From Estimation to Prediction of Genomic Variances: Allowing for Linkage Disequilibrium and Unbiasedness

Nicholas Schreck[*,1] **and Martin Schlather**[*,†]

[*]Chair of Stochastics and Its Applications, University of Mannheim, A5, 6, 68159 Mannheim, Germany, [†]Animal Breeding and Genetics Group, Center for Integrated Breeding Research, University of Goettingen, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany

ABSTRACT The additive genomic variance, the chief ingredient for the heritability, is often underestimated in phenotype-genotype regression models. Various remedies, including different models and estimators, have been proposed in order to improve on what has been coined the "missing heritability". Recently, debates have been conducted whether estimators for the genomic variance include linkage disequilibrium (LD) and how to explicitly account for LD in estimation procedures.

Up-to-now, the genomic variance in random effect models (REM) has been estimated as a parameter of the marginal, i.e. unconditional model. We propose that the genomic variance in REM should be predicted as a conditional random quantity based on the conditional distribution of β. This signifies a paradigm shift from the estimation to the prediction of the genomic variance. This approach is structurally in perfect accordance to the Bayesian regression model (BRM), where the posterior expectation of the genomic variance is estimated based on the posterior of β. We introduce a novel, mathematically rigorously founded predictor for the conditional genomic variance in (g)BLUP, which is structurally close to the Bayesian estimator. The conditioning of the novel predictor on the data is intrinsically tied to the inclusion of the contribution of LD and the predicted effects. In addition to that, the predictor shows much weaker dependence on distribution assumptions than estimators of other approaches, e.g. GCTA-GREML. Last but not least, the predictor, contrasted with the estimator in the unconditional model, enables an innovative approximation of the influence of LD on the genomic variance in the dataset.

An exemplary simulation study based on the commonly used dataset of 1814 mice genotyped for 10346 polymorphic markers substantiates that the bias of the novel predictor is small in all standard situations, i.e. that the predictor for the conditional genomic variance remarkably reduces the "missing heritability".

KEYWORDS unbiased estimation; unbiased prediction;genomic variances; BLUP; whole-genome regression

Introduction

The additive (genotypic) variance is defined as the variance of the breeding value (BV) and is the chief cause of resemblance between relatives and therefore the most important determinant of the response of a population to selection (Falconer and Mackay 1996). In addition to that, the additive variance can be estimated from observations made on the population and is a principal component of the (narrow-sense) heritability, which is one of the main objectives in many genetic studies (Falconer and Mackay 1996). The heritability is eminent, amongst other things, for the prediction of the response to selection in the breeder's equation (Piepho and Moehring 2007; Hill 2010). Although non-additive genomic variation exists, most of the genetic variation is additive, such that it is sufficient to investigate the additive genetic variance (Hill *et al.* 2008). In more detail, epistasis is only important on the gene-level but not for genetic variances (Hill *et al.* 2008), and Zhu *et al.* (2015) show that for human complex traits dominance variation contributes little. Nevertheless, linkage disequilibrium (LD) is an important factor especially when departing from random mating and Hardy-Weinberg equilibrium, which is often the case in animal breeding (Hill *et al.* 2008; Dempfle 2018).

The genomic variance, the genomic equivalent to the genetic variance, is defined as the variance of a trait that can be explained by a linear regression on a set of markers (de los Campos *et al.* 2015). Many authors have been chasing what is sometimes coined "missing heritability" (Maher 2008) which means that only a fraction of the "true" genomic variance can be captured by regression on influential loci. To begin with, researchers have used genome-wide association studies (GWAS) in order to find quantitative trait loci by using fixed effect regression combined with variable selection. After having added the estimated corresponding genomic variances of the single statistically significant loci, they asserted that they could only account for a fraction of the "true" genetic variance. For instance, Maher (2008) found that only 5% instead of the widely accepted heritability estimate of 80% of human height could be explained. Golan *et al.* (2014) state that the "true" genetic variance is generally underestimated when applying variable selection, e.g. GWAS, to genomic datasets which are typically characterized by their high-dimensionality, where the number of variables p is much larger than the number of observations n. It is well known that a lot of traits are influenced by many genes and that at least some loci with tiny effects are missed when using variable selection or even single-marker regression models. Consequently, Yang *et al.* (2010) decided to fit all common markers jointly using genomic best linear unbiased prediction (GBLUP), where they assume the effect vector to vary at random. Then, they estimated the genomic variance using restricted maximum likelihood (REML) in an approach that they termed genome-wide complex trait analysis (GCTA) GREML (Yang *et al.* 2011). They showed that quantifying the combined effect of all single-nucleotide polymorphisms (SNPs) explains a larger part of the heritability than only using certain variants quantified by GWAS methods. They illustrate their results on the dataset on human height by pointing out that they could explain a heritability of about 45%. They concluded that the main reason for the remaining missing heritability was incomplete linkage disequilibrium of causal variants with the genotyped SNPs, which refers to the general difference of genetic and genomic variances (de los Campos *et al.* 2015).

Recently, there has been a general discussion whether estimators for the genomic variance account for linkage disequilibrium (LD) between markers, which is defined as the covariance between additive effects of marker pairs (Bulmer 1971). Some authors argue that estimators similar to GCTA-GREML lack the contribution of LD (de los Campos *et al.* 2015; Kumar *et al.* 2015, 2016; Lehermeier *et al.* 2017) whereas others (Yang *et al.* 2016) resolutely disagree. More specifically, Kumar *et al.* (2015, 2016) state that in GCTA-GREML the contributions of the p markers to the phenotypic value are assumed to be independent normally distributed random variables with equal variances. Thus, they claim that the random contribution made by each marker is not correlated with the random contributions made by any other marker which leads to a negligence of the contribution of LD to the genomic variance. Moreover, Kumar *et al.* (2015, 2016) criticize GCTA-GREML because of the assumption that the estimated genomic relationship matrix (GRM) is treated as a fixed quantity without sampling errors although the GRM is actually a realization of an underlying stochastic process. In a study on the model plant Arabidopsis (The 1001 Genomes Consortium 2016), Lehermeier *et al.* (2017) use Bayesian ridge regression (BRR) to relate the phenotype flowering time to the genomic data. They use an estimator (termed M2) based on the posterior distribution of the marker effects obtained by Markov Chain Monte Carlo (MCMC) methods and show that this estimator explains a larger proportion of the phenotypic variance than the estimator, termed M1, based on GBLUP (VanRaden 2008; Yang *et al.* 2010, 2011). Lehermeier *et al.* (2017) show that the reason for the better performance of the Bayesian estimator for the genomic variance is the explicit inclusion of disequilibrium covariances.

In this article we investigate the additive genomic variance in linear regression models within the framework of quantitative genetics, i.e. the genetic variance stems from the variation of QTL genotypes whereas the effects of alleles on a trait are fixed parameters (Falconer and Mackay 1996; de los Campos *et al.* 2015). The difference of individuals in their genetics values is caused by the inter-individual differences in allele content at QTL (de los Campos *et al.* 2015).

These assumptions are reflected in the fixed effect model (FEM) that we treat in Section Fixed Effect Model (FEM), where β is a deterministic parameter and the genomic variability comes in only through the randomness of the marker content. We show that the genomic variance in FEM explicitly includes the contribution of LD and we derive a nearly unbiased estimator for the genomic variance in this model, i.e. that the remaining bias consists only of possible correlations between the plug-in quantities. However, the FEM is not applicable to genomic datasets that are characterized by their high-dimensionality. As a remedy, Bayesian regression models (BRM) and random effect models (REM) are often used. In this models, tough, the effect vector β is defined as a random variable and therefore these models do not ly within the framework of classical quantitative genetics. We investigate the expression for the genomic variance in FEM, BRM and REM and notice that, in general, the genomic variance strongly depends on the assumptions for the effect vector. In BRM in Section Bayesian Regression Model (BRM), β is assigned a prior distribution and we seek its posterior distribution by means of the likelihood of the data. We show that in this model set-up it is necessary to consider the genomic variance as a random quantity and not as a fixed population parameter. This results in the estimation of the posterior expectation of the (random) genomic variance, which has already been hinted at in Lehermeier *et al.* (2017).

In Section Random Effect Model (REM) we show that up-to-now, the genomic variance in REM has been estimated as a parameter of the marginal, i.e. unconditional model (e.g. GCTA-GREML). By strictly conditioning on the effect vector as in BRM, we constitute a paradigm shift from the estimation of the marginal genomic variance to the prediction of the random conditional genomic variance, which is structurally in perfect accordance to the posterior genomic variance in BRM. Inspired by the prediction of random effects (or in equivalent terminology: the estimation of the realized values of random effects) introduced by Henderson (1984) at the beginning of his chapter on prediction of random variables, we call our procedure the prediction of the genomic variance in REM.

To this end, we introduce a mathematically founded nearly unbiased predictor for the genomic variance that is adapted to the specified model assumptions. The application of the conditional genomic variance explicitly allows for an adaptation of the genomic variance to the data which is caused by the radical change in the structure of the conditional variance-covariance of β compared to the structure of marginal variance-covariance of β (from a diagonal structure to an arbitrary structure). By doing so, we take on the above mentioned critique that GCTA-GREML neglects the contribution of LD due to the diagonal covariance structure of the marginal β (Kumar *et al.* 2015, 2016). We show that the conditional genomic variance explicitly accounts for LD which has special practical relevance in animal breeding (Dempfle 2018), and remarkably reduces the missing heritability. In addition to that, the difference of the novel predictor and the estimator of the marginal genomic variance in REM can be used as an indicator for the contribution of LD to the genomic variance. In general, the conditional genomic variance in REM is structurally similar to the genomic variance in FEM and therefore has an interpretation close to the classical genetic variance from quantitative genetics.

For reasons of clarity, we provide all calculations and detailed derivations in the Appendix. We illustrate our results for ordinary least-squares (OLS) from the class of FEM, for BRR from the class of BRM and for (G)BLUP from the class REM on simulated datasets, where we borrow the covariance structure from the commonly used dataset on 1814 mice that comes with the R-package "BGLR" (Perez and de los Campos 2014).

