma) = \lim_{q\to\infty} g_Q(\gamma)$, it is straightforward that $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = \gamma_Q$.

We consider that the assumption that both q and m increase simultaneously with increasing SNP data density is too strong. Unless we consider the true model to be the infinitesimal model (in which a phenotype is generated by a countable infinite number of QTL, each with very small effect), it is most likely that the number of QTL will be finite (it may still be large, but finite). Thus we assumed a fixed and finite number of QTL for the true model, and therefore $\mathbf{G}_Q \neq \mathbf{A}$. Our main objective was to evaluate how much of the variability in the phenotypes can be captured by $\mathbf{G}$, potentially leading to $\hat{\gamma}$ being biased to γ_Q.

We used simulations to support the theoretical analysis of the REML estimators of $h^2 = \gamma_Q/(\gamma_Q + 1)$ when using genomic models. A preliminary study indicated that 2000 individuals were enough to ensure the asymptotic properties of REML under the true model. The simulations were performed for eight scenarios that differed in population structure (completely unrelated or related individuals) and genetic architecture, in the linkage disequilibrium between QTL and markers, and the MAF of the QTL. We assumed independence between MAF and effect sizes when simulating QTL effects, and phenotypes were simulated using scaled genotypes, with a heritability parameter $h^2 = 0.05, 0.15, \ldots, 0.95$. 20,000 SNPs were simulated and for each scenario we estimated the heritability for 500 replicates that assigned 100 SNPs as QTL, and for 500 replicates that assigned 3000 SNPs as QTL. The algorithms used for the simulations can be found in appendices G to J. For each scenario, the heritability was estimated using the true model (QTL only), and two genomic models, one containing the QTL plus the markers, and one containing the markers only.

Results

Conditions for unbiased or biased REML estimators

The key to evaluating the bias of $\hat{\gamma}$ to the true parameter γ_Q is the term $\sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}$ for every $i = 1, \ldots, n$, in Eq. (5). This term corresponds to the diagonal of $\mathbf{U}'\mathbf{U}_Q\mathbf{\Lambda}_Q\mathbf{U}_Q'\mathbf{U}$. The off-diagonals of this matrix do not feature in the likelihood of the genomic models, as can be seen in Eq. (5), and thus, are not relevant to the bias of $\hat{\gamma}$. Nonetheless, we performed a brief analysis of the off-diagonal elements in our simulations, and verified that their values are centered at zero and within the interval $(-0.15, 0.15)$, regardless of the scenario. To simplify notation, we define:

$$\sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk} = \kappa_i, \quad \forall \quad i = 1, \ldots, n. \tag{6}$$

Here, we set out the conditions for asymptotically unbiased or biased $\hat{\gamma}$, and in the following subsections we give details on how we arrived at these conditions:

(1) $\kappa_i = \lambda_i, \quad \forall\, i = 1, \ldots, n \implies \lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = \gamma_Q$ (unbiased)
(2) $\kappa_i = c, \quad \forall\, i = 1, \ldots, n \implies \lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = 0$ (γ_Q cannot be estimated)
(3) $\kappa_i < \lambda_i$, for most $i = 1, \ldots, n \implies \lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) < \gamma_Q$ (biased: downwards)
(4) $\kappa_i > \lambda_i$, for most $i = 1, \ldots, n \implies \lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) > \gamma_Q$ (biased: upwards)

Note that h^2 and h^2_{gen} are monotone increasing to γ_Q and γ respectively, meaning that if we have $0 < \gamma_- < \gamma_Q < \gamma_+$, such that $h^2_- = \gamma_-/(\gamma_- + 1)$ and $h^2_- = \gamma_+/(\gamma_+ + 1)$, then $h^2_- < h^2 < h^2_+$. Thus, the direction of the bias of $\hat{h}^2_{gen}$ for h^2 is the same as that of the bias of $\hat{\gamma}$ for γ_Q.

Unbiased estimators

We know that $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}_Q) = \gamma_Q$ [6]. This holds for any number q of QTL, regardless of their MAF and correlation between them. Moreover, in the non-observable REML equation, $\kappa_{Qi} = \lambda_{Qi}$. This means that for any set of eigen-values from any $\mathbf{G}$, for n sufficiently large, $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = \gamma_Q$ if, and only if, the structure of the non-observable REML equation is as:

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\lambda_i}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right) = 0, \tag{7}$$

meaning that $\hat{\gamma}$ is unbiased for γ_Q if, and only if, $\kappa_i = \lambda_i$, for all $i = 1, \ldots, n$.

Biased estimators

There are two cases in which $\hat{\gamma}$ is biased to γ_Q: (1) $\kappa_i = c$ for $i = 1, \ldots, n-1$, such that c is a positive constant; (2) $\kappa_i = a_i \neq \lambda_i$ for $i = 1, \ldots, n-1$, such that $a_i > 0$. Note that because the $\mathbf{G}$ matrix is built with centered and scaled genotypes, its eigen-decomposition has $n-1$ degrees of freedom, and $\kappa_n = \lambda_n = 0$ always.

In the first case that $\hat{\gamma}$ is biased to γ_Q, where $\kappa_i = c$ for $i = 1, \ldots, n-1$, we have:

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{c}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right) = 0. \tag{8}$$

Only $\hat{\gamma} = 0$ guarantees the identity in Eq. (8). Therefore, when $\kappa_i = c$ for $i = 1, \ldots, n-1$, no variance from the genomic data can be captured by REML.

We now analyze the second case where $\kappa_i = a_i$ for $i = 1, \ldots, n-1$. If the relationship between a_i and λ_i is linear, *i.e.* $a_i = b\lambda_i$, then Eq. (7) ensures $\lim_{n\to\infty}\mathbb{E}(\hat{\gamma}) = \gamma_Q$. However, because $\sum_{i=1}^{n}\kappa_i = \sum_{i=1}^{n}\lambda_i = n-1$, a_i and λ_i cannot be linearly related, and $\hat{\gamma}$ will be biased to γ_Q. We have now:

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{a_i}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right) = 0. \tag{9}$$

To understand the bias in this case, we will go through some details about $a_i > 0$. Let $a_i = \lambda_i + b_i$, with $b_i \geq -\lambda_i$ and $\sum_{i=1}^{n} b_i = 0$ (because $\sum_{i=1}^{n}\kappa_i = \sum_{i=1}^{n}\lambda_i$). Thus, $\hat{\gamma}$ satisfies

$$\begin{aligned}\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\lambda_i}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right)\\ +\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{b_i}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right) = 0.\end{aligned} \tag{10}$$

From the unbiased case, we know that $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\lambda_i(\lambda_i-\lambda_j)/[(1+\hat{\gamma}\lambda_i)^2(1+\hat{\gamma}\lambda_j)]\,|_{\hat{\gamma}=\gamma_Q} = 0$. This means that if $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} b_i(\lambda_i-\lambda_j)/[(1+\hat{\gamma}\lambda_i)^2(1+\hat{\gamma}\lambda_j)] < 0$, (10) will hold only if $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\lambda_i(\lambda_i-\lambda_j)/[(1+\hat{\gamma}\lambda_i)^2(1+\hat{\gamma}\lambda_j)] > 0$. Because the latter is monotone decreasing on $\hat{\gamma}$, only an estimator $\hat{\gamma} < \gamma_Q$ will result in $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\lambda_i}{(1+\hat{\gamma}\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\hat{\gamma}\lambda_j}\right) > 0$. Now, if $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} b_i(\lambda_i-\lambda_j)/[(1+\hat{\gamma}\lambda_i)^2(1+\hat{\gamma}\lambda_j)] > 0$, then we have analogously that $\hat{\gamma} > \gamma_Q$.

Genomic models for scenarios of interest in quantitative genetics

QTL uncorrelated to markers

In "Appendix 5", we show that for genomic models that include the QTL plus markers, in which markers are uncorrelated to the QTL,

$$\kappa_{QMi} = \lambda_{QMi} + \frac{m}{q}\left(\lambda_{QMi} - 1 - \delta_i\right), \quad \forall \quad i = 1, \ldots, n-1, \text{ with } \mathbb{E}(\delta_i) = 0. \tag{11}$$

Assuming that the number of SNPs is always much larger than the number of individuals ($q + m >>> n$), $\lambda_{QM1} > \cdots > \lambda_{QM,n-1} > \lambda_{QMn} = 0$. Since $\sum_{i=1}^{n} \lambda_{QMi} = n - 1$ and because $\lambda_{QM1}, \ldots, \lambda_{QMn}$ follow the Marčenko-Pastur distribution when n is sufficiently large [18], we have:

$$\left(1 - \sqrt{\frac{n}{q+m}}\right)^2 < \lambda_{QM,n-1} < \cdots < \lambda_{QM1} < \left(1 + \sqrt{\frac{n}{q+m}}\right)^2. \tag{12}$$

Note in Eq. (12) that increasing the number m of markers will concentrate the eigen-values more strongly around 1. Hence, for m very large, $\left(\lambda_{QMi} - 1\right) \to 0$ at a faster rate than m/q increases, and $\kappa_{QMi} \to \lambda_{QMi} - (m/q)\delta_i$. Since $\mathbb{E}(\delta_i) = 0$, the ratio m/q determines only the variance of κ_{QMi} around λ_{QMi}. Therefore, a genomic model that contains QTL plus markers that are uncorrelated to the QTL will yield $\hat{\gamma}_{QM}$ such that $\lim_{n\to\infty} \mathbb{E}(\hat{\gamma}_{QM}) = \gamma_Q$.

We also show in "Appendix 5" that for genomic models that include markers only,

$$\kappa_{Mi} = 1 + \delta_i, \quad \forall \quad i = 1, \ldots, n-1, \text{ with } \mathbb{E}(\delta_i) = 0. \tag{13}$$

Therefore, κ_i is a constant, and a genomic model that only contains markers that are uncorrelated to the QTL in the SNP data will always obtain $\hat{\gamma}_M = 0$, when REML is used.

QTL correlated to markers

In "Appendix 6", we show that for genomic models that include the QTL plus markers, in which markers are correlated to the QTL,

$$\kappa_{QMi} = \lambda_{QMi} + \delta_i, \quad \forall \quad i = 1, \ldots, n-1, \quad \text{with } \mathbb{E}(\delta_i) = \sum_{j=1}^{n}\sum_{l=1}^{n}(\sigma_{ijl,Qjl} - \sigma_{ijl,jl}), \tag{14}$$

where $\sigma_{ijl,Qjl} = \mathbb{C}ov(U_{QMij}U_{QMil}, G_{Qjl})$ and $\sigma_{ijl,jl} = \mathbb{C}ov(U_{QMij}U_{QMil}, G_{QMjl})$. It is intuitive that $\sigma_{ijl,jl} \geq \sigma_{ijl,Qjl}$, and therefore $\mathbb{E}(\delta_i) \leq 0$. Thus, a genomic model that contains all QTL and markers that are correlated to the QTL will yield $\hat{\gamma}_{QM}$ such that $\lim_{n\to\infty} \mathbb{E}(\hat{\gamma}_{QM}) = \gamma_Q$, only when $\sigma_{ijl,jl} \approx \sigma_{ijl,Qjl}$, resulting in $\mathbb{E}(\delta_i) \approx 0$.

$\mathbb{E}(\delta_i)$ depends greatly on the distributions of the minor allele frequencies of the QTL and markers, $f_{MAF(QTL)}$ and $f_{MAF(markers)}$, respectively. When $f_{MAF(QTL)} = f_{MAF(markers)}$, unless the number of QTL is very small (say $q \leq 10$), we find that $\mathbf{G}_Q$ and $\mathbf{G}_{QM}$ are very similar. It is intuitively obvious that in this case $\sigma_{ijl,jl} \approx \sigma_{ijl,Qjl}$. Hence, $\mathbb{E}(\delta_i) \approx 0$, and consequently $\lim_{n\to\infty} \mathbb{E}(\hat{\gamma}_{QM}) = \gamma_Q$. When $f_{MAF(QTL)} \neq f_{MAF(markers)}$, the larger the number m of markers, the more different $\mathbf{G}_Q$ and $\mathbf{G}_{QM}$ will be. Moreover, $\lim_{m\to\infty} \mathbf{G}_{QM} = \lim_{m\to\infty} \mathbf{G}_M$. This difference between $\mathbf{G}_Q$ and $\mathbf{G}_{QM}$ will ensure the inequality $\sigma_{ijl,jl} \geq \sigma_{ijl,Qjl}$. Hence, $\mathbb{E}(\delta_i) < 0$, and consequently $\mathbb{E}(\hat{\gamma}_{QM}) < \gamma_Q$.