Linear Models and the Genomic Variance

We consider the basic additive linear model

$$Y = \mu + X\beta + \varepsilon, \tag{1}$$

where Y is the phenotype of a random individual, μ is a deterministic intercept and β is a p-vector of marker effects. The random allele content at the markers is coded by the random row-p-vector X with expectation $\mathbb{E}[X] = 0$ (in Subsection Notes on the mean-centering of X in the Appendix we consider deviations from this assumption) and covariance matrix Σ_X. The residual ε is assumed to be independent of $X\beta$ and normally distributed with mean 0 and variance σ_ε^2. The additive genomic variance V is then defined as the variance of the genomic value $X\beta$ which consists of the inter-individual differences in allele content at the markers as well as the effects of the markers themselves (de los Campos *et al.* 2015):

$$V := \mathrm{Var}(X\beta). \tag{2}$$

Due to independence of $X\beta$ and ε we can separate the phenotypic variance σ_Y^2 in the genomic variance V and in the residual variance σ_ε^2:

$$\sigma_Y^2 = V + \sigma_\varepsilon^2. \tag{3}$$

Typically, one considers n realizations of model (1), i.e. $y_i = Y|(X = x_{i1}, x_{i2}, ..., x_{ip})$ for $i = 1, ..., n$, which results in the conditional (on X) model

$$y_i = \mu + (\mathbf{X}\beta)_i + \varepsilon_i := \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, ..., n, \tag{4}$$

where $\mathbf{X}$ denotes the $n \times p$ design matrix containing n realizations of the stochastic p-vector X. We consider mean-centered data: $\sum_{i=1}^{n} x_{ij} = 0$ for $j = 1, ..., p$ (in Subsection Notes on the mean-centering of X in the Appendix we consider deviations from this assumption). An unbiased estimator for Σ_X is given by the method-of-moments estimator:

$$\hat{\Sigma}_X := \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}. \tag{5}$$

Fitting linear models to data is based on model (4) and many authors (Piepho and Moehring 2007; Yang *et al.* 2010, 2011; Piepho *et al.* 2012; Janss *et al.* 2012; Lee and Chow 2014; Lehermeier *et al.* 2017) also build their studies on the investigation and estimation of the genomic variance on model (4).

We base our analysis for the (theoretical) expression of the genomic variance on (the theoretical) model (1) for several reasons. First of all, model (1) describes the underlying data-generating process which induces the realized model (4). By defining the genomic variance V given by (2) in model (1), we tackle the criticism of using the realized GRM without accounting for estimation uncertainty issued by Kumar *et al.* (2015, 2016). More importantly, treating X as a random variable represents the interpretations from quantitative genetics that the genetic variability is caused by variation in QTL content. Strikingly, the genomic variance V in the fixed effect model in Section Fixed Effect Model (FEM) constantly equals 0 when building on model (4).

Fixed Effect Model (FEM)

In quantitative genetics the uncertainty about genetic values is assumed to stem from the uncertainty in allele content at the markers, whereas the effects are population parameters and therefore possess no variance (Falconer and Mackay 1996; Gianola *et al.* 2009; de los Campos *et al.* 2015). In accordance to that, we consider β here as a p-vector of fixed effects, i.e. as a deterministic population parameter. Consequently, we calculate the genomic variance V defined in (2) as

$$\begin{aligned} V_\mathrm{f} = \mathrm{Var}(X\beta) &= \beta^\top\Sigma_X\beta \\ &= \sum_{j=1}^{p}\beta_j^2\mathrm{Var}(X_j) + \sum_{i=1}^{p}\sum_{\substack{j=1\\ j\neq i}}^{p}\beta_i\beta_j\mathrm{Cov}(X_i, X_j), \end{aligned} \tag{6}$$

where the expression $\beta^\top\Sigma_X\beta$ is the genomic equivalent of the genetic variance defined in quantitative genetics textbooks for multiple QTL (Lehermeier *et al.* 2017). We have split this term up into the additive locus-specific variance (also called genic variance) and the contribution of linkage disequilibrium between different loci. We notice that the genomic variance V_f is a weighted sum of the variances of the single marker content and the covariance between the content of the markers, where the weights are given by the elements of the fixed population effect vector β.

Replacing β and Σ_X in (6) by unbiased estimators $\hat{\beta}$ and $\hat{\Sigma}_X$ leads to the plug-in estimator

$$\hat{V}_\mathrm{f}^\mathrm{bias} = \hat{\beta}^\top\hat{\Sigma}_X\hat{\beta} \tag{7}$$

for the genomic variance V_f (6). The estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ (7) contains second order products of the random variables $\hat{\beta}_j$, $j = 1, ..., p$, and is therefore a biased estimator for (6). We correct for the covariance of the estimator $\hat{\beta}$ by defining

$$\hat{V}_\mathrm{f} := \hat{\beta}^\top\hat{\Sigma}_X\hat{\beta} - \mathrm{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right) \tag{8}$$

as a less biased estimator for V_f (6), where $\hat{\Sigma}_{\hat{\beta}}$ denotes an unbiased estimator for the variance-covariance matrix $\Sigma_{\hat{\beta}} := \mathrm{Cov}\left(\hat{\beta}\right)$ of $\hat{\beta}$.

In the case that we have more observations (individuals) n than variables (markers) p we can fit the linear model (4) using ordinary least-squares (OLS), for instance. We make the note that if we would base the definition of the genomic variance V given by (2) on the genomic value in model (4), we would obtain $V_\mathrm{f} = \mathrm{Var}(\mathbf{X}\beta) \equiv 0$. As an outcome of the OLS application we obtain the estimated effect vector $\hat{\beta}$ and its estimated covariance matrix $\hat{\Sigma}_{\hat{\beta}}$. Plugging these quantities into $\hat{V}_\mathrm{f}$ (8) we obtain an improved estimator for the genomic variance V_f (6) in OLS and notice that the empirical phenotypic sample variance $\hat{\sigma}_y^2$ splits up as in (3) into the unbiased estimator $\hat{V}_\mathrm{f}$ (8) for the genomic variance in FEM and the unbiased estimator for the residual variance $\hat{\sigma}_\varepsilon^2$:

$$\hat{\sigma}_y^2 = \hat{V}_\mathrm{f} + \hat{\sigma}_\varepsilon^2. \tag{9}$$

This implies that with mean-centered data we can expect the improved estimator for the genomic variance and the estimator for the residual variance to sum up exactly to the phenotypic variance regardless of the data considered. This implies that when using the OLS method to fit a linear model, using the less biased estimator $\hat{V}_\mathrm{f}$ (8) to estimate the genomic variance contribution of all markers is equivalent to simply subtracting the residual variance from the phenotypic variance. We refer to Section FEM in the Appendix for a detailed mathematical derivation of the results in this section.

Bayesian Regression Model (BRM)

Due to the paper of Meuwissen *et al.* (2001) the usage of Bayesian methods has strongly increased in quantitative genetics. The high-dimensionality of genomic data necessitates some way of regularization. The basic idea of adjustment in Bayesian regression models is to express uncertainty of the effect vector β by assigning it a prior distribution. Then, by adapting to the data by means of its likelihood, one attains the posterior distribution of the effect vector.

In the first place, we consider the linear model (1) again where β possesses the prior distribution $p(\beta)$ with prior expectation μ_β (often chosen as 0) and prior variance-covariance matrix Σ_β. The specific form of the distribution of β is not relevant for the following analysis. The genomic variance V given by (2) equals

$$V_\mathrm{b} = \mathrm{Var}(X\beta) = \mu_\beta^\top \Sigma_X \mu_\beta + \mathrm{tr}(\Sigma_X \Sigma_\beta). \tag{10}$$

This expression for the genomic variance is meaningless because we can arbitrarily strongly influence it by the choice of the prior expectation and prior variance-covariance matrix. Instead, we require the genomic variance in BRM to move away from the prior assumptions by adapting to the data. In order to enable this Bayesian learning, we consider the variance of the genomic value $X\beta$ conditional on the effect vector β:

$$W := \mathrm{Var}(X\beta|\beta) = \beta^\top \Sigma_X \beta, \tag{11}$$

which is a quadratic form in the effect vector β. By assigning β a prior distribution, the genomic variance W (11) is assigned a prior distribution with prior expectation $\mathbb{E}[W] = V_\mathrm{b}$.

In the conditional (on X) linear model (4) investigations in BRM are performed on the posterior distribution of β by adapting to the phenotypic data y. We use characteristics of the posterior distribution $p(\beta|y)$ of β to infer the posterior distribution of the genomic variance W given by (11), or equivalently the posterior distribution of the quadratic form W of β. We define the posterior mean of the genomic variance W as

$$W_\mathrm{b} := \mathbb{E}[W|y] = \mathrm{tr}\Big(\Sigma_X \mathbb{E}\left[\beta\beta^\top|y\right]\Big) = \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \mathrm{tr}(\Sigma_X \Sigma_{\beta|y}) \tag{12}$$

and notice that it comprises the posterior expectation $\mu_{\beta|y} := \mathbb{E}[\beta|y]$ and the posterior variance-covariance matrix $\Sigma_{\beta|y} := \mathrm{Var}(\beta|y)$ of β. The expression W_b structurally resembles the prior expectation V_b of W but includes the posterior mean as well as the posterior covariance of β instead of the prior moments. Structurally, the expressions W and W_b resemble the genomic variance V_f given by (6) in the FEM in Section Fixed Effect Model (FEM) and explicitly include the contribution of LD, where the role of the weights for the covariance terms of X $\big($formerly played by $\beta_i\beta_j$, $i \neq j$, in V_f, see equation (6)$\big)$is taken over by the off-diagonal elements of the matrix of the posterior second moments $\mathbb{E}[\beta\beta^\top|y]$ of β. Hence, W_b can be split up in the genic variance and a part including the contribution of LD similar to V_f in (6).

There are many different approaches to fit the conditional model (4) in BRM that mainly differ in the choice of the prior distribution for the effect vector β. Most of the time, the posterior distribution of β is approximated using MCMC methods. Then, it is possible to estimate characteristics of the (posterior) effect vector by the mean value or the empirical variance of the resulting Markov chain. In this context, we denote the sequence of the Markov chain of the estimated effects, after discarding the burn-in iterations and after thinning the chain, by the sequence of p-vectors $\big(\hat{\beta}^{(m)}\big)_{m=1,\ldots,M}$. These vectors are draws from the distribution $p(\beta|y)$. Consequently, we express the quantities $\mu_{\beta|y}$ and $\Sigma_{\beta|y}$ by their empirical counterparts, namely the estimated posterior mean (also often just termed the estimated effects) $\hat{\mu}_{\beta|y}$ of β

$$\hat{\mu}_{\beta|y} = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}^{(m)} \tag{13}$$

and the estimated posterior covariance

$$\begin{aligned}\hat{\Sigma}_{\beta|y} = &\frac{1}{M-1}\sum_{m=1}^{M} \hat{\beta}^{(m)}\left(\hat{\beta}^{(m)}\right)^\top \\ &- \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{m=1}^{M} \hat{\beta}^{(k)}\left(\hat{\beta}^{(m)}\right)^\top.\end{aligned} \tag{14}$$

We propose to plug (13) and (14) into the estimator

$$\widehat{W}_\mathrm{b} = \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\hat{\mu}_{\beta|y}}) + \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\beta|y}) \tag{15}$$

for the mean of the posterior genomic variance W_b (12) in BRM, where $\hat{\Sigma}_{\hat{\mu}_{\beta|y}}$ denotes an unbiased estimator for the covariance $\Sigma_{\hat{\mu}_{\beta|y}}$ of the estimated effects $\hat{\mu}_{\beta|y}$.