We show in "Appendix 4" that $\mathbf{G}_Q = \mathbf{A} + \mathbf{\Delta}_Q$, with the number of QTL q being fixed, $\mathbf{G}_{QM} = \mathbf{A} + \mathbf{\Delta}_{QM}$, and $\mathbf{G}_M = \mathbf{A} + \mathbf{\Delta}_M$. The relationship between individuals increases the speed of convergence of $\mathbf{\Delta}_{QM} \to \mathbf{0}$ (when $f_{MAF(QTL)} \neq f_{MAF(markers)}$) and of $\mathbf{\Delta}_M \to \mathbf{0}$, when m increases. Therefore, $\mathbb{V}ar(\delta_{QMij})$ and $\mathbb{V}ar(\delta_{Mij})$ decreases when the number of generations increases. Finally, we find that $\mathbf{G}$ and $\mathbf{G}_Q$ are more similar when the number of generations increases. Thus, the choice of populations with increasing relationship between individuals tends to reduce the bias of heritability estimators (when these are biased). However, when $f_{MAF(QTL)} \neq f_{MAF(markers)}$, stronger relationships are necessary, so the downward-bias becomes less perceptible. The explanation for this is related to the range of the MAF. For any $\mathbf{G} = \mathbf{A} + \mathbf{\Delta}$, $\mathbb{V}ar(\delta_{ij})$ for SNPs with low MAF is lower than $\mathbb{V}ar(\delta_{ij})$ for SNPs with high MAF, and populations with more closely related individuals are necessary to compensate for this difference in $\mathbb{V}ar(\delta_{ij})$ across different MAF ranges.

We also show in "Appendix 6" that for genomic models that include the markers only,

$$\kappa_{Mi} = 1 + \sum_{j=1}^{n} U_{Mij}^2 \delta_{Qjj} + \sum_{j=1}^{n}\sum_{l\neq j} U_{Mij}U_{Mil}\left(a_{jl} + \delta_{Qjl}\right), \quad \forall \quad i = 1, \ldots, n-1. \tag{15}$$

Equation (15) is not so straightforward to understand analytically. Assume that we randomly pick just a few markers. These markers will most likely be in low LD with the QTL and $\kappa_{Mi} \approx 0$, as shown in the previous section. When the density of marker data increases, we obtain Eq. (15). We show in "Appendix 6" that $\lim_{m\to\infty} \kappa_{Mi} = \lim_{m\to\infty} \kappa_{QMi}$. This means that $\lim_{n\to\infty} \mathbb{E}(\hat{\gamma}_M) \leq \lim_{n\to\infty} \mathbb{E}(\hat{\gamma}_{QM})$, with equality only when $m \to \infty$.

Table 1 Relationship between γ and the REML estimators obtained from data with QTL plus markers ($\hat{\gamma}_{QM}$) and markers only ($\hat{\gamma}_{M}$), assuming q fixed and finite, m very large, m > q, and q + m >>> n

Scenario	Population	MAF	QTL/markers	$\lim_{n\to\infty}$	
				$\mathbb{E}(\hat{\gamma}_{QM})$	$\mathbb{E}(\hat{\gamma}_{M})$
1	1 generation*	$f_{MAF_{QTL}} = f_{MAF_{makers}}$	Complete LE	γ	0
2	1 generation*	$f_{MAF_{QTL}} \neq f_{MAF_{makers}}$	Complete LE	γ	0
3	1 generation*	$f_{MAF_{QTL}} = f_{MAF_{makers}}$	LD	γ	$<<< \mathbb{E}(\hat{\gamma}_{QM})$
4	2 generations	$f_{MAF_{QTL}} = f_{MAF_{makers}}$	LD	γ	$<< \mathbb{E}(\hat{\gamma}_{QM})$
5	10 generations	$f_{MAF_{QTL}} = f_{MAF_{makers}}$	LD	γ	$< \mathbb{E}(\hat{\gamma}_{QM})^{\dagger}$
6	1 generation*	$f_{MAF_{QTL}} \neq f_{MAF_{makers}}$	LD	$<<< \gamma$	$<<< \mathbb{E}(\hat{\gamma}_{QM})$
7	2 generations	$f_{MAF_{QTL}} \neq f_{MAF_{makers}}$	LD	$<< \gamma$	$<< \mathbb{E}(\hat{\gamma}_{QM})$
8	10 generations	$f_{MAF_{QTL}} \neq f_{MAF_{makers}}$	LD	$< \gamma^{\dagger}$	$< \mathbb{E}(\hat{\gamma}_{QM})^{\dagger}$

*Completely unrelated individuals

† $\lim_{g\uparrow} h^2_M = \lim_{g\uparrow} h^2_{QM} = h^2$ for a large number g of generations (strongly related individuals)

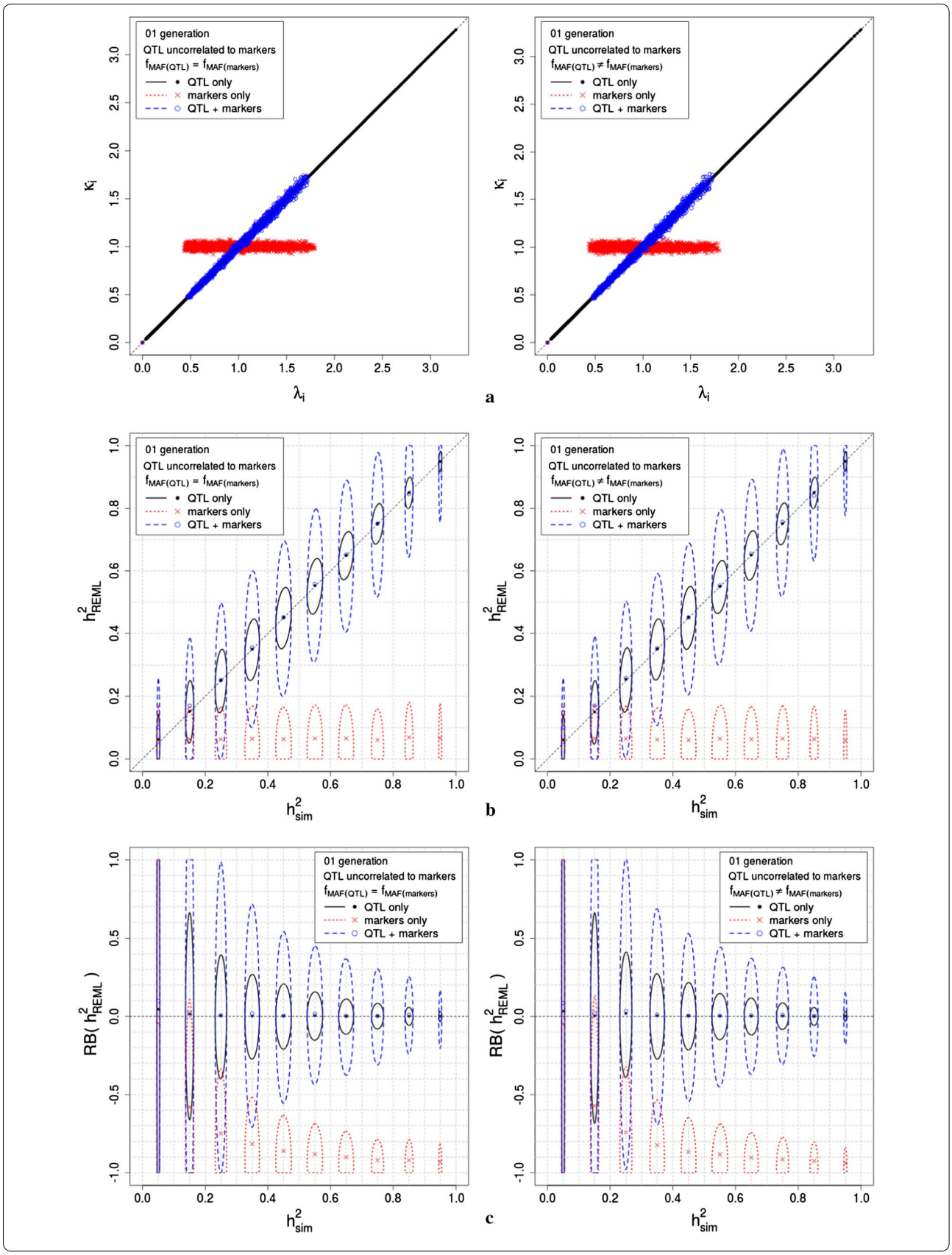
Simulations

Table 1 summarizes what is expected regarding the estimation of the heritability for a set of scenarios that are relevant in quantitative genetics studies, based on the theory detailed in the section Conditions for unbiased REML estimators. The REML estimation of γ was performed on data containing QTL plus markers ($\hat{\gamma}_{QM}$) and markers only ($\hat{\gamma}_{M}$). It is known that $\hat{\gamma}_{M} < \hat{\gamma}_{QM}$, since the markers alone cannot capture more genetic variability than SNP data that contains both QTL and markers [12]. However, for scenarios in which markers are in LD with the QTL $\lim_{m\to\infty} \hat{\gamma}_{M} = \lim_{m\to\infty} \hat{\gamma}_{QM}$. When individuals are completely unrelated, such convergence is most probably unrealistic even with sequence data (although $\hat{\gamma}_{M}$ approaches $\hat{\gamma}_{QM}$). In populations with strongly related individuals $\hat{\gamma}_{M} \longrightarrow \hat{\gamma}_{QM}$ for m finite and sufficiently large.

Figure 1 presents the simulation results for scenarios 1 and 2, Fig. 2 presents the simulation results for scenarios 3, 4 and 5, and Fig. 3 presents the simulation results for scenarios 6, 7 and 8. All three figures show the results for simulations that assigned 3000 SNPs as QTL. The results for the simulations that assigned 100 SNPs as QTL differed from those presented in Figs. 1, 2 and 3 only by a larger variance around the same means. In all three figures, panel (a) shows the relationship between λ_i and κ_i, for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); the relationship in the simulated data agreed with the theory in the section Genomic models for scenarios of interest in quantitative genetics, for QTL uncorrelated and correlated to markers, respectively. In all three figures, panel (b) presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the heritabilities estimated using REML ($h^2_{REML} = \gamma_{REML}/(1+\gamma_{REML})$), for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); h^2_{sim} was very stable around the simulation parameters, and h^2_{REML} had confidence intervals that agreed with the results in Table 1. Note that when QTL were correlated with markers, in scenarios 3 to 8, the variability of h^2_{REML} was smaller than that of h^2_{REML} when QTL were uncorrelated with markers, in scenarios 1 and 2. In all three figures, panel (c) presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the relative bias of h^2_{REML} ($RB(h^2_{REML}) = (h^2_{REML} - h^2_{sim})/h^2_{sim}$). Note that when QTL were correlated with markers, in scenarios 3 to 8, the variability of $RB(h^2_{REML})$ was smaller than that of $RB(h^2_{REML})$ when QTL were uncorrelated with markers, in scenarios 1 and 2. Note as well, that the variability of $RB(h^2_{REML})$ decreases when h^2_{sim} increases. For scenarios 6, 7 and 8, in which QTL were correlated with markers and $f_{MAF(QTL)} \neq f_{MAF(makers)}$, we found that increasing the number of generations (thus, increasing the relationship between simulated individuals) decreases the bias of estimation and the variability of h^2_{REML} and $RB(h^2_{REML})$.

(See figure on next page)

Fig. 1 Simulation results for scenarios 1 and 2, consisting of one generation of completely unrelated individuals, with QTL and markers in complete linkage equilibrium (LE), for both $f_{MAF_{QTL}} = f_{MAF_{markers}}$ and $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$. Simulations were performed with 3000 QTL generating the phenotypes, replicated 500 times. **a** shows the relationship between λ_i and κ_i, for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); **b** presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the heritabilities estimated using REML ($h^2_{REML} = \gamma_{REML}/(1+\gamma_{REML})$), for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); **c** presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the relative bias of h^2_{REML} ($RB(h^2_{REML}) = (h^2_{REML} - h^2_{sim})/h^2_{sim}$)


01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} = f_{MAF(markers)}$
QTL only
markers only
QTL + markers
κ_i
λ_i
a
01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} \neq f_{MAF(markers)}$
QTL only
markers only
QTL + markers
κ_i
λ_i
01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} = f_{MAF(markers)}$
QTL only
markers only
QTL + markers
h^2_{REML}
h^2_{sim}
b
01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} \neq f_{MAF(markers)}$
QTL only
markers only
QTL + markers
h^2_{REML}
h^2_{sim}
01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} = f_{MAF(markers)}$
QTL only
markers only
QTL + markers
$RB(h^2_{REML})$
h^2_{sim}
c
01 generation
QTL uncorrelated to markers
$f_{MAF(QTL)} \neq f_{MAF(markers)}$
QTL only
markers only
QTL + markers
$RB(h^2_{REML})$
h^2_{sim}