A detailed mathematical derivation of the results in this section has been provided in Subsection BRM in the Appendix, where we also show that plugging (13) and (14) into (12) is equivalent to using

$$\widehat{W}_\mathrm{Post} := \frac{1}{M}\sum_{m=1}^{M} \left(\hat{\beta}^{(m)}\right)^\top \hat{\Sigma}_X \hat{\beta}^{(m)}, \tag{16}$$

which can explicitly be interpreted as an estimator for the posterior mean of the genomic variance W in (11), in which the empirical mean is taken over the realizations in every MCMC sample. The estimator $\widehat{W}_\mathrm{Post}$ is called M2 in Lehermeier *et al.* (2017) and it has already been mentioned that this approach draws inferences on the posterior distribution of the genomic variance, whereas the estimator M1 mentioned in Lehermeier *et al.* (2017) equals $\mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\beta|y})$. By strictly deriving $\widehat{W}_\mathrm{b}$, see (15), as well as its corresponding theoretical expression W_b in (12) we have also derived a relation of the estimators used in Lehermeier *et al.* (2017):

$$\mathrm{M2} = \widehat{W}_\mathrm{Post} \approx \widehat{W}_\mathrm{b} \approx \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} + \mathrm{M1}, \tag{17}$$

because $\hat{\Sigma}_{\hat{\mu}_{\beta|y}} \approx 0$.

Random Effect Model (REM)

Gianola *et al.* (2009) shows on toy examples how the genomic variance changes when treating the effect vector as random instead of treating it as a fixed quantity. In random effect models we assume that the effect vector β in model (1) is a normally distributed random variable with mean 0 and diagonal variance-covariance matrix with equal variances σ_β^2, which is equivalent to modeling the single p components of β as independent random variables $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2), j = 1, ..., p$.

We obtain the marginal genomic variance V in model (1) as

$$V_\mathrm{r} = \mathrm{Var}(X\beta) = \mathrm{tr}(\Sigma_X \Sigma_\beta) = \sigma_\beta^2 \sum_{j=1}^{p} \mathrm{Var}(X_j). \tag{18}$$

The effect vector in the linear model (4) in REM is treated as random. Consequently, it is not possible to estimate β but we have to predict β in the terminology introduced by Henderson (1984), which means that the realized values of the random effects are estimated. Common approaches to find a best linear unbiased predictor (BLUP) for β are based on the mixed model equations (Henderson 1984) or in general on maximum-likelihood approaches (Searle *et al.* 1992). The variance components σ_ε^2 and σ_β^2 in model (4) are usually estimated using restricted maximum likelihood (REML) (Patterson and Thompson 1971).

Subsequently, the marginal genomic variance V_r (18) can be estimated by

$$\hat{V}_\mathrm{r} = \hat{\sigma}_\beta^2 \mathrm{tr}(\hat{\Sigma}_X) = \frac{1}{n-1} \hat{\sigma}_\beta^2 \mathrm{tr}\left(\mathbf{X}^\top \mathbf{X}\right). \tag{19}$$

The derivation of the genomic variance in REM is often entirely based on the realized model (4) (Piepho and Moehring 2007; Yang *et al.* 2010, 2011; Piepho *et al.* 2012) which results in the genomic variance-covariance matrix

$$\mathrm{Cov}(\mathbf{X}\beta) = \mathbf{X}\mathbf{X}^\top \sigma_\beta^2. \tag{20}$$

The (realized) genomic variance is estimated as the mean value of the n genomic variances on the diagonal of (20):

$$\hat{V}_\mathrm{r}^\mathrm{real} = \frac{1}{n} \mathrm{tr}\left(\widehat{\mathrm{Cov}}(\mathbf{X}\beta)\right) = \frac{1}{n} \mathrm{tr}\left(\mathbf{X}\mathbf{X}^\top\right) \hat{\sigma}_\beta^2. \tag{21}$$

Due to the properties of the trace, $\hat{V}_\mathrm{r}^\mathrm{real}$ is (approximately) equivalent to the estimated genomic variance $\hat{V}_\mathrm{r}$ in (19) that is based on the marginal variance V_r, see (18) in model (1). Thus, there is no difference in considering the stochastic model (1) or the realized model (4) for the expression of the marginal genomic variance in REM.

Due to computational advantages, it is common (Piepho and Moehring 2007; VanRaden 2008; Piepho 2009; Yang *et al.* 2010, 2011; Speed *et al.* 2012; Janss *et al.* 2012; Lee and Chow 2014; Fernando *et al.* 2017) to consider the so-called linear equivalent model (Henderson 1984)

$$y \overset{\mathrm{d}}{=} \mu + g + \varepsilon, \tag{22}$$

to model (4), i.e. y in (22) is equally distributed as y in model (4). In the linear model (22) the n-vector of genomic values

$$g \overset{\mathrm{d}}{=} \mathbf{X}\beta \sim \mathcal{N}\left(0, \sigma_\beta^2 \mathbf{X}\mathbf{X}^\top\right) \tag{23}$$

is called breeding-value (BV) (VanRaden 2008; Hill 2010) and describes the expected performance of a progeny. The covariance matrix $\sigma_\beta^2 \mathbf{X}\mathbf{X}^\top$ of g can be replaced by some sort of equivalent genomic relationship matrix (GRM) $\mathbf{G}$ (VanRaden 2008) which is the reason why a model fit in REM based on model (22) is called genomic BLUP (GBLUP). For high-dimensional data where $p \gg n$, it is computationally more efficient to investigate the n-vector g and its $n \times n$ covariance matrix than the p-vector β and its corresponding $p \times p$ covariance matrix. Basically,

$$\sigma_\beta^2 \mathbf{X}\mathbf{X}^\top = \frac{1}{p} \mathbf{X}\mathbf{X}^\top (p\sigma_\beta^2) =: \mathbf{G}\sigma_g^2, \tag{24}$$

where $\sigma_\mathrm{g}^2 := p\sigma_\beta^2$ and $\mathbf{G} := \frac{1}{p}\mathbf{X}\mathbf{X}^\top$. The estimated equivalent genomic variance $\hat{V}_\mathrm{r}^\mathrm{equi}$ in the equivalent model (22) equals

$$\hat{V}_\mathrm{r}^\mathrm{equi} = \frac{1}{n} \mathrm{tr}(\mathbf{G}) \hat{\sigma}_g^2, \tag{25}$$

where additionally the mean trace of the GRM $\mathbf{G}$ is often standardized to equal 1. Consequently, the marginal genomic variance $\hat{V}_\mathrm{r}^\mathrm{equi}$ in the equivalent model (22) equals $\hat{\sigma}_g^2$. This approach is termed GCTA-GREML (Yang *et al.* 2010, 2011). The estimated genomic variance $\hat{\sigma}_g^2$ is equivalent to the estimated genomic variance $\hat{V}_\mathrm{r}^\mathrm{real}$ (21) and therefore also in accordance with the marginal genomic variance V_r given by (18).

No matter which of the equivalent approaches $\hat{V}_\mathrm{r}$ (19), $\hat{V}_\mathrm{r}^\mathrm{real}$ (21) or $\hat{V}_\mathrm{r}^\mathrm{equi}$ (25) to estimate the marginal genomic variance V_r (18) is used, they are similar to the first part of expression V_f (6), namely $\sum_{j=1}^{p} \beta_j^2 \mathrm{Var}(X_j)$. But instead of weighting the variances of the allele content by different components of the (fixed) effect vector β, the weights in V_r, see (18), equal the variance component σ_β^2 for every locus. More strikingly, the covariances between the different loci take no part in V_r in (18) but they do so in V_f in (6). Nevertheless, it is not clear how strong the disequilibrium covariances are involved in the estimation of $\hat{\sigma}_\beta^2$ or $\hat{\sigma}_g^2$ in the REML equations and implicitly influence the estimates $\hat{V}_\mathrm{r}$ (19), $\hat{V}_\mathrm{r}^\mathrm{real}$ (21) and $\hat{V}_\mathrm{r}^\mathrm{equi}$ (25). The assumptions on the marginal distribution of β (especially on its covariance structure) are very influential and cause the marginal genomic variance V_r in (18) to be unsatisfactory. This is similar the genomic variance V_b in (10) in Section Bayesian Regression Model (BRM) that is arbitrarily strongly influenced by the prior moments of β.

As a consequence, we consider the genomic variance V (2) conditional on the effect vector β, analogously to Section Bayesian Regression Model (BRM):

$$W := \mathrm{Var}(X\beta|\beta) = \beta^\top \Sigma_X \beta = \mathrm{tr}(\Sigma_X \beta\beta^\top), \tag{11}$$

which is a quadratic form in the normally distributed effect vector β. The genomic variance W is a random variable with $\mathbb{E}[W] = V_\mathrm{r}$. Investigations on the random variable W have to be done similar to investigations on the random effect β in REM, namely by a strict conditioning on the phenotypic data y in accordance to the prediction (Henderson 1984) of the effect vector β, where the BLUP $\mu_{\beta|y} := \mathbb{E}[\beta|y]$ of β is given by the conditional expectation of β on y (Searle *et al.* 1992). Consequently, we define an unbiased predictor for the random genomic variance W in (11) as the expectation of the random variable W conditional on the data y:

$$W_\mathrm{r} := \mathbb{E}[W|y] = \mathrm{tr}(\Sigma_X \mathbb{E}[\beta\beta^\top|y]) = \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \mathrm{tr}(\Sigma_X \Sigma_{\beta|y}), \tag{26}$$

where we have used the conditional variance-covariance matrix $\Sigma_{\beta|y} := \mathrm{Var}(\beta|y)$ of β additional to the BLUP $\mu_{\beta|y}$.