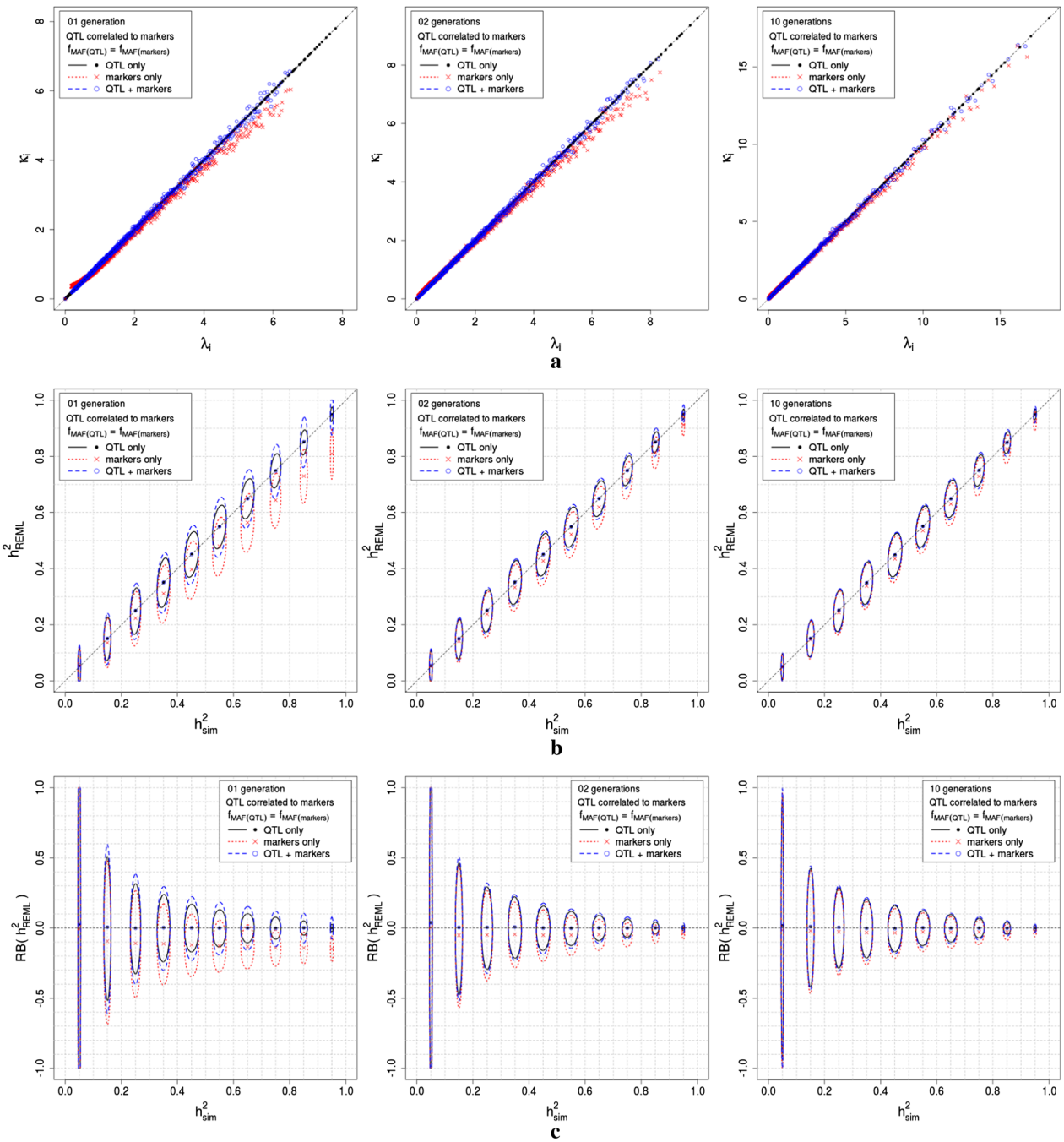


Fig. 2 Simulation results for scenarios 3, 4 and 5, consisting of one generation of completely unrelated individuals and two and 10 generations of related individuals, with QTL and markers in LD, for $f_{MAF_{QTL}} = f_{MAF_{markers}}$. Simulations were performed with 3000 QTL generating the phenotypes, replicated 500 times. **a** shows the relationship between λ_i and κ_i, for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); **b** presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the heritabilities estimated using REML ($h^2_{REML} = \gamma_{REML}/(1+\gamma_{REML})$), for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); **c** presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the relative bias of h^2_{REML} ($RB(h^2_{REML}) = (h^2_{REML} - h^2_{sim})/h^2_{sim}$)

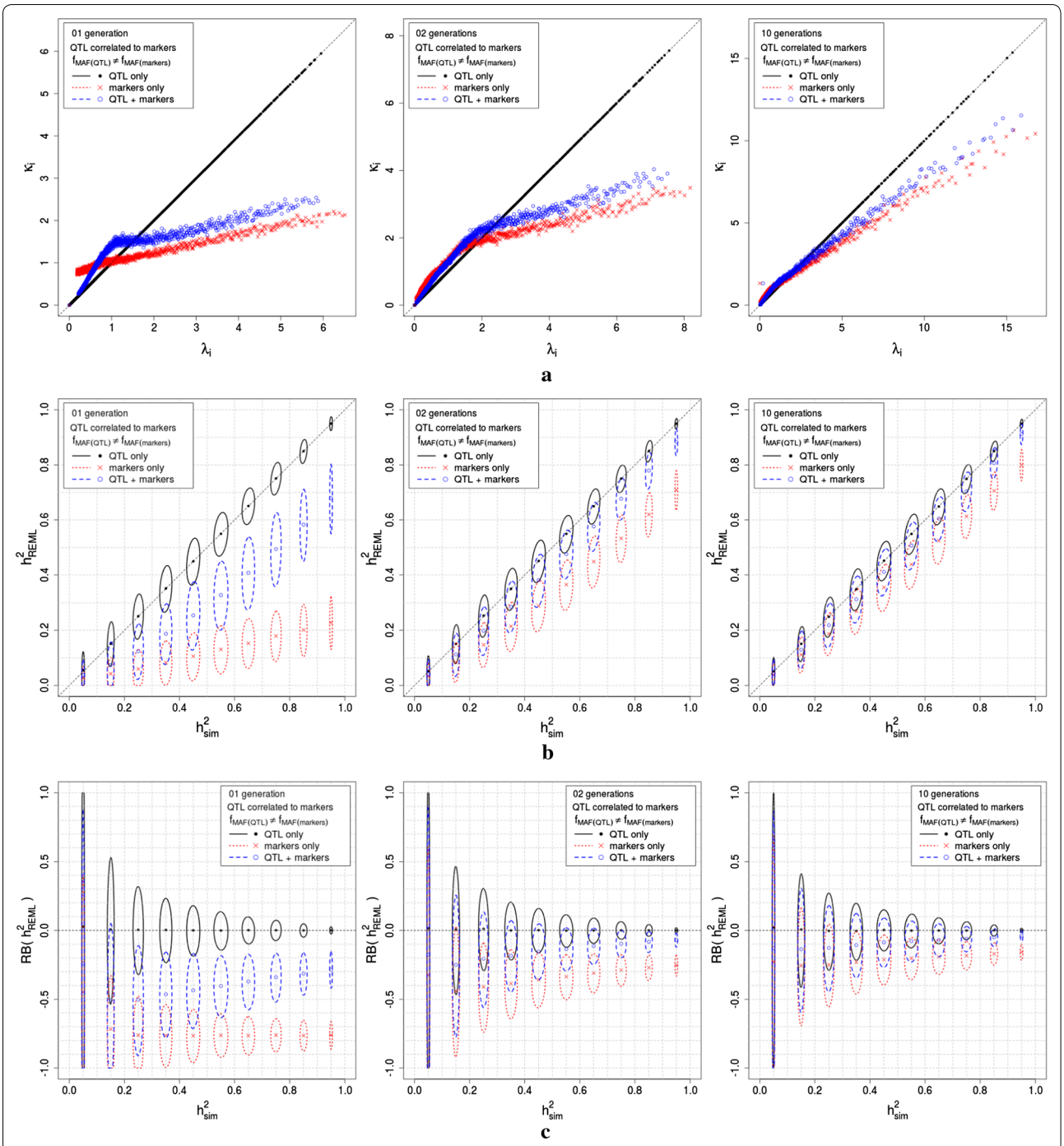


Fig. 3 Simulation results for scenarios 6, 7 and 8, consisting of one generation of completely unrelated individuals and two and 10 generations of related individuals, with QTL and markers in LD, for $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$. Simulations were performed with 3000 QTL generating the phenotypes, replicated 500 times. Panel (a) shows the relationship between λ_i and κ_i, for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); panel (b) presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the heritabilities estimated using REML ($h^2_{REML} = \gamma_{REML}/(1+\gamma_{REML})$), for the true model (QTL only) and for both genomic models evaluated (QTL plus markers and markers only); panel (c) presents the confidence ellipses for the simulated heritabilities ($h^2_{sim} = \gamma_{sim}/(1+\gamma_{sim})$), with a simulation parameter $h^2 = 0.05, 0.15, \ldots, 0.95$, and the relative bias of h^2_{REML} ($RB(h^2_{REML}) = (h^2_{REML} - h^2_{sim})/h^2_{sim}$)

Discussion

We have performed a theoretical analysis of the likelihood equations of genomic models based on splitting these equations into components in order to isolate and identify those that contribute to incorrect inferences. We have shown that the term in the likelihood equations that is responsible for producing potentially biased heritability estimators ($\hat{h}^2_{gen}$) is in fact a measure that evaluates whether **G** provides a proper description of $\mathbf{G}_Q$ or not.

The key measure to evaluate whether bias will arise in the REML heritability estimators is $\kappa_i = \sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}$, for every $i = 1, \ldots, n$, as we have shown in the section Conditions for unbiased REML estimators. Elements $\kappa_1, \ldots, \kappa_n$ correspond to the diagonal of $\mathbf{U}'\mathbf{U}_Q\mathbf{\Lambda}_Q\mathbf{U}_Q'\mathbf{U}$, and the condition for unbiased $\hat{h}^2_{gen}$ is that the portion of variance explained by the i-th component of **G** (defined by λ_i) is equivalent to the sum of weighted correlations between its corresponding eigen-vector and the eigen-vectors of $\mathbf{G}_Q$, *i.e.* $\kappa_i = \lambda_i$. This identity is equivalent to saying that $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ are similar matrices in the general linear group of $\mathbf{U}_Q'\mathbf{U}$. Evaluation of this similarity of $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ is much more informative than a direct comparison of the elements of $\mathbf{G}_Q$ with those of **G**, or a comparison of their eigen-values (see Additional file 1, which presents the distribution of λ_i and κ_i, Additional files 2, 3 and 4, which present the scatterplots of λ_i versus $(\mathbf{y}'\mathbf{U}_i)^2$, and Additional files 5 and 6, which present respectively a scenario where $\lambda_{QMi} \neq \lambda_{Qi}$ and $\kappa_{QMi} = \lambda_{QMi}$ with $\mathbb{E}(\hat{h}^2_{QM}) = h^2$, and a scenario where $\lambda_{Mi} = \lambda_{Qi}$ and $\kappa_{Mi} \neq \lambda_{Mi}$ with $\mathbb{E}(\hat{h}^2_M) \neq h^2$). Hence, in the scenarios studied here, we can detect when a genomic relationship estimated by the SNPs correctly represents the true genetic relationships by verifying whether $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ are similar matrices in the general linear group of $\mathbf{U}_Q'\mathbf{U}$, by comparing κ_i with λ_i. This comparison allows us to determine the presence and direction of the bias, as described in the section Conditions for unbiased REML estimators.

Scenarios in which QTL and markers are in LE have been explored theoretically in other studies, with particular emphasis on the effect of the eigen-values of **G** on the likelihood of the (misspecified) genomic model [8, 13]. Although Kumar et al. [8] discussed the relevance of the difference between the eigen-vectors of **G** and $\mathbf{G}_Q$, they did not relate this difference expressed by the correlation between the eigen-vectors, implicit in κ_i, to the portion of variance explained by each λ_i and λ_{Qk}. The performance of genomic models in estimating heritability was assessed mainly by describing the sensitivity of the likelihood to changes in the eigen-values, under the Marčenko–Pastur distribution [18]. Indeed, the likelihood depends sensitively on all the eigen-values, but evaluating the likelihood given a change in each eigen-value separately is not as informative as evaluating the REML equation given a change in the distribution of all eigen-values simultaneously.