The predictor W_r for the genomic variance W is structurally in perfect accordance with the posterior genomic variance W_b in (12) and consequently has the same interpretation as V_f, see (6), similar to the genetic variance in quantitative genetics. Most importantly, opposed to the marginal genomic variance V_r in (18), the predicted genomic variance W_r includes the contribution of disequilibrium covariances similar to V_f in (6). This is done by weighting the covariances of X with the off-diagonals of the matrix of the conditional second moment $\mathbb{E}[\beta\beta^\top|y]$ of β. Hence, W_r can be split up in the genic variance and a part including the contribution of LD similar to V_f in (6).

We examine the covariance of $\beta|y$ in model (4) more closely and obtain:

$$\Sigma_{\beta|y} = \sigma_\beta^2 \mathbb{1}_{p\times p} - \sigma_\beta^2 \mathbf{X}^\top(\sigma_\beta^2 \mathbf{X}\mathbf{X}^\top + \sigma_\varepsilon^2 \mathbb{1}_{n\times n})^{-1}\mathbf{X}\sigma_\beta^2. \qquad (27)$$

The marginal covariance structure $\sigma_\beta^2 \mathbb{1}_{p\times p}$ of β (components independent with equal variances) in V_r in (18) changes drastically when considering the conditional (on y) covariance structure $\Sigma_{\beta|y}$ of β. In this conditional approach, the single components of $\beta|y$ are not equally and independently distributed, but posses an arbitrary covariance structure by adapting to the data by means of the likelihood of the data similar to the posterior covariance $\Sigma_{\beta|y}$ in Section Bayesian Regression Model (BRM). By introducing the concept of the prediction of conditional genomic variance, we tackle one of the central points of critique of GCTA-GREML issued by Kumar *et al.* (2015, 2016).

We substitute the variance components implicitly included in W_r in (26) by estimates and obtain a nearly unbiased predictor for the conditional genomic variance :

$$\widehat{W}_\mathrm{r} = \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} + \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\beta|y}), \qquad (28)$$

or

$$\widehat{W}_\mathrm{r}^{\mathrm{equi}} := \frac{1}{n-1}\hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} + \frac{1}{n-1}\mathrm{tr}(\hat{\Sigma}_{g|y}), \qquad (29)$$

when the prediction procedure is based on the equivalent model (22).

By considering the genomic variance in REM as random and deriving a predictor for this random variable, we also bridge the gap between the estimation of the posterior mean of the genomic variance in BRM and estimation of the marginal variance in REM that has been observed in Lehermeier *et al.* (2017).

For a detailed mathematical derivation of the results in this chapter we refer to Subsection REM in the Appendix. For an extension of the REM to the mixed-effect model (MEM) we refer to Subsection MEM in the Appendix.

Statistical Analysis

In this section we compare the performance of the estimators $\hat{V}_\mathrm{f}^{\mathrm{bias}}$ in (7) and $\hat{V}_\mathrm{f}$ in (8) from Section Fixed Effect Model (FEM), the performance of the estimator $\widehat{W}_\mathrm{b}$ in (15) from Section Bayesian Regression Model (BRM) and the performance of the estimator $\hat{V}_\mathrm{r}$ in (19)as well as the predictor $\widehat{W}_\mathrm{r}$ in (28) from Section Random Effect Model (REM) with respect to the genomic variance V_f defined in (6). We have already mentioned in Section Fixed Effect Model (FEM) that the genomic variance V_f is the genomic equivalent of the genetic variance as defined in quantitative genetics for multiple QTL and explicitly accounts for the contribution of LD.

We executed all calculations in this section with the free software R (R Development Core Team 2008).

Preparation of Datasets

We considered the mice dataset that comes with the *R*-package BGLR (Perez and de los Campos 2014). The data originally stem from an experiment from Valdar *et al.* (2006a,b) in a mice population. The dataset contains $p = 10346$ polymorphic markers that were measured in $n = 1814$ mice. The trait under consideration was body mass index (BMI). In order to compare the estimators from the FEM with the ones from BRM and REM we created a second dataset (the reduced mice dataset) where we included only the first $\tilde{p} = 0.6n \approx 1088$ markers, such that $\tilde{p} < n$ holds true.

We used the $n \times p$ ($n \times \tilde{p}$) matrix $\mathbf{X}$ coding the marker content from the mice (reduced mice) dataset to obtain a realistic LD-structure for the further analysis. In order to obtain modified datasets with different QTL-to-marker densities we assigned k out of the p ($\tilde{p}$) markers to be QTL. We attributed each designated QTL with a corresponding "true" (fixed) effect k-vector β_k. Then, we calculated the "true" genomic variance V_k as

$$V_k = \beta^\top \Sigma_{X_k} \beta, \qquad (30)$$

using formula (6) for V_f from Section Fixed Effect Model (FEM) because it resembles the genetic variance as defined in quantitative genetics. We denote by X_k the restriction of the marker content data to the (designated) QTL content. For each k, we calculated the covariance matrix Σ_{X_k} applying the method of moments estimator (5) to the QTL content data $\mathbf{X}_k$ with all individuals (serves as the whole population). It has been claimed that the main source of missing heritability is imperfect LD between markers and QTL (Yang *et al.* 2010) which we exclude by explicitly assigning markers to be QTL. In addition to that, the genomic variance under consideration is purely additive and the variance-covariance matrix of the QTL content is given. Consequently, the performance of the estimators and of the predictor depends only on their ability to represent the genomic variance V_k for all k.

In order to investigate the estimation (prediction) procedures for each k for different levels of heritabilities

$$h_k^2 := \frac{V_k}{V_k + \sigma_\varepsilon^2} \in \{0.2, 0.5, 0.8\},$$

we set the true error variance σ_ε^2 equal to 1 and multiplied the "true" effect vector β_k by the constant c_k, where

$$c_k^2 := \frac{h_k^2}{1 - h_k^2}\frac{\sigma_\varepsilon^2}{V_k}.$$

This results in considering "true" genomic variances of $V_k \in \{0.25, 1, 4\}$ for each QTL-marker ratio k/p ($k/\tilde{p}$). We drew n realizations of ε from a normal distribution with mean 0 and standard deviation $\sigma_\varepsilon = 1$, calculated the phenotypic values y_k using the additive linear model (4), and hence obtained several modified genomic datasets with phenotypic and genotypic values for each V_k and h_k^2).

Model-fitting and Genomic Variance Calculation

Given the phenotypic and genotypic data described in Subsection Preparation of Datasets, we fitted the OLS model using the R-function "lm" and obtained the estimated effect vector

$\hat{\beta}$ as well as the estimated error variances $\hat{\sigma}_{\varepsilon}^2$. We used these in order to calculate the biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ in (7) as well as the nearly unbiased estimator $\hat{V}_{\mathrm{f}}$ in (8). The OLS method is not well-defined in applications where the number of markers p is larger than the number of individuals n. Therefore we applied this method only to the reduced mice dataset.

We fitted the BRR model with the function "BGLR" with the specification of the model equal to "BRR" in the R-package "BGLR" (Perez and de los Campos 2014). We decided to use 30000 iterations of the Markov chain and discarded the first 10000 as burn-in, after we had exemplarily checked the convergence of the resulting Markov chain and asserted convergence in every case. We kept only every fifth realization of the remaining chain in order to obtain approximate independence. This left us with $M = 4000$ state values that are assumed to be representative of the posterior distribution. As a result of the application we obtained estimators $\hat{\mu}$ for the intercept, $\hat{\sigma}_{\varepsilon}^2$ for the residual variance, and a $M \times p$ ($M \times \tilde{p}$) matrix with realizations of the estimated effect vector $\hat{\beta}^{(m)}$, $m = 1, ..., M$, in every state m of the considered Markov chain. We plugged these into $\widehat{W}_{\mathrm{post}}$, see (16), in order to calculate an estimator for the posterior expectation of the genomic variance W_{b} defined in (12).

We fitted the (G)BLUP model in its equivalent form (22) as in Section Random Effect Model (REM) by using the R-package "sommer" (Covarrubias-Pazaran 2017) and in particular its function "mmer". We obtained the predicted effects $\hat{\mu}_{g|y}$ and the estimated variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_{\varepsilon}^2$. We used these quantities in order to calculate the estimator $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ in (25) for V_{r} and the predictor $\widehat{W}_{\mathrm{r}}^{\mathrm{equi}}$ in (29) for the conditional genomic variance W_{r}. Despite the explicit implementation of $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ and $\widehat{W}_{\mathrm{r}}^{\mathrm{equi}}$ we use the equivalent quantities $\hat{V}_{\mathrm{r}}$, see (19), and $\widehat{W}_{\mathrm{r}}$, see (28), to describe the simulation studies in order to emphasize the derivation using the stochastic data-generating process X.

Performance Indexes

We compared each estimator $\hat{V}$ for the genomic variance V_k in (30) with respect to the absolute value of the relative bias

$$\mathrm{rBias}(\hat{V}) := \frac{|\mathbb{E}[\hat{V}] - V_k|}{V_k} \tag{31}$$

and their relative root-mean-squared-error

$$\mathrm{rRMSE}(\hat{V}) := \sqrt{\frac{\mathbb{E}\left[(\hat{V} - V_k)^2\right]}{V_k}}. \tag{32}$$

For the analysis in Subsection Variation of QTL-Allocations. we define the relative contribution rLD of LD to the genomic variance V_k as

$$\mathrm{rLD}(V_k) := \frac{\sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} \beta_i^{(k)} \beta_j^{(k)} \mathrm{Cov}(X_i^{(k)}, X_j^{(k)})}{V_k} \tag{33}$$

and the indicator I_{r} in REM for the contribution of LD to the genomic variance as

$$I_{\mathrm{r}} := \frac{\widehat{W}_{\mathrm{r}} - \hat{V}_{\mathrm{r}}}{\widehat{W}_{\mathrm{r}}}. \tag{34}$$

Variation of Observational Data

We randomly selected k QTL as described in Subsection Preparation of Datasets and fixed them for the further analysis, where we chose the number k from the sets $K_{\mathrm{m}} := \{10, 100, 500, 1000, 2000, 5000, 10000\}$ for the mice dataset and $K_{\mathrm{rm}} := \{10, 50, 100, 200, 500, 1000\}$ for the reduced mice dataset. For practical reasons of creating effect vectors with shapes of realizations of normal distributions or the heavier-tailed gamma distribution, we chose the "true" effect vector β_k as a realization (i.e. fixed value) according to the distributions depicted in Table 1. Formally, we considered an unknown data-generating process X with n realized p-vectors contained in the design matrices $\mathbf{X}$. We randomly selected $\tilde{n} = 0.8n$ out of the n realizations (individuals) 500 times for each combination of k and h^2 which imitates drawing from the data-generating process X. In each iteration, we calculated the estimators and the predictor in the OLS, BRR and (g)BLUP models as described in Subsection Model-fitting and Genomic Variance Calculation.

The estimation performance of the biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ compared to the improved estimator $\hat{V}_{\mathrm{f}}$ from FEM in the reduced mice dataset is depicted in Figure 1 for a heritability of 0.2 ($V_k = 0.25$ for all k). The biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ performs drastically worse than the improved estimator $\hat{V}_{\mathrm{f}}$. This behavior of $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ is very similar for all considered h^2 which emphasizes the importance of the bias-correction in the FEM. For reasons of clarity we abstain from depicting the estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ in the further analysis.

We compared the performance of the remaining estimators and the predictor for the genomic variances in the reduced mice dataset for $h^2 = 0.2$ in Figure 2, for $h^2 = 0.5$ in Figure 3 and for $h^2 = 0.8$ in Figure 4. The estimated variances are averaged over the 500 realizations and are depicted in subject to the number of QTL k which also determines the QTL-marker ratios $k/\tilde{p}$.

The bias-corrected estimator $\hat{V}_{\mathrm{f}}$ given by (8) performs best and is very close to the "true" value of the genomic variance for all levels of heritabilities h^2 and numbers of QTL k. This is expected because the "true" genomic variance is calculated according to the genomic variance in the FEM, the genomic equivalent of the genetic variance as defined in quantitative genetics.

The estimator $\widehat{W}_{\mathrm{b}}$, see(15), from the BRM overestimates the "true" genomic variance for $h^2 = 0.2$ for about over 10%. The performance of the estimator improves with larger heritability and for $h^2 = 0.8$ the estimator is very close to the "true" value for all k. Possible reasons for the overestimation by $\widehat{W}_{\mathrm{b}}$ for small h^2 are dependencies between the states of the Markov chain, such that the approximation (40) is not good enough and that the model-fit gets worse such that the plugged-in state values are not representative of the posterior distribution (although the MCMC algorithm had converged).

The estimation performance of $\hat{V}_{\mathrm{r}}$ given by (19) depends on the QTL-marker ratio such that with increasing number of QTL's k the underestimation of $\hat{V}_{\mathrm{r}}$ drastically increases, whereas for a small QTL-marker ratio, the estimator $\hat{V}_{\mathrm{r}}$ tends to overestimate the "true" genomic variance. The performance of the estimator strongly declines with increasing heritability, such that for $h^2 = 0.2$ the relative bias amounts to about 4%, for $h^2 = 0.5$ to $5 - 15\%$ and for $h^2 = 0.8$ to $5\% - 20\%$.

The novel predictor $\widehat{W}_{\mathrm{r}}$ defined in (28) from the REM overestimates the "true" genomic variance for $h^2 = 0.2$ but nevertheless performs better than the estimators from the REM and the BRM. The predictor $\widehat{W}_{\mathrm{r}}$ performs relatively independent of the QTL-marker ratio and its performance advantage upon $\hat{V}_{\mathrm{r}}$ increases with increasing h^2. Although the "true" genomic variance is calculated according to the FEM, the performance of $\widehat{W}_\mathrm{r}$ can more than compete with the estimators $\hat{V}_\mathrm{f}$ from FEM and $\widehat{W}_\mathrm{b}$ from the BRM. We put special emphasis on the performance improvement of the novel predictor $\widehat{W}_\mathrm{r}$ versus the estimator $\hat{V}_\mathrm{r}$ in the case of higher heritability (Figure 4). This resembles the study of the missing heritability (Maher 2008; Yang *et al.* 2010) and the novel predictor remarkably reduced the missing heritability in REM in our simulation study. The number of covariances that contribute to the genomic variance V_k depends quadratically on k $(k^2 - k)$ and we draw the conclusion that the increasing bias of $\hat{V}_\mathrm{r}$ in (19) with increasing k is due to the quadratic increase in the number of missed covariances. In contrast to that, the estimators $\hat{V}_\mathrm{f}$ in (8), $\widehat{W}_\mathrm{b}$ in (15) and the predictor $\widehat{W}_\mathrm{r}$ in (28) fluctuate around the "true" value of the genomic variance independent on the number of covariances.

The performance of the estimators and the predictor from BRM and REM in the full mice dataset is very similar to the performance in the reduced mice dataset such that we can also assert the improved performance of $\widehat{W}_\mathrm{b}$ and $\widehat{W}_\mathrm{r}$ in the case of $p \gg n$. In addition to that, we compared the estimators and the predictor with respect to relative root-mean-squared-error and assert similar behavior as when investigating the estimation performance. We conclude that treating the genomic variance as random is also advantageous with respect to the precision of the estimators and the predictor.

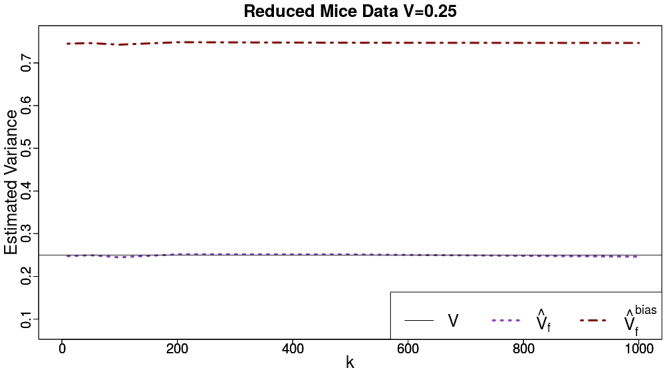

Figure 1 Estimated variance in the FEM (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V_k independently of the QTL-to-marker ratio.

Table 1 **Sources of Effect vector β in Subsection Variation of Observational Data**

K	β
10	$(1, 0.3, -0.5, 5, -2.4, 0.1, -0.6, 1.3, -2, -1.7)^\top$
50	$\mathcal{U}[-2.6, 3]$
100	$\mathcal{G}(0.1, 5)$[a]
200	$\mathcal{N}(0.1, 0.38^2)$
500	$\mathcal{N}(0.2, 1)$
1000	$\mathcal{G}(0.03, 8)$
2000	$\mathcal{N}(0.1, 0.38^2)$
5000	$\mathcal{G}(0.03, 8)$
10000	$\mathcal{N}(0.1, 1)$

[a] Gamma distribution with shape and scale parameters

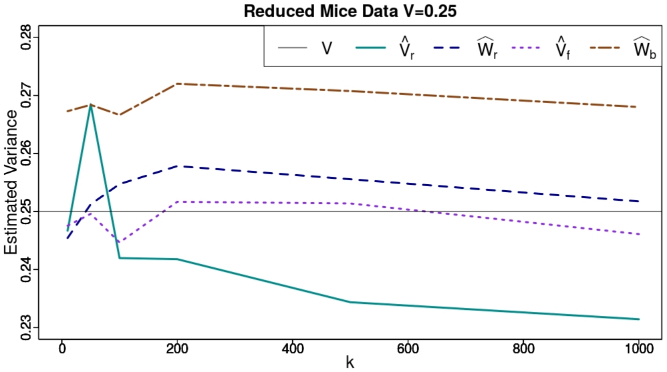

Figure 2 Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ from FEM performs best followed by the predictor $\widehat{W}_\mathrm{r}$ for the conditional genomic variance in REM which slightly overestimates V_k. The estimator $\hat{V}_\mathrm{r}$ underestimates V_k and the bias of the estimators increases with k. The estimator for the posterior genomic variance $\widehat{W}_\mathrm{b}$ constantly overestimates V by around 10%.

Variation of QTL-Allocations

In Subsection Variation of Observational Data we investigated the performance of the estimators and the predictor of the genomic variance for a fixed QTL-allocation and varying observations. Hence, it is possible that the conclusions made depend strongly on the specific QTL-allocation and the corresponding implied LD-structure, and cannot be generalized. Consequently, we considered the whole dataset of individuals and conducted the analysis in this section for different QTL-allocations for each level of heratibility and number of QTL k. In order to do so, we undertook 2000 iterations of randomly choosing the actual QTL-allocations for every level of heritability h^2 and each number of QTL $k \in K$, where $K_\mathrm{m} = \{10, 100, 500, 1000, 2000, ..., 10000\}$ for the mice dataset and $K_\mathrm{rm} = \{10, 50, 100, 200, ..., 1000\}$ for the

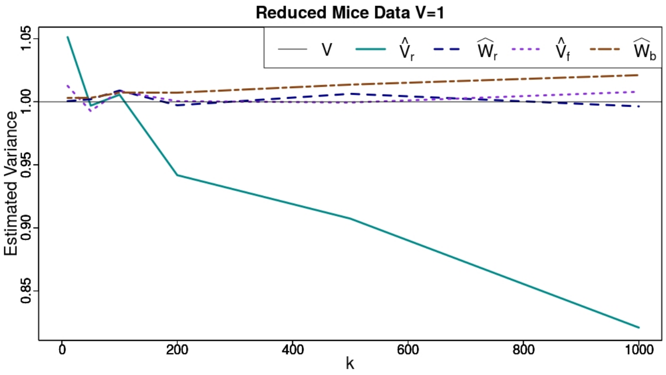

Figure 3 Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_f$ from the FEM and the predictor $\widehat{W}_r$ from the REM are very close to the "true" V_k for all k. The estimator $\widehat{W}_b$ for the posterior mean of genomic variance also performs well but slightly overestimate V_k with increasing k. The estimator $\hat{V}_r$ drastically underestimates V_k and the bias of the estimator strongly increases with k.

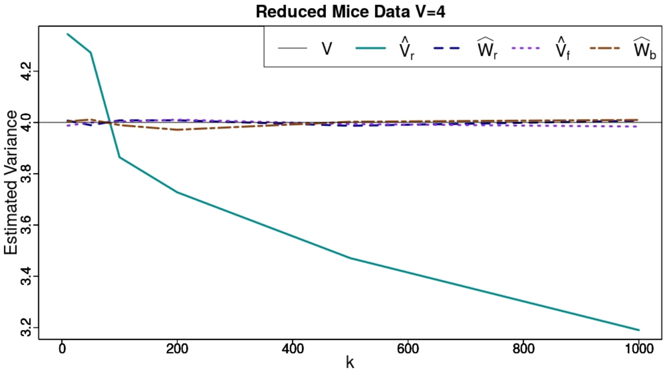

Figure 4 Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_f$ from the FEM, the predictor $\widehat{W}_r$ from the REM and the estimator $\widehat{W}_b$ are constantly very close to V_k for all QTL-marker ratios $k/\tilde{p}$. The estimator $\hat{V}_r$ from the REM drastically underestimates V_k and the bias of the estimator strongly increases with k.

reduced mice dataset. We used $\beta = (1, ..., 1)_k$ as the "true" effect vector in V_k, see (30), prior to scaling by c_k, in order to weight all locus-specific variances as well as all disequilibrium covariances equally.