Assuming that the number of individuals (n) is sufficiently large, and that the numbers of QTL and markers (q and m) both increase with increasing density of SNP data such that $\lim_{n,q,m\to\infty} n/(q+m) = c_1 \in (0,1)$ and $\lim_{q,m\to\infty} q/m = c_2 \in (0,1)$, Jiang et al. [13] used the Marčenko–Pastur distribution of the eigenvalues to evaluate the limiting behavior of the term $\mathbf{PGP}/\mathrm{tr}(\mathbf{PG}) - \mathbf{P}^2/\mathrm{tr}(\mathbf{P})$ in the REML Eq. (3) of the genomic model, proving that under these assumptions, $\hat{h}^2_{gen}$ is unbiased and consistent. Although Jiang et al. [13] stated that $\hat{h}^2_{gen}$ still remains unbiased and consistent when QTL and markers are in LD, we have raised a particular concern about inferences of h^2 in the case when MAF(QTL) $\neq$ MAF(markers), such as when QTL are rare mutations, for which the estimate of heritability are empirically known to be biased [1, 4, 8, 19]. Two remarks about the approach used in [13] must be made at this stage. First, the limiting behavior of $\mathbf{PGP}/\mathrm{tr}(\mathbf{PG}) - \mathbf{P}^2/\mathrm{tr}(\mathbf{P})$ relies strongly on the Marč enko–Pastur distribution, which holds only when the SNPs are in complete LE (and thus individuals are unrelated, since family relationships would induce LD). The second remark is that LD between markers and QTL and the distribution of their MAFs may alter correlations between phenotypes and genotypes, implied in $\mathbf{y}'\mathbf{PGPy}$ and $\mathbf{y}'\mathbf{Py}$, and this was not evaluated by Jiang et al. [13], whereas in our study the correlations between phenotypes and genotypes are implied in κ_i (see "Appendix 2 and 3"). Using our approach and the result that $\lim_{q,m\to\infty} \mathbf{G} = \lim_{q\to\infty} \mathbf{G}_Q = \mathbf{A}$ [17], we demonstrated in section Genomic models for scenarios of interest in quantitative genetics, that the conclusions from Jiang et al. [13] about $\hat{h}^2_{gen}$ are mathematically true, even when QTL and markers are in LD. However, in populations of unrelated individuals with QTL as rare mutations, an unrealistically large number (tending to infinity) of QTL would be necessary to ensure $\lim_{q\to\infty} \mathbf{G}_Q = \mathbf{A}$, and when we assume that the number of QTL is finite, $\lim_{m\to\infty} \mathbf{G} \neq \mathbf{G}_Q$.

With the knowledge that the method proposed by Yang et al. [1] may yield a biased $\hat{h}^2_{gen}$ under some scenarios, several approaches have been proposed for solving the problem of biased estimates, exploring different genetic architectures of the trait and population structure. The theory presented in our study can be adapted to provide an explanation for the success of those approaches, and we offer an overview on how that can be done for five approaches.

First, addressing the different MAF of the SNPs, Speed et al. [4] suggests a weighting of the SNPs by their MAF,

which would give the same weighting to terms involving γ in the non-observable REML functions, owing to the change in the definition of the heritability. A G-matrix obtained using the SNPs suitably weighted according to the scenario will improve the relationship between κ_i and λ_i, reducing the bias of $\hat{h}^2_{gen}$. The definition of a suitable weight must be explored further, and the theory provided in this study provides a tool that can be used for such investigations. The genomic model in (1) can be generalized to assume different weights to the SNPs by simply changing the assumption for the distribution of $\mathbf{b}$ to $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}\sigma^2_b)$, such that $\mathbf{D}$ is a diagonal matrix of weights. The G-matrix must then be defined as $\mathbf{G} = \mathbf{WDW}'/\text{tr}(\mathbf{D})$ to ensure that properties indicated in "Appendix 1" and theoretical evaluations in the Results section hold.

Second, and with the same objective of distinguishing SNPs by their MAF, Yang et al. [19] suggested a method that is analogous to that proposed in [1] by fitting the model with several genomic variance components, each of them relative to groups of SNPs with MAF values within the same range. In this approach, assuming the components to be independent, we can obtain a non-observable REML equation for each genomic variance component to be estimated, and the analysis then follows exactly as we have presented here. This approach also generalizes the genomic model in Eq. (1) by assuming $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}\sigma^2_b)$, as described in the previous paragraph, with the difference that the values on the diagonal of $\mathbf{D}$ are to be estimated. Indeed the method is capable of removing the bias of $\hat{h}^2_{gen}$. However, as observed by [19], the increase in the number of variance components may increase the variance of $\hat{h}^2_{gen}$, especially when independence between the components cannot be ensured, and, depending on the scenario evaluated, the estimates may be less reliable than those obtained by fitting a single genomic variance component.

Edwards et al. [20] suggested a third approach, which fits genomic models by including a variance component for SNPs grouped based on genomic features (i.e. genes and their gene ontology) to the model, which requires the use of prior information about the genomic data. The results of their study showed that a relevant amount of variance was attributed to the significant feature, and $\hat{h}^2_{gen}$ in this approach can be evaluated with a non-observable REML equation for each component (SNPs grouped based on genomic features and SNPs not grouped based on genomic features), just as we suggest for evaluating $\hat{h}^2_{gen}$ obtained with the approach proposed by [19]. It is important to note that the genomic feature model works better than a single component when the feature component is enriched for the QTL; otherwise, this model can also lead to problems in the estimation of h^2. The advantage of grouping SNPs based on genomic features instead of MAF is that there are fewer variance components, reducing the variance of $\hat{h}^2_{gen}$. The use of prior genomic information to fit genomic models with mutiple genomic variance components was previously suggested by Speed and Balding [21], who included a dynamic procedure to define a suitable partition of SNPs.

A fourth approach considers the situation where prior genomic feature information is absent and Bayesian mixture models, such as BayesB [22] or BayesR [23], are reasonable solutions for assigning different distributions to groups of SNP effects [20]. Again, non-observable REML equations for each component can be used to evaluate $\hat{h}^2_{gen}$ based on the assumptions posed by the Bayesian mixture models, and the assumptions can be tuned using the information from our suggested theoretical analysis.

A fifth approach includes related individuals to study populations, which can greatly reduce the bias of $\hat{h}^2_{gen}$, when it exists. This is because rare QTL induce genetic relationships between individuals. In populations of nominally unrelated individuals the common markers will disguise those induced genetic relationships at the QTL ($\lim_{m\to\infty} \mathbf{G}_{QM} = \mathbf{A} = \mathbf{I}_n \neq \mathbf{G}_Q$), drastically reducing the correlations between eigen-vectors $(1/n)\mathbf{U}'_i\mathbf{U}_{Qk}$ and resulting in $\kappa_i < \lambda_i$. Conversely, in populations of related individuals, assuming no selection, the induced genetic relationships at the QTL will better reflect the kinship matrix ($\mathbf{G}_Q \approx \mathbf{A}$), improving the correlation between eigen-vectors $(1/n)\mathbf{U}'_i\mathbf{U}_{Qk}$ and resulting in less biased $\hat{h}^2_{gen}$.

A last point to be raised in this discussion, concerns the direction of the bias of $\hat{h}^2_{gen}$. We show in the section Genomic models for scenarios of interest in quantitative genetics, with our theoretical analysis, that when we consider a single genomic component in the model to estimate heritability (assuming SNP effects are all *i.i.d.*), when it exists, the bias of the estimator will tend to be downward. An exception is observed when $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$ and the number of QTL is smaller than the number of individuals. If genomic models are fitted including the QTL and markers, such that the total number of SNPs in the genomic data is smaller than the number of individuals, the heritability estimator will present an upwards bias. This fact is related to $\text{rank}(\mathbf{G}_{QM}) < n - 1$, as eigen-values that are zero are overestimated by κ_i. Increasing the number of markers will make $\text{rank}(\mathbf{G}_{QM})$ approach $n - 1$, forcing only the last eigen-value to zero and the overestimation will no longer be present (see Additional file 7).

When multiple genomic components are considered in the model, overestimation of heritability may be observed even when the total number of SNPs is larger than the number of individuals. When different variance parameters are estimated for each component, the

multiple components approach is explicit, and overestimation of heritability will relate to components with a rank lower than n − 1. When a single variance parameter is estimated for SNPs associated with pre-determined weights, the multiple component approach is implicit, and overestimation of heritability as observed in [24] may relate to κ_i overestimating the largest λ_i. Associating pre-determined weights to the SNPs in the genomic model may inflate correlations $\mathbf{U}_i'\mathbf{U}_{Qk}$ for eigen-vectors $\mathbf{U}_i$ that are associated with the highest eigen-values, resulting in $\kappa_i > \lambda_i$, while correlations $\mathbf{U}_i'\mathbf{U}_{Qk}$ for eigen-vectors $\mathbf{U}_i$ that are associated with the lowest eigen-values will be deflated, resulting in $\kappa_i < \lambda_i$ (see Additional file 8).

Conclusions

In a Gaussian setup, the likelihood of a genomic model is misspecified with respect to that of the true model that conceptually generated the data. When used for inferring variance parameters the misspecified likelihood may yield biased estimators of those parameters, and inferences must be interpreted with caution. Misspecification of the likelihood is due to the difference between the covariance structures of the data specified by the misspecified and true models ($\mathbf{G}$ and $\mathbf{G}_Q$), and our study shows that the bias of REML estimators of variance parameters is linked to the relationship between the eigen-values and eigen-vectors of both models, occurring when $\kappa_i = \sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk} \neq \lambda_i$. Moreover, comparison of κ_i with the eigen-value λ_i not only identifies the potential bias of variance components estimators, but is also a very informative method for comparing $\mathbf{G}$ with $\mathbf{G}_Q$. The eigen-vectors reflect how each individual contributes to the proportion of variance explained by the components of $\mathbf{G}$ and $\mathbf{G}_Q$ (defined by λ_i and λ_{Qk}), and if the contributions are similar, then $\kappa_i \approx \lambda_i$, meaning that the covariance structures of the data specified by the genomic and the true models are equivalent. In mathematical terms, $\kappa_i = \lambda_i$ is the same as stating that $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ are similar matrices in the general linear group of $\mathbf{U}_Q'\mathbf{U}$. We have evaluated the similarity of $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ in a set of scenarios of interest to quantitative genetics studies, identifying those for which inferences must be interpreted with caution. Because of the many factors related to the genetic architecture that influence the similarity of $\mathbf{\Lambda}_Q$ and $\mathbf{\Lambda}$ (LD between QTL and markers, presence and number of QTL in the SNP data, MAF, relationship between individuals) and the lack of information about the QTL, quantifying the bias (when it exists) of the estimators of variance parameters, is not trivial. Although the quantification of this bias is complex, we can determine that in genomic models that consider a single genomic component to estimate heritability (assuming SNP effects are all *i.i.d.*), the bias of the estimator will tend to be downward, when it exists.

Additional files

Additional file 1. Distribution of λ_i and κ_i for the eight scenarios, when considering 3000 QTL and 2000 individuals. Distributions are presented for **G** assuming the true model (QTL only) and assuming both genomic models evaluated (QTL plus markers and markers only).

Additional file 2. Scatterplot of λ_i versus $(\mathbf{y}'\mathbf{U}_i)^2$ for one replicate on all the eight scenarios, when considering 3000 QTL and 2000 individuals. Distributions are presented for **G** assuming the true model (QTL only).

Additional file 3. Scatterplot of λ_i versus $(\mathbf{y}'\mathbf{U}_i)^2$ for one replicate on all the eight scenarios, when considering 3000 QTL and 2000 individuals. Distributions are presented for **G** assuming the genomic model with markers only.

Additional file 4. Scatterplot of λ_i versus $(\mathbf{y}'\mathbf{U}_i)^2$ for one replicate on all the eight scenarios, when considering 3000 QTL and 2000 individuals. Distributions are presented for **G** assuming the genomic model with QTL plus markers.

Additional file 5. Comparison of **G** to $\mathbf{G}_Q$ for 2000 completely unrelated individuals by: (1) direct comparison of values, (2) comparison of eigen-values (λ_i), and (3) comparison of λ_i to κ_i, and comparison of $\hat{h}^2_{REML}$ to h^2 simulated with parameter 0.25,0.5,0.75, for 100 replicates of each simulation parameter. $\mathbf{G}_Q$ consisted of 100 QTL and **G** consisted of these 100 QTL plus 1900 markers in complete LE with the QTL, such that $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$.

Additional file 6. Comparison of **G** to $\mathbf{G}_Q$ for 2000 completely unrelated individuals by: (1) direct comparison of values, (2) comparison of eigen-values (λ_i), and (3) comparison of λ_i to $\boldsymbol{\kappa_i}$, and comparison of $\hat{h}^2_{REML}$ to h^2 simulated with parameter 0.25,0.5,0.75, for 100 replicates of each simulation parameter. $\mathbf{G}_Q$ consisted of 3000 QTL and **G** consisted of 3000 markers in complete LE with the QTL, such that $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$.

Additional file 7. 95\% confidence intervals, based on 100 replicates, for the absolute bias of $\hat{h}^2_{REML}$ to h^2. Genomic data consisted of QTL plus markers for 2000 completely unrelated individuals. The number of markers in the SNP data varied from 0 to 19,900, and 100 QTL were used to simulate the phenotypes with h^2 parameter 0.6. QTL and markers were in LD, such that $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$.

Additional file 8. Relationship between λ_i and $\boldsymbol{\kappa_i}$, and absolute bias of $\hat{h}^2_{REML}$ to h^2. Genomic data consisted of QTL plus markers for 2000 completely unrelated individuals. 100 QTL were used to simulate the phenotypes with h^2 parameter 0.05,0.15,...,0.95. QTL and markers were in LD, such that $f_{MAF_{QTL}} \neq f_{MAF_{markers}}$. Markers were weighted according to their LD with the QTL.

Authors' contributions

BCDC conceived the study, performed the calculus, simulations and analysis, and wrote the manuscript. ACS helped to conceive the study, helped with the analysis, and contributed to the manuscript. PS conceived the study, helped with the calculus and the analysis, and contributed to the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Daniel Sorensen, who provided fruitful discussions and insights relevant to this work, and Gustavo de los Campos, who contributed with the initial ideas for the simulations. This research was supported and funded by the Center for Genomic Selection in Animals and Plants funded by the Danish Council for Strategic Research (contract number 12-132452).