We compared the performance of the estimator $\hat{V}_f^{\text{bias}}$ in (7) to the improved estimator $\hat{V}_f$ in (8) in the reduced mice dataset in Figure 5. Similar to Subsection Variation of Observational Data we notice that the bias-corrected estimator $\hat{V}_f$ behaves much better than the estimator $\hat{V}_f^{\text{bias}}$. In addition to that, $\hat{V}_f$ fluctuates around the true value of $V = 0.25$ for all k, which indicates that the performance is independent of the QTL-marker ratio. We observed similar behavior for $h^2 = 0.5$ and $h^2 = 0.8$. In Figures 6 ($h^2 = 0.2$), 7 ($h^2 = 0.5$) and 8 ($h^2 = 0.8$) we depict the average of the estimators and the predictor over all considered QTL-allocations in the reduced mice dataset for different number of QTL k for fixed number of markers $\tilde{p} = 1088$. We notice that the behavior of all considered quantities over k is more bumpy compared to the analysis for a fixed QTL-allocation. This indicates that the QTL-allocation influences the estimators and the predictor. The general conduct of the estimator $\hat{V}_f$, see (8), the estimator $\widehat{W}_b$, see (15) and the predictor $\widehat{W}_r$, see (28), is similar and independent of the level of heritability, as we notice that these quantities have spikes and slabs for the same k (same QTL-allocations) for each h^2. This indicates that $\hat{V}_f$, $\widehat{W}_b$, and $\widehat{W}_r$ are in accordance and confirms that they can be used to estimate the genomic variance as defined in the FEM (and quantitative genetics). The estimator $\hat{V}_f$ fluctuates around the "true" value of the genomic variance, wheres the estimator $\widehat{W}_b$ constantly overestimates the V_k for all k for small h^2 (as in Subsection Variation of Observational Data). The predictor $\widehat{W}_r$ fluctuates around the "true" value of the genomic variance for $h^2 = 0.2$ and slightly overestimates for larger heritabilities, but performs at least as good as $\hat{V}_f$ and $\widehat{W}_b$. The estimator $\hat{V}_r$ from REM underestimates the "true" value of the genomic variance in all cases where the bias increases with increasing k regardless of h^2. Compared the behavior in Subsection Variation of Observational Data where only one QTL-allocation was examined, the estimator $\hat{V}_r$ underestimates V also for small k. The difference to the novel predictor $\widehat{W}_r$ is striking. Especially for $h^2 = 0.8$ the estimator $\hat{V}_r$ accounts for less than half of the true genomic variance, which is in accordance with observations of the missing heritability (Maher 2008; Yang *et al.* 2010). The missed covariances increase quadratically in k which explains the increasing bias of the estimator $\hat{V}_r$. This simulation studies indicates that the novel predictor $\widehat{W}_r$ in (28) as well as the estimator $\widehat{W}_b$ in (15) are possible solutions to the missing heritability, because of their explicit inclusion of LD.

For each level of heritability h^2 and each number of QTL k we considered 2000 different QTL-allocations and each of them defines a specific LD-structure. Consequently, the "true" value of the genomic variance for each QTL-allocation can be distinguished by a different relative contribution of LD to V as defined in rLD defined in (33). We depict the empirical covariance of this relative contribution of LD with the value of $\hat{V}_r$, $\widehat{W}_r$, the indicator I_r given by (34), the relative bias (31) of $\hat{V}_r$ and the relative bias of $\widehat{W}_r$ for each h^2 and k in Figure 9. The correlation of $\hat{V}_r$ with the relative contribution of LD is negative (about -0.75) which indicates that the larger the contribution of LD becomes, the smaller the estimator becomes. This is clearly contrasted by the novel predictor $\widehat{W}_r$ which is approximately uncorrelated with the contribution of LD. In addition to that, the relative bias of $\hat{V}_r$ is positively correlated (about 0.75) with the relative contribution of LD which demonstrates that the larger the contribution of LD, the larger the bias of the estimator becomes. This is once again contrasted by the relative bias of $\widehat{W}_\mathrm{r}$ that is approximately uncorrelated to the contribution of LD. Strikingly, the empirical correlation of the indicator I_r, which can be calculated using only $\hat{V}_\mathrm{r}$ and $\widehat{W}_\mathrm{r}$, is positively correlated with the relative contribution of LD to the genomic variance. As a consequence, I_r constitutes a novel approximation of the relative contribution of LD to the genomic variance.

In addition to the analysis for the reduced mice dataset we compared the estimators and the predictor in the full mice dataset where $p \gg n$. The performance of the estimator $\widehat{W}_\mathrm{b}$, $\hat{V}_\mathrm{r}$, and the predictor $\widehat{W}_\mathrm{r}$ are very similar to the performance in the reduced mice dataset.

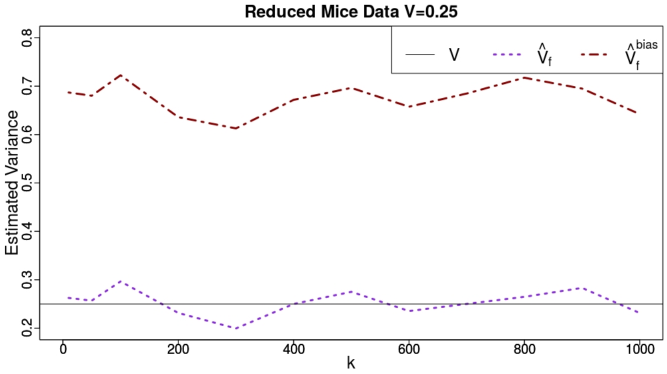

Figure 5 Estimated variance in the FEM (mean value over different QTL allocations) in the reduced mice dataset for different numbesr of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V independently of the QTL-to-marker ratio.

Results

We defined the genomic variance V in (2) as the variance of the genomic value for a stochastic marker content X in Section Linear Models and the Genomic Variance. We noticed that the expression for the genomic variance V as a fixed population parameter strongly depends on the models assumptions on the effect vector β. As a consequence, we distinguished the analysis of the genomic variance between the FEM, the BRM and the REM.

The genomic variance V_f in the FEM in Section Fixed Effect Model (FEM), where the effect vector β is a deterministic parameter, is the genomic equivalent of the genetic variance in quantitative genetics and explicitly includes the contribution of LD. We noticed that the simple plug-in estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$, given by (7), is clearly biased and introduced the improved estimator $\hat{V}_\mathrm{f}$ in (8) that is unbiased for V if the plugged-in estimators are uncorrelated. The FEM can be applied when the number of effects is small or after the application of variable selection or reductions methods. In the simulation studies on the reduced mice dataset in Section Statistical Analysis we showed that the bias-corrected estimator largely improved the estimation of genomic variance V in FEM (Figures 1 and 5).

The genomic variance V_b, given by (10) in the BRM in Section Bayesian Regression Model (BRM), proved to be meaningless

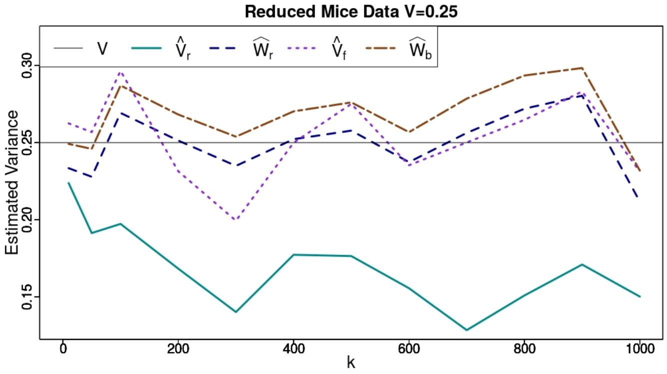

Figure 6 Estimated variance (mean value over different QTL allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ from the FEM performs similar to the predictor $\widehat{W}_\mathrm{r}$ from the REM and they are both close to the true V of 0.25. The estimator $\widehat{W}_\mathrm{b}$ from the BRM performs solidly but constantly sightly overestimates the "true" genomic variance. The estimator $\hat{V}_\mathrm{r}$ from the REM underestimates V by around 40% and the bias of the estimator tends to increase with the number of QTL k.

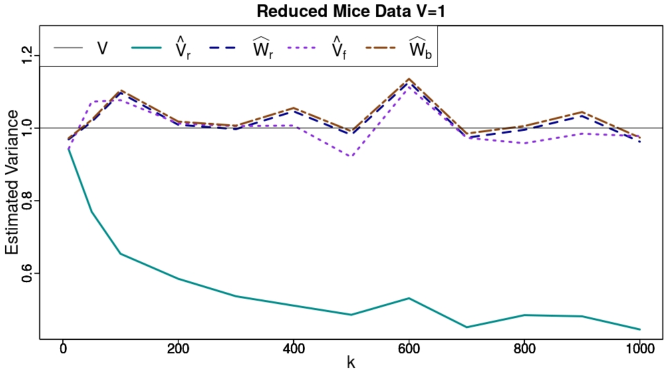

Figure 7 Estimated variance (mean value over different QTL allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_\mathrm{f}$ from the FEM performs similar to the predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM and they are all very close to V. The estimator $\hat{V}_\mathrm{r}$ from the REM underestimates V increasingly with the number of QTL k and by at least 40% starting at a QTL-marker ratio of 10%.

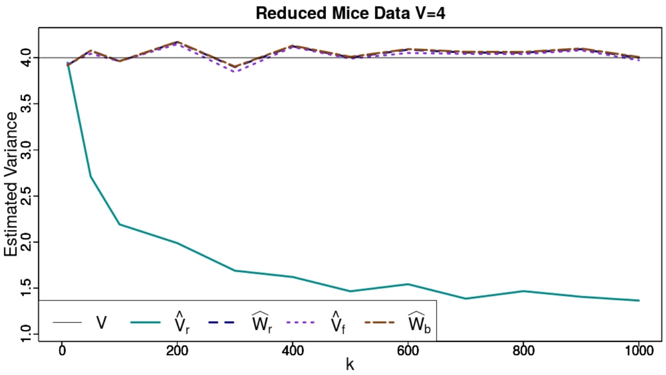

Figure 8 Estimated variance (mean value over different QTL allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V_k equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_f$ from the FEM, the predictor $\widehat{W}_r$ from the REM and the estimator $\widehat{W}_b$ from the BRM are very close to the true V of 4. The estimator $\hat{V}_r$ from the REM drastically underestimates V and the bias of the estimator tends to increase with the number of QTL's k. For a large QTL-marker ratio, $\hat{V}_r$ can only recover about 40% of the true genomic variance.