Competing interests

The authors declare that they have no competing interests.

Appendix 1: REML equation to REML function

Given any SNP genotypes matrix $\mathbf{W}$ with s SNPs, we define $\mathbf{G} = s^{-1}\mathbf{W}\mathbf{W}' = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$ where $\mathbf{U}$ and $\boldsymbol{\Lambda}$ are the eigen-decomposition matrices of $\mathbf{G}$, the variance matrix $\mathbf{V} = \gamma\mathbf{G} + \mathbf{I}_n$, such that n is the number of individuals in the population, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{1}_n\left(\mathbf{1}_n'\mathbf{V}^{-1}\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\mathbf{V}^{-1}$. Using the eigen-decomposition of $\mathbf{G}$, it is straightforward that $\mathbf{V}^{-1} = \mathbf{U}(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}\mathbf{U}'$. Thus,

$$\mathbf{P} = \mathbf{U}\left[(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1} - \frac{(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}\mathbf{U}'\mathbf{1}_n\mathbf{1}_n'\mathbf{U}(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}}{\mathbf{1}_n'\mathbf{U}(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}\mathbf{U}'\mathbf{1}_n}\right]\mathbf{U}', \quad (16)$$

and since $(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}\mathbf{U}'\mathbf{1}_n\mathbf{1}_n'\mathbf{U}(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}$ has elements $n^2\bar{U}_i\bar{U}_j/\left[(1+\gamma\lambda_i)(1+\gamma\lambda_j)\right]$, and the scalar $\mathbf{1}_n'\mathbf{U}(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1}\mathbf{U}'\mathbf{1}_n = n^2\sum_{i=1}^n \bar{U}_i^2/(1+\gamma\lambda_i)$,

$$\mathbf{P} = \mathbf{U}\left\{(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1} - \left(\sum_{i=1}^n \frac{\bar{U}_i^2}{1+\gamma\lambda_i}\right)^{-1}\left[\frac{\bar{U}_i\bar{U}_j}{(1+\gamma\lambda_i)(1+\gamma\lambda_j)}\right]_{i,j=1}^n\right\}\mathbf{U}'. \quad (17)$$

We use the following properties of the eigen-decomposition, when n is sufficiently large, to go through the next steps:

ED1. $\lambda_1 \geq \cdots \geq \lambda_{n-1} \geq \lambda_n = 0$;

ED2. $s < n \Longrightarrow \lambda_i = 0, \forall\, i > s$;

ED3. $\sum_{i=1}^n \lambda_i = n - 1$;

ED4. $\mathbf{U} = [\mathbf{U}_1 \cdots \mathbf{U}_n]: \ \mathbf{U}_i' = [U_{i1} \cdots U_{in}], \quad \forall\, i = 1, \ldots, n$;

ED5. $\lambda_i > 0 \Longrightarrow \mathbf{1}_n'\mathbf{U}_i = \sum_{j=1}^n U_{ij} = n\bar{U}_i = 0$;

ED6. $\sum_{i=1}^n \left(\sum_{j=1}^n U_{ij}\right)^2 = n \Longrightarrow \sum_{i=1}^n \bar{U}_i^2 = n^{-2}\sum_{i=1}^n \left(\sum_{j=1}^n U_{ij}\right)^2 = n^{-1}$

Thus, we have now that

$$\mathbf{P} = \mathbf{U}\left\{(\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1} - \left(\sum_{i=1}^n \bar{U}_i^2\mathbb{I}_{\{\lambda_i=0\}}\right)^{-1}\left[\bar{U}_i\bar{U}_j\right]_{i,j=1}^n\right\}$$
$$\mathbf{U}' = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}', \quad (18)$$

where $\boldsymbol{\Gamma} = (\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-1} - n\left[\bar{U}_i\bar{U}_k\right]_{i,k=1}^n$. Therefore,

$$\mathbf{y}'\left[\frac{\mathbf{PGP}}{\mathrm{tr}(\mathbf{PG})} - \frac{\mathbf{P}^2}{\mathrm{tr}(\mathbf{P})}\right]\mathbf{y} = \mathbf{y}'\left[\frac{\mathbf{U}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\mathbf{U}'}{\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Lambda})} - \frac{\mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}'}{\mathrm{tr}(\boldsymbol{\Gamma})}\right]$$
$$\mathbf{y} = \frac{\mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\mathbf{U}'\mathbf{y}}{\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Lambda})} - \frac{\mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}'\mathbf{y}}{\mathrm{tr}(\boldsymbol{\Gamma})}, \quad (19)$$

where,

(1) $\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma} = \mathbf{diag}\left(\frac{\lambda_1}{(\gamma\lambda_1+1)^2}, \ldots, \frac{\lambda_n}{(\gamma\lambda_n+1)^2}\right)$,
$\Longrightarrow \mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\mathbf{U}'\mathbf{y} = \sum_{i=1}^n \left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2\lambda_i$

(2) $\boldsymbol{\Gamma}^2 = (\gamma\boldsymbol{\Lambda} + \mathbf{I}_n)^{-2} - n\left[\bar{U}_i\bar{U}_k\right]_{i,k=1}^n \Longrightarrow$,
$\mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}'\mathbf{y} = \sum_{i=1}^n \left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2 - n\bar{y}^2$

(3) $\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Lambda}) = \sum_{i=1}^n \frac{\lambda_i}{\gamma\lambda_i+1} = \sum_{i=1}^{n-1} \frac{\lambda_i}{\gamma\lambda_i+1}$,

(4) $\mathrm{tr}(\boldsymbol{\Gamma}) = \sum_{i=1}^n \frac{1}{\gamma\lambda_i+1} - 1 = \sum_{i=1}^{n-1} \frac{1}{\gamma\lambda_i+1}$.

Now, being $\hat{\gamma}$ the solution of $\mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}\mathbf{U}'\mathbf{y}/\mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Lambda}) - \mathbf{y}'\mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}'\mathbf{y}/\mathrm{tr}(\boldsymbol{\Gamma}) = 0$, then,

$$\left(\sum_{j=1}^{n-1}\frac{\lambda_j}{\gamma\lambda_j+1}\right)^{-1}\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2\lambda_i - \left(\sum_{j=1}^{n-1}\frac{1}{\gamma\lambda_j+1}\right)^{-1}\left[\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2 - n\bar{y}^2\right] = 0, \quad (20)$$

multiplying the identity by $\left[\sum_{j=1}^{n-1}\lambda_j/(\gamma\lambda_j+1)\right]\left[\sum_{j=1}^{n-1}1/(\gamma\lambda_j+1)\right]$,

$$\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2\lambda_i\left(\sum_{j=1}^{n-1}\frac{1}{\gamma\lambda_j+1}\right) - \left(\sum_{j=1}^{n-1}\frac{\lambda_j}{\gamma\lambda_j+1}\right)\left[\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2 - n\bar{y}^2\right] = 0, \quad (21)$$

rewriting with $\sum_{i=1}^{n-1}(\mathbf{y}'\mathbf{U}_i)^2/(\gamma\lambda_i+1)^2$ in evidence,

$$\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2\left[\sum_{j=1}^{n-1}\frac{\lambda_i}{\gamma\lambda_j+1} - \sum_{j=1}^{n-1}\frac{\lambda_j}{\gamma\lambda_j+1}\right] + n\bar{y}^2\sum_{j=1}^{n-1}\frac{\lambda_j}{\gamma\lambda_j+1} = 0, \quad (22)$$

and symplifying further,

$$\sum_{i=1}^{n-1}\left(\frac{\mathbf{y}'\mathbf{U}_i}{\gamma\lambda_i+1}\right)^2\left[\sum_{j=1}^{n-1}\frac{\lambda_i-\lambda_j}{\gamma\lambda_j+1}\right] + n\bar{y}^2\sum_{j=1}^{n-1}\frac{\lambda_j}{\gamma\lambda_j+1} = 0. \quad (23)$$

Finally, $\hat{\gamma}$ is the root of what we now refer to as REML function,

$$g(\gamma) = \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{(\mathbf{y}'\mathbf{U}_i)^2}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) + n\bar{y}^2\sum_{j=1}^{n-1}\frac{\lambda_j}{1+\gamma\lambda_j}. \quad (24)$$

Appendix 2: Correlations between phenotypes and eigen-vectors

As defined by the true model, the phenotypes are $\mathbf{y} = \mathbf{1}_n\mu + \mathbf{W}_Q\mathbf{b}_Q + \boldsymbol{\varepsilon}_\mathbf{Q}$. Since $\boldsymbol{\varepsilon}_\mathbf{Q}$ is assumed (and simulated) to be independent from all the other elements, $\boldsymbol{\varepsilon}'_\mathbf{Q}\mathbf{U}_i = n\hat{\mathbb{C}}ov(\boldsymbol{\varepsilon}_\mathbf{Q}, \mathbf{U}_i) \to 0$ for n sufficiently large. Thus,

$$\mathbf{y}'\mathbf{U}_i = \mathbf{1}'_n\mathbf{U}_i\mu + \mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i + \boldsymbol{\varepsilon}'_\mathbf{Q}\mathbf{U}_i \longrightarrow \mathbf{1}'_n\mathbf{U}_i\mu + \mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i, \tag{25}$$

and the squared term,

$$\left(\mathbf{y}'\mathbf{U}_i\right)^2 \longrightarrow \left(n\mu\bar{U}_i\right)^2 + 2n\mu\bar{U}_i\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_{Qi} + \left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2. \tag{26}$$

Assuming the true model, $\mathbf{U}_i = \mathbf{U}_{Qi} \Longrightarrow \bar{U}_{Qi} = 0$, $\forall \ \lambda_{Qi} > 0$ and $\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i = 0$, $\forall \ \lambda_{Qi} = 0$. Therefore, $2n\mu\bar{U}_{Qi}\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_{Qi} = 0$, $\forall \ i = 1, \ldots, n$. Assuming the genomic model, we will assume that the number of SNPs is greater than the number of individuals, *i.e.* $s > n$, meaning that $\lambda_1 \geq \cdots \geq \lambda_{n-1} > \lambda_n = 0$, $\bar{U}_1 = \cdots = \bar{U}_{n-1} = 0$, and $\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_n = 0$. Therefore, $2n\mu\bar{U}_{Qi}\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_{Qi} = 0, \forall\, i = 1, \ldots, n$. Finally,

$$\left(\mathbf{y}'\mathbf{U}_i\right)^2 \longrightarrow \left(n\mu\bar{U}_i\right)^2 + \left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2. \tag{27}$$

Appendix 3: Non-observable REML function

When we use the result from (27) into the REML function (24), we obtain the following:

$$\begin{aligned} g(\gamma) = &\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\left(\frac{n\mu\bar{U}_i}{1+\gamma\lambda_i}\right)^2\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) \\ &+ \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) \\ &+ n\bar{y}^2\sum_{j=1}^{n-1}\frac{\lambda_j}{1+\gamma\lambda_j}, \end{aligned} \tag{28}$$

in which we rewrite the elements of the first term as

$$\begin{aligned} &\left(\frac{n\mu\bar{U}_i}{1+\gamma\lambda_i}\right)^2\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) \\ &= n^2\mu^2\left[\left(\frac{\bar{U}_i}{1+\gamma\lambda_i}\right)^2\frac{\lambda_i}{1+\gamma\lambda_j} - \left(\frac{\bar{U}_i}{1+\gamma\lambda_i}\right)^2\frac{\lambda_j}{1+\gamma\lambda_j}\right]. \end{aligned} \tag{29}$$

Now, since $\bar{U}_i = 0, \ \forall \ \lambda_i > 0 \Longrightarrow (\bar{U}_i)^2\lambda_i = 0, \ \forall\, i = 1, \ldots, n$. Thus,

$$\left(\frac{n\mu\bar{U}_i}{1+\gamma\lambda_i}\right)^2\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) = -n^2\mu^2\left(\frac{\bar{U}_i}{1+\gamma\lambda_i}\right)^2\frac{\lambda_j}{1+\gamma\lambda_j}. \tag{30}$$

Because $\bar{U}_i \neq 0$ only when $\lambda_i = 0$ and $\sum_{i=1}^{n}\bar{U}_i^2 = n^{-1}$,

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\left(\frac{n\mu\bar{U}_i}{1+\gamma\lambda_i}\right)^2\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) = -n\mu^2\sum_{j=1}^{n-1}\frac{\lambda_j}{1+\gamma\lambda_j}. \tag{31}$$

When n is sufficiently large, $\bar{y} \longrightarrow \mu$, and $n\bar{y}^2\sum_{j=1}^{n-1}\lambda_j/(1+\gamma\lambda_j) - n\mu^2\sum_{j=1}^{n-1}\lambda_j/(1+\gamma\lambda_j) \longrightarrow 0$. Therefore:

$$g(\gamma) \longrightarrow \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right). \tag{32}$$

We now move to understand some properties of $\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i$, keeping in mind that in the assumptions (and simulations) we make in our study the QTL are independent from each other, and the QTL effects are independent from all the other elements. Another important assumption used here is that $\mathbf{W}'_{Q1}\mathbf{U}_i, \ldots, \mathbf{W}'_{Qq}\mathbf{U}_i$ are iid, which allows us to define

$$\begin{aligned} &(1)\ \mathbb{E}\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right) = 0, \\ &(2)\ \mathbb{V}ar\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right) = \sigma^2_{T_Q}\mathbb{V}ar\left(\mathbf{W}'_{Qj}\mathbf{U}_i\right) \\ &\qquad = \mathbb{E}\left(\left[\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right]^2\right). \end{aligned}$$

Thus, we have $\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2$ as an estimator,

$$\begin{aligned} \left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right)^2 &= \hat{\mathbb{E}}\left(\left[\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right]^2\right) \\ &= \hat{\mathbb{V}}ar\left(\mathbf{b}'_Q\mathbf{W}'_Q\mathbf{U}_i\right) = \hat{\sigma}^2_{T_Q}\hat{\mathbb{V}}ar\left(\mathbf{W}'_{Qj}\mathbf{U}_i\right). \end{aligned} \tag{33}$$

And finally, the non-observable REML function:

$$g(\gamma) \longrightarrow \hat{\sigma}^2_{T_Q}\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\frac{\hat{\mathbb{V}}ar\left(\mathbf{W}'_{Qj}\mathbf{U}_i\right)}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right). \tag{34}$$

Replacing $\hat{\mathbb{V}}ar\left(\mathbf{W}'_{Qj}\mathbf{U}_i\right)$ with an estimator $\hat{\mathbb{V}}ar\left(\mathbf{W}'_{Qj}\mathbf{U}_i\right) = q^{-1}\mathbf{U}'_i\mathbf{W}_Q\mathbf{W}'_Q\mathbf{U}_i - q^{-2}(\mathbf{1}'_q\mathbf{W}'_Q\mathbf{U}_i)^2$, $= \mathbf{U}'_i\mathbf{U}_Q\boldsymbol{\Lambda}_Q\mathbf{U}'_Q\mathbf{U}_i - q^{-2}(\mathbf{1}'_q\mathbf{W}'_Q\mathbf{U}_i)^2$

$$g(\gamma) \longrightarrow \hat{\sigma}^2_{T_Q} \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \frac{\left[\mathbf{U}_i'\mathbf{U}_Q\boldsymbol{\Lambda}_Q\mathbf{U}_Q'\mathbf{U}_i - q^{-2}(\mathbf{1}_q'\mathbf{W}_Q'\mathbf{U}_i)^2\right]}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right), \tag{35}$$

and for n sufficiently large, $q^{-2}(\mathbf{1}_q'\mathbf{W}_Q'\mathbf{U}_i)^2 <<< \mathbf{U}_i'\mathbf{U}_Q\boldsymbol{\Lambda}_Q\mathbf{U}_Q'\mathbf{U}_i = \sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}$ with $\sum_{k=1}^{n}(\mathbf{U}_n'\mathbf{U}_{Qk})^2\lambda_{Qk} = 0$ due to properties of the eigen-decomposition. So we can approximate the non-observable REML function,

$$g(\gamma) \longrightarrow \hat{\sigma}^2_{T_Q} \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \frac{\sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right), \tag{36}$$

and since our interest in evaluating $g(\hat{\gamma}) = 0$, it is equivalent to evaluate $\hat{\sigma}^{-2}_{T_Q} g(\hat{\gamma}) = 0$, thus:

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \frac{\sum_{k=1}^{n}(\mathbf{U}_i'\mathbf{U}_{Qk})^2\lambda_{Qk}}{(1+\gamma\lambda_i)^2}\left(\frac{\lambda_i-\lambda_j}{1+\gamma\lambda_j}\right) = 0. \tag{37}$$

We call this function non-observable because we have now written it as a function of $\mathbf{U}_Q$ and $\boldsymbol{\Lambda}_Q$, that cannot be observed directly from phenotypes and genomic data.

Appendix 4: Summations as deviations from the A-matrix

Let $\mathbf{G}_1 = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1 = \mathbf{A} + \boldsymbol{\Delta}_1$ and $\mathbf{G}_2 = \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{U}_2 = \mathbf{A} + \boldsymbol{\Delta}_2$ be two G-matrices based on any set of SNPs, such that $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ are how much $\mathbf{G}_1$ and $\mathbf{G}_2$ respectively, deviate from $\mathbf{A}$, the matrix of expected relationships between the individuals, expressed as correlations. We have then the following:

$$\begin{aligned}
&\sum_{k=1}^{n}\left(\mathbf{U}_{1i}'\mathbf{U}_{2k}\right)^2\lambda_{2k}\\
&= \sum_{j=1}^{n}U_{1ij}^2\sum_{k=1}^{n}U_{2kj}^2\lambda_{2k} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}\sum_{k=1}^{n}U_{2kj}U_{2kl}\lambda_{2k}\\
&= \sum_{j=1}^{n}U_{1ij}^2G_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}G_{2jl}\\
&= \sum_{j=1}^{n}U_{1ij}^2\left(a_{jj}+\delta_{2jj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}\left(a_{jl}+\delta_{2jl}\right)\\
&= \sum_{j=1}^{n}U_{1ij}^2\left(1+\delta_{2jj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}\left(a_{jl}+\delta_{2jl}\right)\\
&= \sum_{j=1}^{n}U_{1ij}^2 + \sum_{j=1}^{n}U_{1ij}^2\delta_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}\left(a_{jl}+\delta_{2jl}\right)\\
&= 1 + \sum_{j=1}^{n}U_{1ij}^2\delta_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1il}\left(a_{jl}+\delta_{2jl}\right),
\end{aligned} \tag{38}$$

and

$$\begin{aligned}
&\sum_{k=1}^{n}\mathbf{U}_{1i}'\mathbf{U}_{2k}\mathbf{U}_{2k}'\mathbf{U}_{1r}\lambda_{2k}\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj}\sum_{k=1}^{n}U_{2kj}^2\lambda_{2k} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}\sum_{k=1}^{n}U_{2kj}U_{2kl}\lambda_{2k}\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj}G_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}G_{2jl}\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj}\left(a_{jj}+\delta_{2jj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}\left(a_{jl}+\delta_{2jl}\right)\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj}\left(1+\delta_{2jj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}\left(a_{jl}+\delta_{2jl}\right)\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj} + \sum_{j=1}^{n}U_{1ij}U_{1rj}\delta_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}\left(a_{jl}+\delta_{2jl}\right)\\
&= \sum_{j=1}^{n}U_{1ij}U_{1rj}\delta_{2jj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{1ij}U_{1rl}\left(a_{jl}+\delta_{2jl}\right),
\end{aligned} \tag{39}$$

such that $a_{jl} \in [0, 0.5]$ is the expected relationship between the j-th and l-th individuals (in a non-inbred population), $a_{jj} = 1$ (the expected relationship of and individual to itself, in a non-inbred population). Because of the way we defined the G-matrices, and $\mathbf{G}_1, \mathbf{G}_2 \rightarrow \mathbf{A}$ when the number of SNPs is sufficiently large [17], it is intuitive that $\mathbb{E}(\delta_{*jj}) = \mathbb{E}(\delta_{*jl}) = 0$, with $\mathbb{V}\mathrm{ar}(\delta_{*jj})$ and $\mathbb{V}\mathrm{ar}(\delta_{*jl})$ proportionally inverse to the number of SNPs.

Appendix 5: QTL and markers in complete linkage equilibrium

E.1 QTL + markers

We will use the following identities relating $\mathbf{G}_{QM}$, $\mathbf{G}_Q$ and $\mathbf{G}_M$ to understand the properties of the REML estimators for the scenarios we simulated. Using the eigen-decompositions, we have:

$$\begin{aligned}
\mathbf{G}_{QM} &= \frac{q}{q+m}\mathbf{G}_Q + \frac{m}{q+m}\mathbf{G}_M\\
&= \mathbf{U}_Q\left(\frac{q}{q+m}\boldsymbol{\Lambda}_Q + \frac{m}{q+m}\mathbf{U}_Q'\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}_M'\mathbf{U}_Q\right)\mathbf{U}_Q'.
\end{aligned} \tag{40}$$

Now, because $\mathbf{G}_{QM} = \mathbf{U}_{QM}\boldsymbol{\Lambda}_{QM}\mathbf{U}_{QM}'$, using that identity in Eq. (40) we have:

$$\mathbf{U}_{QM}\boldsymbol{\Lambda}_{QM}\mathbf{U}'_{QM} = \mathbf{U}_Q\left(\frac{q}{q+m}\boldsymbol{\Lambda}_Q + \frac{m}{q+m}\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q\right)\mathbf{U}'_Q, \tag{41}$$

hence,

$$\boldsymbol{\Lambda}_{QM} = \mathbf{U}'_{QM}\mathbf{U}_Q\left(\frac{q}{q+m}\boldsymbol{\Lambda}_Q + \frac{m}{q+m}\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q\right) \times \mathbf{U}'_Q\mathbf{U}_{QM}. \tag{42}$$

We must now understand what happens to the term $\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q$ in Eq. (42), in order to understand what will happen the REML estimator of heritability. Applying the results from Eqs. (38) and (39), and because in these scenarios we only considered completely unrelated individuals (implying that $a_{jl} = 0$), we have that the elements in the diagonal and off-diagonal of $\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q$ are respectively,

$$\sum_{k=1}^{n}\left(\mathbf{U}'_{Qi}\mathbf{U}_{Mk}\right)^2\lambda_{Mk} = 1 + \sum_{j=1}^{n}U^2_{Qij}\delta_{Mjj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{Qij}U_{Qil}\delta_{Mjl}, \tag{43}$$

and

$$\sum_{k=1}^{n}\mathbf{U}'_{Qi}\mathbf{U}_{Mk}\mathbf{U}'_{Mk}\mathbf{U}_{Qr}\lambda_{Mk} = \sum_{j=1}^{n}U_{Qij}U_{Qrj}\delta_{Mjj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{Qij}U_{Qrl}\delta_{Mjl}. \tag{44}$$

Because the markers are in complete LE with the QTL, and remembering that $\mathbb{E}(\delta_{Mjj}) = \mathbb{E}(\delta_{Mjl}) = 0$, we have:

$$\mathbb{E}\left(\sum_{k=1}^{n}\left[\mathbf{U}'_{Qi}\mathbf{U}_{Mk}\right]^2\lambda_{Mk}\right) = 1 + \sum_{j=1}^{n}\mathbb{E}\left(U^2_{Qij}\delta_{Mjj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}\mathbb{E}\left(U_{Qij}U_{Qil}\delta_{Mjl}\right) = 1, \tag{45}$$

and

$$\mathbb{E}\left(\sum_{k=1}^{n}\mathbf{U}'_{Qi}\mathbf{U}_{Mk}\mathbf{U}'_{Mk}\mathbf{U}_{Qr}\lambda_{Mk}\right) = \sum_{j=1}^{n}\mathbb{E}\left(U_{Qij}U_{Qrj}\delta_{Mjj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}\mathbb{E}\left(U_{Qij}U_{Qrl}\delta_{Mjl}\right) = 0. \tag{46}$$

Therefore, $\mathbb{E}\left(\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q\right) = \mathbf{I}_n$, meaning that we can write $\mathbf{U}'_Q\mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}'_M\mathbf{U}_Q = \mathbf{I}_n + \boldsymbol{\delta}$, such that $\boldsymbol{\delta}$ is a matrix of random errors around $\mathbf{I}_n$, with $\mathbb{E}(\boldsymbol{\delta}) = \mathbf{0}$. Back to Eq. (42), we have now:

$$\boldsymbol{\Lambda}_{QM} = \mathbf{U}'_{QM}\mathbf{U}_Q\left[\frac{q}{q+m}\boldsymbol{\Lambda}_Q + \frac{m}{q+m}(\mathbf{I}_n + \boldsymbol{\delta})\right]\mathbf{U}'_Q\mathbf{U}_{QM}, \tag{47}$$

Therefore the elements in the diagonal of $\boldsymbol{\Lambda}_{QM}$ obey the following identity,

$$\lambda_{QMi} = \frac{q}{q+m}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk} + \frac{m}{q+m}(1+\delta_i), \tag{48}$$

with $\delta i = \sum_{j=1}^{n}U^2_{Qij}\delta_{Mjj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{Qij}U_{Qil}\delta_{Mjl}$, from Eq. (43), such that $\mathbb{E}(\delta_i) = 0$, from Eq. (45). Finally, isolating $\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk}$ in Eq. (48):

$$\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk} = \lambda_{QMi} + \frac{m}{q}\left(\lambda_{QMi} - 1 - \delta_i\right). \tag{49}$$