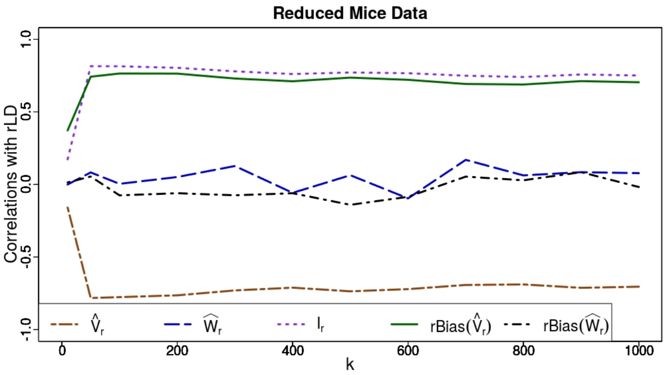

Figure 9 Empirical correlations with the relative contribution of LD to the "true" genomic variance in the reduced mice dataset for different numbers of QTL k for fixed number of markers $\tilde{p} \approx 1088$. Values are averaged over the different levels of $h^2 \in \{0.2, 0.5, 0.8\}$. The correlation of the estimator $\hat{V}_r$ with the relative contribution of LD is about -0.7 except for $k = 10$, whereas the correlation of the predictor $\widehat{W}_\mathbf{r}$ fluctuate around 0. The correlation of the relative bias of $\hat{V}_r$ is about 0.7 except for $k = 10$ which indicates that the larger the contribution of LD to the genomic variance, the larger the bias of $\hat{V}_r$ becomes. Contrary to that, the bias of the predictor $\widehat{W}_\mathbf{r}$ is approximately uncorrelated to the relative contribution of LD. The quantity I_r is positively correlated $(0.7 - 0.8)$ to the relative contribution of LD which makes it an usable indicator for the relative contribution of LD to the genomic variance.

because of its dependence on characteristics of the prior distribution of the effect vector β. In order to include characteristics of the posterior distribution of β, we considered the genomic variance as a random variable W in (11) (conditional on the effect vector β) with prior expectation V_b. We inferred its posterior expectation W_b in (12), which includes the contribution of LD and has an interpretation similar to the genomic variance in the FEM. By doing so, we laid the theoretical foundations for the estimation of the posterior genomic variance in BRM and proved that the estimator for the mean of the posterior genomic variance $\widehat{W}_{\text{Post}}$, given by (16), (Lehermeier *et al.* 2017) is nearly unbiased for the expectation of the random variable W. In Section Statistical Analysis we illustrated that the estimator $\widehat{W}_b$ performs similarly to the estimator $\hat{V}_f$ from the FEM (Figures 4, 7, 8) although it tends to overestimate in cases of low heritability (Figures 2, 3, 6). This enables an estimation of the genomic variance as defined in the FEM for high-dimensional genomic datasets ($p \gg n$).

In Section Random Effect Model (REM) we showed that, up-to-now, the genomic variance in REM has been treated as the parameter V_r, given by (18), and that popular estimation methods (e.g. GCTA-GREML) are based on the marginal covariance matrix of the effect vector β which leads to a negligence of the contribution of LD. In perfect accordance with the BRM, we introduced the novel concept of the random genomic variance W in (11) in REM by conditioning on the effect vector β. Similar to the prediction of the random effect β, the random genomic variance W has to be predicted by a strict conditioning on the phenotypic data by means of its likelihood. We derived a nearly unbiased predictor $\widehat{W}_r$ in (28) for the random genomic variance in REM that is based on the covariance of the conditional distribution of β given the data y. By adapting to the data, this approach explicitly allows for the contribution of LD and remarkably reduces the missing heritability of $\hat{V}_r$ in REM. In Section Statistical Analysis we illustrated that $\widehat{W}_r$ performs drastically superior to the estimator $\hat{V}_r$ (19) (Figures 3, 4, 6, 7, 8). Furthermore, the novel predictor $\widehat{W}_r$ performs at least as good as the estimator used for the mean of the posterior genomic variance in BRM $\widehat{W}_b$ and similar to the estimator $\hat{V}_f$. This enables an estimation of the genomic variance as defined in the FEM in the REM.

In Section Statistical Analysis we compared the estimators and predictors for the genomic variance with respect to their ability to estimate the genomic equivalent of the genetic variance in quantitative genetics in the mice dataset as well as the reduced mice dataset. In Subsection Variation of Observational Data we designated a fixed subset of markers to be QTL and investigated the performance of the estimator and the predictor for varying observations, which simulated drawing from the data-generating process of X. Because the investigation had only been executed for one fixed QTL setting and corresponding fixed population effect vector β, we investigated the dependence of the performance of the estimators and the predictor of the genomic variance on the QTL-allocation in Subsection Variation of QTL-Allocations. We asserted that the performance of the estimators and of the predictor as described above is consistent when varying the number of observations as well as when varying the underlying QTL allocation. In the end, we introduced an innovative indicator I_r in (34) of the contribution of LD on the genomic variance (Figure 9) by comparing the estimator $\hat{V}_r$ and the predictor $\widehat{W}_r$. This comparison added to the conclusion that the improved performance of the novel predictor $\widehat{W}_r$ compared to the estimator $\hat{V}_r$ is caused by the inclusion of LD.

Discussion

The additive genetic variance and the narrow sense heritability are clearly and uniquely defined, but nevertheless estimation procedures give different results (Chen 2016). The estimation of the genomic variance (especially in REM) varies even when using the same marker data to calculate different genomic relationship matrices. (Legarra 2015; Fernando *et al.* 2017). We showed in Subsection Notes on the mean-centering of X in the Appendix that transformations of the input marker-matrix $\mathbf{X}$ change the estimate of the genomic variance when using estimators similar to GCTA-GREML like $\hat{V}_\mathrm{r}$ (19), $\hat{V}_\mathrm{r}^\mathrm{real}$ (21), or $\hat{V}_\mathrm{r}^\mathrm{equi}$ (25). The estimate of the genomic variance depends on the specific form on the estimated GRM and whether one considers mean-centered data or not. This critique goes hand in hand with Kumar *et al.* (2015, 2016) that state that the GRM in GCTA-GREML is an estimate of the underlying data-generating process but is treated as a fixed quantity, which makes the calculation of the genomic variance as in (20) (Yang *et al.* 2010, 2011) invalid. As a solution, we built our analysis on the data-generating process of the marker data by considering X as a random vector in model (1), which is also vital to generate a genomic variance in the FEM where the effects are fixed population parameters. We showed that treating the marker data as random is not enough because of the approximate equivalence of $\hat{V}_\mathrm{r}$ in (19) and $\hat{V}_\mathrm{r}^\mathrm{real}$ in (21) (as well as its equivalent forms). We introduced the random genomic variance W in (11) by explicitly conditioning on the effect vector. As a consequence, the (random) genomic variance depends on the variance-covariance structure of the data-generating process X and is consequently independent of linear transformations of the marker content. In addition to that, the novel predictor $\widehat{W}_\mathrm{r}$, given by (28), for this random quantity is based on the conditional (on the data y) distribution of the effects which implies a departure from the marginal variance-covariance structure of β (diagonal, with equal variances σ_β^2) to the arbitrary variance-covariance structure of β conditional on y. This approach is in accordance with the estimation of the posterior mean of the genomic variance $\widehat{W}_\mathrm{b}$ given by (15) in BRM (Lehermeier *et al.* 2017) and tackles yet another central point of critique on GCTA-GREML issued by Kumar *et al.* (2015, 2016), namely that the single marker effects are treated as independent random variables with equal variances.

In the theoretical expression of the genomic variance V_r in (18) LD does not contribute. However, when using the REML algorithm to estimate the variance component σ_β^2, LD implicitly contributes to estimated variances similar to GCTA-GREML like $\hat{V}_\mathrm{r}$ (19), $\hat{V}_\mathrm{r}^\mathrm{real}$ (21), or $\hat{V}_\mathrm{r}^\mathrm{equi}$ (25). Nevertheless, as we have noticed in Figure 9, the bias of $\hat{V}_\mathrm{r}$ is still very much correlated with with the contribution of LD. Our approach of considering the genomic variance W in BRM and REM as a random variable conditional on the effect vector can be considered as an extension of the genomic variance from the FEM to high-dimensional datasets. This is intrinsically tied to an explicit contribution of LD to the genomic variance. To be more specific, we have noticed in Figure 9 that the bias of $\widehat{W}_\mathrm{r}$ (28) is approximately uncorrelated with the contribution of LD. This led us to deducing the indicator I_r, given (34), for the contribution of LD to the genomic variance.

The estimation and prediction of the effects β in high-dimensional datasets using the BRM and the REM is executed by adapting to the data by means of its likelihood, which possibly results in an over-adjustment. In accordance to that, estimating the posterior mean of the conditional variance in BRM or predicting the conditional genomic variance in REM bears the risk of over-adjustment to the data. In our simulation study in Section Statistical Analysis we have assumed a very simplistic model and excluded, e.g., the influence of imperfect linkage between the markers and the QTL. This removed one of the main sources of the missing heritability claimed in literature, e.g. Yang *et al.* (2010). The stability of the novel predictor $\widehat{W}_\mathrm{r}$ as well as of the estimator of the posterior mean $\widehat{W}_\mathrm{b}$ has still to be further tested in more complex scenarios and for different datasets with different LD-structures. Specifically, it would be of interest to apply the novel predictor $\widehat{W}_\mathrm{r}$ to the dataset of human height (not available to us), which is characterized by a large heritability of 80%, and compare the prediction performance of the genomic variance to the estimation of the genomic variance performed in Yang *et al.* (2010).

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft, DFG, project number SCHL 1865/4-1, as well as the 'RTG 1953 - Statistical Modeling of Complex Systems and Processes'.

We thank Hans-Peter Piepho for his great effort, insight and expertise that greatly assisted the research, although he may not agree with all of the interpretations and conclusions of this paper. We would also like to show our gratitude to Chris-Carolin Schön and Leo Dempfle for important remarks and advice on the paper. We are also grateful to Henner Simianer, Torsten Pook and Jonas Brehmer for their comments on earlier versions of the manuscript.

Literature Cited

Bulmer, M., 1971 The effect of selection on genetic variability. American Naturalist **105**: 201–211.