E.2 Markers only

The genomic model containing the markers only is simpler than the genomic model containing the QTL and markers. Applying the result from Eq. (38), and because in these scenarios we only considered completely unrelated individuals (implying that $a_{jl} = 0$), we have:

$$\sum_{k=1}^{n}\left(\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk} = 1 + \sum_{j=1}^{n}U^2_{Mij}\delta_{Qjj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{Mij}U_{Mil}\delta_{Qjl} = 1 + \delta_i, \tag{50}$$

with $\delta_i = \sum_{j=1}^{n}U^2_{Mij}\delta_{Qjj} + \sum_{j=1}^{n}\sum_{l\neq j}U_{Mij}U_{Mil}\delta_{Qjl}$. Because the markers are in complete LE with the QTL, and remembering that $\mathbb{E}(\delta_{Qjj}) = \mathbb{E}(\delta_{Qjl}) = 0$, we have:

$$\mathbb{E}(\delta_i) = \sum_{j=1}^{n}\mathbb{E}\left(U^2_{Mij}\delta_{Qjj}\right) + \sum_{j=1}^{n}\sum_{l\neq j}\mathbb{E}\left(U_{Mij}U_{Mil}\delta_{Qjl}\right) = 0. \tag{51}$$

Appendix 6: QTL and markers in linkage disequilibrium

F.1 QTL + markers

We will use the following identities relating $\mathbf{G}_{QM}$, $\mathbf{G}_{Q}$ and $\mathbf{G}_{M}$ to understand the properties of the REML estimators for the scenarios we simulated. Using the eigen-decompositions, we have:

$$\begin{aligned}\mathbf{G}_{QM} &= \frac{q}{q+m}\mathbf{G}_{Q} + \frac{m}{q+m}\mathbf{G}_{M} \\ &= \left(1 - \frac{m}{q+m}\right)\mathbf{U}_{Q}\boldsymbol{\Lambda}_{Q}\mathbf{U}'_{Q} + \frac{m}{q+m}\mathbf{U}_{M}\boldsymbol{\Lambda}_{M}\mathbf{U}'_{M}.\end{aligned} \tag{52}$$

Now, because $\mathbf{G}_{QM} = \mathbf{U}_{QM}\boldsymbol{\Lambda}_{QM}\mathbf{U}'_{QM}$, using that identity in Eq. (52) we have:

$$\begin{aligned}\mathbf{U}_{QM}\boldsymbol{\Lambda}_{QM}\mathbf{U}'_{QM} &= \left(1 - \frac{m}{q+m}\right)\mathbf{U}_{Q}\boldsymbol{\Lambda}_{Q}\mathbf{U}'_{Q} \\ &+ \frac{m}{q+m}\mathbf{U}_{M}\boldsymbol{\Lambda}_{M}\mathbf{U}'_{M},\end{aligned} \tag{53}$$

hence,

$$\begin{aligned}\boldsymbol{\Lambda}_{QM} &= \left(1 - \frac{m}{q+m}\right)\mathbf{U}'_{QM}\mathbf{U}_{Q}\boldsymbol{\Lambda}_{Q}\mathbf{U}'_{Q}\mathbf{U}_{QM} \\ &+ \frac{m}{q+m}\mathbf{U}'_{QM}\mathbf{U}_{M}\boldsymbol{\Lambda}_{M}\mathbf{U}'_{M}\mathbf{U}_{QM}.\end{aligned} \tag{54}$$

Therefore the elements in the diagonal of $\boldsymbol{\Lambda}_{QM}$ obey the following identity,

$$\begin{aligned}\lambda_{QMi} &= \left(1 - \frac{m}{q+m}\right)\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk} \\ &+ \frac{m}{q+m}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Mk}\right)^{2}\lambda_{Mk}.\end{aligned} \tag{55}$$

Isolating $\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk}$ in Eq. (55) gives us the following:

$$\begin{aligned}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk} &= \lambda_{QMi} \\ &+ \frac{m}{q+m}\left[\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk} - \sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Mk}\right)^{2}\lambda_{Mk}\right].\end{aligned} \tag{56}$$

Applying Eq. (38) to $\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk}$ and $\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Mk}\right)^{2}\lambda_{Mk}$, we have:

$$\begin{aligned}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk} &= 1 + \sum_{j=1}^{n}U^{2}_{QMij}\delta_{Qjj} \\ &+ \sum_{j=1}^{n}\sum_{l\neq j}U_{QMij}U_{QMil}\left(a_{jl} + \delta_{Qjl}\right),\end{aligned} \tag{57}$$

and

$$\begin{aligned}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Mk}\right)^{2}\lambda_{Mk} &= 1 + \sum_{j=1}^{n}U^{2}_{QMij}\delta_{Mjj} \\ &+ \sum_{j=1}^{n}\sum_{l\neq j}U_{QMij}U_{QMil}\left(a_{jl} + \delta_{Mjl}\right).\end{aligned} \tag{58}$$

Finally, using the identities from Eqs. (57) and (58) in Eq. (56):

$$\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk} = \lambda_{QMi} + \delta_{i}, \tag{59}$$

with $\delta_{i} = \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}U_{QMij}U_{QMil}\left(\delta_{Qjl} - \delta_{Mjl}\right)$.

To understand how δ_{i} will affect the relationship between $\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^{2}\lambda_{Qk}$ and λ_{QMi}, it is fundamental to understand $\mathbb{E}(\delta_{i})$.

$$
\begin{aligned}
\mathbb{E}(\delta_i) &= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{E}\left(U_{QMij}U_{QMil}\left[\delta_{Qjl}-\delta_{Mjl}\right]\right)\\
&= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left[\mathbb{E}\left(U_{QMij}U_{QMil}\delta_{Qjj}\right)-\mathbb{E}\left(U_{QMij}U_{QMil}\delta_{Mjj}\right)\right]\\
&= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left[\mathbb{C}ov\left(U_{QMij}U_{QMil},\delta_{Qjj}\right)-\mathbb{C}ov\left(U_{QMij}U_{QMil},\delta_{Mjj}\right)\right]\\
&= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left[\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjj}-a_{jj}\right)-\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Mjj}-a_{jj}\right)\right]\\
&= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left[\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjj}\right)-\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Mjj}\right)\right]\\
&= \frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjj}\right)\\
&\quad -\frac{m}{q+m}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{C}ov\left(U_{QMij}U_{QMil},\left[1+\frac{q}{m}\right]G_{QMjj}-\frac{q}{m}G_{Qjj}\right)\\
&= \frac{m}{q+m}\left(1+\frac{q}{m}\right)\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjj}\right)\\
&\quad -\frac{m}{q+m}\left(1+\frac{q}{m}\right)\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{QMjj}\right)\\
&= \sum_{j=1}^{n}\sum_{l=1}^{n}\left[\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjl}\right)-\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{QMjl}\right)\right]\\
&= \sum_{j=1}^{n}\sum_{l=1}^{n}\left(\sigma_{ijl,Qlj}-\sigma_{ijl,jl}\right)
\end{aligned}
\tag{60}
$$

Thus, $\mathbb{E}(\delta_i)=\sum_{j=1}^{n}\sum_{l=1}^{n}\left(\sigma_{ijl,Qlj}-\sigma_{ijl,jl}\right)$. It is intuitive that $\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{QMjl}\right)\geq\mathbb{C}ov\left(U_{QMij}U_{QMil},G_{Qjl}\right)$, therefore $\mathbb{E}(\delta_i)\leq 0$.

F.2 Markers only

When we use the genomic model containing the markers only, from Eq. (38) it is straightforward that:

$$
\begin{aligned}
\sum_{k=1}^{n}\left(\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk} &= 1+\sum_{j=1}^{n}U_{Mij}^2\delta_{Qjj}\\
&\quad +\sum_{j=1}^{n}\sum_{l\neq j}U_{Mij}U_{Mil}\left(a_{jl}+\delta_{Qjl}\right).
\end{aligned}
\tag{61}
$$

If we randomly pick just a few markers, they will most likely be in low LD with the QTL ($\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\approx 0$ for any $i,k=1,\ldots,n$) and $\sum_{k=1}^{n}\left(\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk}\approx 0$. When the density of marker data increases, $\sum_{k=1}^{n}\left(\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk}$ also increases, but Eq. (61) is not so straightforward to understand analytically. Therefore, we use the identity $\mathbf{G}_{QM}=q(q+m)^{-1}\mathbf{G}_Q+m(q+m)^{-1}\mathbf{G}_M$ to help us understand what happens when we use a genomic model that contains only markers in LD with the QTL in the SNP data. When the number of markers is sufficiently large, $\lim_{m\to\infty}\mathbf{G}_M=\lim_{m\to\infty}\mathbf{G}_{QM}$, and consequently $\lim_{m\to\infty}\sum_{k=1}^{n}\left(\mathbf{U}'_{Mi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk}=\lim_{m\to\infty}\sum_{k=1}^{n}\left(\mathbf{U}'_{QMi}\mathbf{U}_{Qk}\right)^2\lambda_{Qk}$.

This means that when we have a genomic model with markers only, $\mathbf{G}_{\mathrm{M}}$ will be able to explain, at most, the same amount of variability that $\mathbf{G}_{\mathrm{QM}}$ can explain.

Appendix 7: Algorithm used to simulate genotype data

Given a sample size n, a number of markers m and a number of QTL q:

(1) Generate $\boldsymbol{\theta}_{(\mathrm{m})\times 1} = \begin{bmatrix}\boldsymbol{\theta}_{\mathrm{Q}} & \boldsymbol{\theta}_{\mathrm{M}}\end{bmatrix}$ a vector of MAF for all the m markers, such that:

(a) if QTL/marker are common variants: $\theta_{\mathrm{Qj}} \overset{\mathrm{iid}}{\sim} \mathrm{U}(0.05, 0.5) \forall \mathrm{j} = 1, \ldots, \mathrm{q}$, $\theta_{\mathrm{Mj}} \overset{\mathrm{iid}}{\sim} \mathrm{U}(0.05, 0.5) \forall \mathrm{j} = 1, \ldots, \mathrm{m}$;

(b) if QTL/marker are rare variants: $\theta_{\mathrm{Qj}} \overset{\mathrm{iid}}{\sim} \mathrm{U}(0.01, 0.03) \forall \mathrm{j} = 1, \ldots, \mathrm{q}$, $\theta_{\mathrm{Mj}} \overset{\mathrm{iid}}{\sim} \mathrm{U}(0.01, 0.03) \forall \mathrm{j} = 1, \ldots, \mathrm{m}$;

(2) Generate $\mathrm{n} \times (\mathrm{q} + \mathrm{m})$ genotype matrix $\mathbf{Z}$ containing both QTL and markers:

(a) if the SNPs are independent: $\mathrm{Z}_{1\mathrm{j}}, \ldots, \mathrm{Z}_{\mathrm{nj}} \overset{\mathrm{iid}}{\sim} \mathrm{B}(2, \theta_{\mathrm{j}}), \forall \mathrm{j} = 1, \ldots, \mathrm{q} + \mathrm{m}$;

(b) if the SNPs are correlated, use a routine to generate the correlated binomial data (details in next algorithm), such that:

- $(\mathrm{Z}_{1\mathrm{j}}, \ldots, \mathrm{Z}_{\mathrm{nj}}) \mid \boldsymbol{\theta} \overset{\mathrm{iid}}{\sim} \mathrm{B}(2, \theta_{\mathrm{j}})$;
- $\mathbb{C}\mathrm{or}(\mathrm{Z}_{\mathrm{ij}}, \mathrm{Z}_{\mathrm{il}}) = \rho_{\mathrm{jl}} \forall \mathrm{i} = 1, \ldots, \mathrm{n}$;

(3) Obtain the standardized genotype matrix $\mathbf{W}$,

(a) $\hat{\theta}_{\mathrm{j}} = (1/2\mathrm{n}) \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Z}_{\mathrm{ij}} \forall \mathrm{j} = 1, \ldots, \mathrm{q} + \mathrm{m}$,

(b) $\mathrm{w}_{\mathrm{ij}} = \frac{\mathrm{z}_{\mathrm{ij}} - 2\hat{\theta}_{\mathrm{j}}}{\sqrt{2\hat{\theta}_{\mathrm{j}}(1-\hat{\theta}_{\mathrm{j}})}} \forall \mathrm{i} = 1, \ldots, \mathrm{n}, \mathrm{j} = 1, \ldots, \mathrm{q} + \mathrm{m}$,

(4) Use the group of q simulated QTL to create $\mathbf{W}_{\mathrm{Q}}$,

(5) Use the group of m simulated markers to create $\mathbf{W}_{\mathrm{M}}$.

Appendix 8: Algorithm used to simulate correlated SNP data

In order to simulate correlated binary data, we have looked into approaches such as the Bahadur's representation [25], and the algorithms of [26–28] and [29]. Finally, we have opted for the use of a method that involves the multivariate normal distribution, as proposed by [30].

(1) Given the vector $\boldsymbol{\theta}_{(\mathrm{q}+\mathrm{m})\times 1}$ of MAF, calculate $\mu_{\mathrm{j}} = \Phi^{-1}(\theta_{\mathrm{j}}) \forall \mathrm{j} = 1, \ldots, \mathrm{q} + \mathrm{m}$,

(2) For a matrix $\boldsymbol{\Xi} = [\xi_{\mathrm{kj}}]_{\mathrm{k,j}=1}^{\mathrm{q}+\mathrm{m}}$, simulate $\mathbf{X}^{(1)}_{\mathrm{n}\times(\mathrm{p}+\mathrm{m})}, \mathbf{X}^{(2)}_{\mathrm{n}\times(\mathrm{q}+\mathrm{m})} \overset{\mathrm{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Xi})$,

(3) $\forall \mathrm{i} = 1, \ldots, \mathrm{n}, \mathrm{j} = 1, \ldots, \mathrm{q} + \mathrm{m}$:

(a) $\mathrm{Z}^{(1)}_{\mathrm{ij}} = \mathbb{I}\{\mathrm{X}^{(1)}_{\mathrm{ij}} > 0\}$: $\mathbb{P}(\mathrm{X}^{(1)}_{\mathrm{ij}} > 0) = \theta_{\mathrm{j}}$;

(b) $\mathrm{Z}^{(2)}_{\mathrm{ij}} = \mathbb{I}\{\mathrm{X}^{(2)}_{\mathrm{ij}} > 0\}$: $\mathbb{P}(\mathrm{X}^{(2)}_{\mathrm{ij}} > 0) = \theta_{\mathrm{j}}$;

(3) $\mathrm{Z}_{\mathrm{ij}} = \mathrm{Z}^{(1)}_{\mathrm{ij}} + \mathrm{Z}^{(2)}_{\mathrm{ij}}$, such that:

$$\text{(a)} \quad \rho_{\mathrm{kj}} = \mathbb{C}\mathrm{or}(\mathrm{Z}_{\mathrm{ik}}, \mathrm{Z}_{\mathrm{ij}}) = \mathbb{C}\mathrm{or}(\mathrm{Z}^{(1)}_{\mathrm{ik}}, \mathrm{Z}^{(1)}_{\mathrm{ij}}) = \frac{\mathbb{E}\left(\mathrm{Z}^{(1)}_{\mathrm{ik}}\mathrm{Z}^{(1)}_{\mathrm{ij}}\right) - \theta_{\mathrm{k}}\theta_{\mathrm{j}}}{\sqrt{\theta_{\mathrm{k}}(1-\theta_{\mathrm{k}})\theta_{\mathrm{j}}(1-\theta_{\mathrm{j}})}},$$

$\forall \mathrm{k} \neq \mathrm{j}$;

$$\text{(b)} \quad \mathbb{E}\left(\mathrm{Z}^{(1)}_{\mathrm{ik}}\mathrm{Z}^{(1)}_{\mathrm{ij}}\right) = \mathbb{P}\left(\mathrm{X}^{(1)}_{\mathrm{ik}} > 0, \mathrm{X}^{(1)}_{\mathrm{ij}} > 0\right) = \mathbb{P}\left(\mathrm{X}^{(1)}_{\mathrm{ik}} < 0, \mathrm{X}^{(1)}_{\mathrm{ij}} < 0\right) + \theta_{\mathrm{j}} + \theta_{\mathrm{k}} - 1$$

Details about how to relate input matrix $\boldsymbol{\Xi}$ to the output correlation matrix $\boldsymbol{\Sigma}$ between SNPs are in "Appendix 10".

Appendix 9: Algorithm used to simulate phenotype data

Setting $\sigma^2_{\mathrm{b_Q}} = (1/\mathrm{q})\mathrm{h}^2$ and $\sigma^2_{\varepsilon_{\mathrm{Q}}} = 1 - \mathrm{h}^2$:

(1) Generate $\mathbf{b}_{\mathrm{Q}} \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathrm{q}}\sigma^2_{\mathrm{b_Q}})$,

(2) Generate $\boldsymbol{\varepsilon}_{\mathrm{Q}} \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathrm{n}}\sigma^2_{\varepsilon_{\mathrm{Q}}})$,

(3) Define a value to μ,

(4) Calculate $\mathbf{y} = \mathbf{1}\mu + \mathbf{W}_{\mathrm{Q}}\mathbf{b}_{\mathrm{Q}} + \boldsymbol{\varepsilon}_{\mathrm{Q}}$.

Appendix 10: Correlation structure between the SNPs

Our aim was to simulate genotypes according to a particular correlation structure between the SNPs, defined by $\boldsymbol{\Sigma}$. It is very important to note that $\boldsymbol{\Sigma} \neq \boldsymbol{\Xi}$. Thus, we need to describe here how we defined the elements in matrix $\boldsymbol{\Xi}$, so that after the data was simulated, we obtained the desired $\boldsymbol{\Sigma}$. Without loss of generality, we can define straight away that the diagonal elements of $\boldsymbol{\Xi}$ are all 1s.

Keeping in mind that $\theta_{\mathrm{j}} \overset{\mathrm{indep.}}{\sim} \mathrm{U}(\mathrm{a_j}, \mathrm{b_j})$, $\mu_{\mathrm{j}} = \Phi^{-1}(\mathrm{t_j})$ and the joint cumulative normal distribution

$$\mathbb{P}\left(\mathrm{X}^{(1)}_{\mathrm{ik}} < 0, \mathrm{X}^{(1)}_{\mathrm{ij}} < 0\right) = \Phi_2(-\mu_{\mathrm{j}}, -\mu_{\mathrm{k}}, \xi_{\mathrm{kj}}) = \int_{-\infty}^{-\mu_{\mathrm{j}}} \int_{-\infty}^{-\mu_{\mathrm{k}}} \frac{1}{2\pi\sqrt{1-\xi^2_{\mathrm{jk}}}} \exp\left[\frac{-(\mathrm{u}^2 - 2\xi_{\mathrm{kj}}\mathrm{uv} + \mathrm{v}^2)}{2(1-\xi^2_{\mathrm{jk}})}\right] \mathrm{dudv},$$

which has to be evaluated numerically [31], and defining $\boldsymbol{\Sigma} = [\mathbb{E}_{\boldsymbol{\theta}}(\rho_{\mathrm{jk}})]_{\mathrm{k,j}=1}^{\mathrm{p}+\mathrm{m}}$, such that $\mathbb{E}_{\boldsymbol{\theta}}(\rho_{\mathrm{jk}}) = \rho_{|\mathrm{k}-\mathrm{j}|}$, for all $\mathrm{j} \neq \mathrm{k}$:

$$
\begin{aligned}
\rho_{|j-k|} =& \mathbb{E}_{\boldsymbol{\theta}}\left(\mathbb{C}\mathrm{or}(Z_{ik}, Z_{ij})\right) = \int_{a_j}^{b_j}\int_{a_k}^{b_k} \mathbb{C}\mathrm{or}(Z_{ik}, Z_{ij}) f_{\theta_j}(t_j) f_{\theta_k}(t_k) dt_k dt_j \\
=& \int_{a_j}^{b_j}\int_{a_k}^{b_k} \frac{\left[\mathbb{P}(X_{ik}^{(1)} > 0, X_{ij}^{(1)} > 0) - t_k t_j\right]}{\sqrt{t_k(1-t_k)t_j(1-t_j)}} f_{\theta_j}(t_j) f_{\theta_k}(t_k) dt_k dt_j \\
=& \int_{a_j}^{b_j}\int_{a_k}^{b_k} \frac{\left[\Phi_2(-\mu_j, -\mu_k, \xi_{kj}) + t_j + t_k - t_j t_k - 1\right]}{\sqrt{t_k(1-t_k)t_j(1-t_j)}} f_{\theta_j}(t_j) f_{\theta_k}(t_k) dt_k dt_j \\
=& \int_{a_j}^{b_j}\int_{a_k}^{b_k} \left[\frac{\Phi_2(-\mu_j, -\mu_k, \xi_{kj})}{\sqrt{t_k(1-t_k)t_j(1-t_j)}} - \sqrt{\frac{(1-t_j)(1-t_k)}{t_j t_k}}\right] f_{\theta_j}(t_j) f_{\theta_k}(t_k) dt_k dt_j \\
=& \frac{1}{(b_j - a_j)(b_k - a_k)} \int_{a_j}^{b_j}\int_{a_k}^{b_k} \frac{\Phi_2(-\mu_j, -\mu_k, \xi_{kj})}{\sqrt{t_k(1-t_k)t_j(1-t_j)}} dt_k dt_j \\
& - \left[\frac{\sqrt{u(1-u)}}{(b_j - a_j)} - \frac{\mathrm{arctg}\left(\sqrt{\frac{1-u}{u}}\right)}{(b_j - a_j)}\right]\Bigg|_{a_j}^{b_j} \times \left[\frac{\sqrt{u(1-u)}}{(b_k - a_k)} - \frac{\mathrm{arctg}\left(\sqrt{\frac{1-u}{u}}\right)}{(b_k - a_k)}\right]\Bigg|_{a_k}^{b_k}.
\end{aligned} \tag{62}
$$

Thus, when we have the value for $\rho_{|k-j|}$ and the MAF's θ_j and θ_k, along with their uniform distributions boundaries (a_j, b_j) and (a_k, b_k), from Eq. (62) we define the function of ξ_{kj},

$$
\begin{aligned}
f(\xi_{kj}) =& -\rho_{|j-k|} + \frac{1}{(b_j - a_j)(b_k - a_k)} \\
& \int_{a_j}^{b_j}\int_{a_k}^{b_k} \frac{\Phi_2(-\mu_j, -\mu_k, \xi_{kj})}{\sqrt{t_k(1-t_k)t_j(1-t_j)}} dt_k dt_j \\
& - \left[\frac{\sqrt{u(1-u)}}{(b_j - a_j)} - \frac{\mathrm{arctg}\left(\sqrt{\frac{1-u}{u}}\right)}{(b_j - a_j)}\right]\Bigg|_{a_j}^{b_j} \\
& \times \left[\frac{\sqrt{u(1-u)}}{(b_k - a_k)} - \frac{\mathrm{arctg}\left(\sqrt{\frac{1-u}{u}}\right)}{(b_k - a_k)}\right]\Bigg|_{a_k}^{b_k}.
\end{aligned} \tag{63}
$$

Finally, each element ξ_{kj} from the correlation matrix $\boldsymbol{\Xi}$, that must be used to simulate $\mathbf{X}_{n\times(q+m)}^{(1)}$, $\mathbf{X}_{n\times(q+m)}^{(2)} \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Xi})$, is the root of $f(\xi_{kj})$, which can be obtained by using numerical methods.

Publisher's Note

Received: 19 December 2017 Accepted: 10 July 2018
Published online: 06 August 2018

References

1. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 2010;42:565–9.
2. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
3. Golan D, Rosset S. Accurate estimation of heritability in genome wide studies using random effects models. Bioinformatics. 2011;27:i317–23.
4. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91:1011–21.
5. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971;58:545–54.
6. Jiang J. REML estimation: asymptotic behavior and related topics. Ann Stat. 1996;24:255–86.
7. Hill WG, Weir BS. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb.). 2011;93:47–64.
8. Kumar SK, Feldman MW, Rehkopf DH, Tuljapurkar S. Limitations of GCTA as a solution to the missing heritability problem. Proc Natl Acad Sci USA. 2016;113:E61-70.
9. Yang J, Lee H, Wray NR, Goddard ME, Vissher PM. Commentary on: Limitations of GCTA as a solution to the missing heritability problem. bioRXiv eprint. 2016. https://doi.org/10.1101/036574.
10. Kumar SK, Feldman MW, Rehkopf DH, Tuljapurkar S. Response to commentary on: Limitations of GCTA as a solution to the missing heritability problem. bioRXiv eprint. 2016. https://doi.org/10.1101/039594.
11. de los Campos G, Sorensen D. A commentary on pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14:894.
12. de los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015;11:e1005048.
13. Jiang J, Li C, Paul D, Yang C, Zhao H. High-dimensional genome-wide association study and misspecified mixed model analysis. arXiv eprint. 2014. arXiv:1404.2355.

14. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
15. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc. 1977;72:320–38.
16. Jiang J. Linear and generalized linear mixed models and their applications. New York: Springer; 2007.
17. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 2012;8:e1002685.
18. Marčenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. Math USSR-Sbornik. 1967;1:457–83.
19. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015;47:1114–20.
20. Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P. Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in drosophila melanogaster. Genetics. 2016;203:1871–83.
21. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24:1550–7.
22. Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for Bayes B type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009;41:2.
23. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.
24. Speed D, Cai N, The UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49:986–92.
25. Bahadur RR. A representation of the joint distribution of responses to ***n*** dichotomous items. In: Solomon H, editor. Studies in item analysis and prediction. Stanford: Stanford University Press; 1961. p. 158–68.
26. Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variables. Am Stat. 1991;45:302–4.
27. Lee AJ. Generating random binary deviates having fixed marginal distributions and specified degrees of association. Am Stat. 1993;47:209–15.
28. Gange SJ. Generating multivariate categorical variates using the iterative proportional fitting algorithm. Am Stat. 1995;49:134–8.
29. Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. Am Stat. 1996;50:306–10.
30. Leisch F, Weingessel A, Hornik K. On the generation of correlated artificial binary data. Working Papers SFB "Adaptive Information Systems and Modelling in Economics and Management Science" 13. SFB Adaptive Information Systems and Modelling in Economics and Management Science. WU Vienna University of Economics and Business, Vienna. 1998.
31. Meyer C. Recursive numerical evaluation of the cumulative bivariate normal distribution. J Stat Softw. 2013;52:1–14.