Chen, G. B., 2016 On the reconciliation of missing heritability for genome-wide association studies. European Journal of Human Genetics **24**: 1810–1816.

Covarrubias-Pazaran, G., 2017 *Solving Mixed Model Equations in R*.

de los Campos, G., D. Sorensen, and D. Gianola, 2015 Genomic heritability: What is it? PLoS Genetics **11**: e1005048.

Dempfle, L., 2018 Personal Communication.

Falconer, D. and T. Mackay, 1996 *Introduction into Quantitative Genetics*. Fourth edition.

Fernando, R., H. Cheng, X. Sun, and D. Garrick, 2017 A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. Journal of Animal Breeding and Genomics **134**: 213–223.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the bayesian alphabet. Genetics **183**: 347–363.

Golan, D., E. S. Lander, and S. Rosset, 2014 Measuring missing heritability: Inferring the contribution of common variants. Proceedings of the National Academy of Sciences **111**: E5272–E5281.

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*.

Hill, W. G., 2010 Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B **365**: 73–85.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genetics **4**.

Janss, L., G. de los Campos, N. Sheehan, and D. Sorensen, 2012 Inferences from genomic models in stratified populations. Genetics **192**: 693–704.

Kumar, S. K., M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar, 2015 Limitations of GCTA as a solution of the missing heritability problem. PNAS pp. E61–E70.

Kumar, S. K., M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar, 2016 Respone to "commentary on limitations of GCTA as a solution to the missing heritability problem". bioRxiv: http://dx.doi.org/10.1101/039594.

Lee, J. J. and C. C. Chow, 2014 Conditions for the validity of SNP-based heritability estimation. Human Genetics **133**: 1011–1022.

Legarra, A., 2015 Comparing estimates of genetic variance across different relationship models. Theoretical Population Biology **107**: 26–30.

Lehermeier, C., G. de los Campos, V. Wimmer, and C.-C. Schön, 2017 Genomic variance estimates: With or without disequlibrium covariances? Journal of Animal Breeding and Genomics **134**: 232–241.

Maher, B., 2008 Personal genomes: The case of the missing heritability. Nature **456**: 18–21.

Meuwissen, T., B. Hayes, and M. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Patterson, H. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Perez, P. and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics **198**: 483–495.

Piepho, H.-P., 2009 Ridge regression and extensions for genomewide selection in maize. Crop Science **49**: 1165–1176.

Piepho, H.-P. and J. Moehring, 2007 Computing heritability and selection response from unbalanced plant breeding trials. Genetics **177**: 1881–1888.

Piepho, H.-P., J. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, *et al.*, 2012 Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Science **52**: 1093–1104.

R Development Core Team, 2008 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. Wiley Interscience.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics **91**: 1011–1021.

The 1001 Genomes Consortium, 2016 1135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell **166**: 481–491.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, *et al.*, 2006a Genome-wide genetic association of comlex traits in heterogeneous stock mice. Nature Genetics **38**: 879–887.

Valdar, W., L. C. Solberg, D. Gauguier, W. O. Cookson, J. N. P. Rawlins, *et al.*, 2006b Genetic and environment effect on complex traits in mice. Genetics **174**: 959–984.

VanRaden, P., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of heritability for human height. National Genetics **42(7)**: 565–569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: A tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Yang, J., S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher, 2016 Commentary on "Limitations of GCTA as a solution to the missing heritability problem". bioRxiv: http://dx.doi.org/10.1101/036574.

Zhu, Z., A. Bakshi, A. A. Vinkhuyzen, G. Hemani, S. H. Lee, *et al.*, 2015 Dominance genetic variation contributes little to the missing heritability for human complex traits. The American Journal of Human Genetics **96**: 377–385.

Appendix

FEM

We derive the results depicted in Section Fixed Effect Model (FEM). We consider β in the stochastic model

$$Y = \mu + X\beta + \varepsilon, \tag{1}$$

as a deterministic p-vector containing the allele effects and express the genomic variance as

$$V_{\mathrm{f}} := \mathrm{Var}(X\beta) = \beta^{\top}\Sigma_X\beta. \tag{6}$$

Replacing β and Σ_X in (6) by unbiased estimators $\hat{\beta}$ and $\hat{\Sigma}_X$ leads to the plug-in estimator

$$\hat{V}_{\mathrm{f}}^{\mathrm{bias}} = \hat{\beta}^{\top}\hat{\Sigma}_X\hat{\beta}. \tag{7}$$

This estimator has expectation:

$$\begin{aligned}
\mathbb{E}\left[\hat{V}_{\mathrm{f}}^{\mathrm{bias}}\right] &= \mathbb{E}\left[\hat{\beta}^{\top}\hat{\Sigma}_X\hat{\beta}\right] = \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\left[\hat{\sigma}_{ij}^{X}\hat{\beta}_i\hat{\beta}_j\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\left[\mathrm{Cov}\left(\hat{\sigma}_{ij}^{X},\hat{\beta}_i\hat{\beta}_j\right) + \mathbb{E}\left[\hat{\sigma}_{ij}^{X}\right]\mathbb{E}\left[\hat{\beta}_i\hat{\beta}_j\right]\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^{X},\hat{\beta}_i\hat{\beta}_j\right) + \sigma_{ij}^{X}\left[\sigma_{ij}^{\hat{\beta}} + \mathbb{E}\left[\hat{\beta}_i\right]\mathbb{E}\left[\hat{\beta}_j\right]\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\sigma_{ij}^{X}\beta_i\beta_j + \sigma_{ij}^{X}\sigma_{ij}^{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^{X},\hat{\beta}_i\hat{\beta}_j\right) \\
&= \beta^{\top}\Sigma_X\beta + \mathrm{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^{X},\hat{\beta}_i\hat{\beta}_j\right),
\end{aligned} \tag{35}$$

where we denote by $\sigma_{ij}^{X} := \mathrm{Cov}(X_i, X_j)$ the covariance between the random variables X_i and X_j and by $\sigma_{ij}^{\hat{\beta}} := \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ the covariance between the random variables $\hat{\beta}_i$ and $\hat{\beta}_j$.
The estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ (7) is biased by the amount

$$\begin{aligned}
\mathrm{Bias}\left(\hat{V}_{\mathrm{f}}^{\mathrm{bias}}\right) &: = \mathbb{E}\left[\hat{V}_{\mathrm{f}}^{\mathrm{bias}}\right] - V_{\mathrm{f}} \\
&\overset{(35),(6)}{=} \mathrm{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^{X},\hat{\beta}_i\hat{\beta}_j\right),
\end{aligned}$$

where only $\operatorname{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right)$ is amenable to estimation. Consequently, we define the bias-corrected estimator $\hat{V}_{\text{f}}$ for the genomic variance V_{f} (6) as

$$\hat{V}_{\text{f}} := \hat{\beta}^\top\hat{\Sigma}_X\hat{\beta} - \operatorname{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right), \tag{8}$$

where $\hat{\Sigma}_{\hat{\beta}}$ is an unbiased estimator for $\Sigma_\beta = \operatorname{Cov}(\hat{\beta})$. We first investigate the bias-correction term $\operatorname{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}})$ and find that

$$\begin{aligned}\mathbb{E}\left[\operatorname{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right)\right] &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\left[\hat{\sigma}_{ij}^X\hat{\sigma}_{ij}^{\hat{\beta}}\right]\\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\sigma}_{ij}^{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\left[\hat{\sigma}_{ij}^X\right]\mathbb{E}\left[\hat{\sigma}_{ij}^{\hat{\beta}}\right]\\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\sigma}_{ij}^{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}^X\sigma_{ij}^{\hat{\beta}}\\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\sigma}_{ij}^{\hat{\beta}}\right) + \operatorname{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right). \end{aligned}\tag{36}$$

We examine the estimator $\hat{V}_{\text{f}}$ (8):

$$\begin{aligned}\mathbb{E}\left[\hat{V}_{\text{f}}\right] &\overset{(8)}{=} \mathbb{E}\left[\hat{\beta}^\top\hat{\Sigma}_X\hat{\beta}\right] - \mathbb{E}\left[\operatorname{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right)\right]\\ &\overset{(35),(36)}{=} \beta^\top\Sigma_X\beta + \operatorname{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\beta}_i\hat{\beta}_j\right)\\ &\quad - \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\sigma}_{ij}^{\hat{\beta}}\right) - \operatorname{tr}\left(\Sigma_X\Sigma_{\hat{\beta}}\right)\\ &= \beta^\top\Sigma_X\beta + \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\beta}_i\hat{\beta}_j - \hat{\sigma}_{ij}^{\hat{\beta}}\right). \end{aligned}\tag{37}$$

The estimator $\hat{V}_{\text{f}}$ (8) is biased by the amount

$$\begin{aligned}\operatorname{Bias}\left(\hat{V}_{\text{f}}\right) :&= \mathbb{E}\left[\hat{V}_{\text{f}}\right] - V_{\text{f}}\\ &\overset{(37),(6)}{=} \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{Cov}\left(\hat{\sigma}_{ij}^X,\hat{\beta}_i\hat{\beta}_j - \hat{\sigma}_{ij}^{\hat{\beta}}\right),\end{aligned}$$

that is caused only by dependencies between the unbiased plug-in estimators $\hat{\Sigma}_X$, $\hat{\beta}$ and $\hat{\Sigma}_{\hat{\beta}}$. If they are pairwise uncorrelated, the estimator V_{f} is unbiased. We call estimators that are biased only due to correlations between plugged-in estimators "nearly unbiased".

If we fit the conditional (on X) linear model

$$y_i = \mu + (\mathbf{X}\beta)_i + \varepsilon_i := \mu + \sum_{j=1}^{p}x_{ij}\beta_j + \varepsilon_i, \quad i = 1,\ldots,n, \tag{4}$$

by OLS, we can express

$$y = \mu + \mathbf{X}\beta + \varepsilon = \hat{\mu} + \mathbf{X}\hat{\beta} + \hat{\varepsilon},$$

where $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y$ and $\hat{\varepsilon} = y - \mathbf{X}\hat{\beta}$.

It holds true that

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma_\varepsilon^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

Subsequently, an unbiased estimator $\hat{\Sigma}_{\hat{\beta}}$ for the variance of $\hat{\beta}$ in OLS is given by:

$$\hat{\Sigma}_{\hat{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\hat{\sigma}_\varepsilon^2, \tag{38}$$

which leads to

$$\begin{aligned}\operatorname{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right) &\overset{(5),(38)}{=} \operatorname{tr}\left(\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\hat{\sigma}_\varepsilon^2\right)\\ &= \frac{1}{n-1}\hat{\sigma}_\varepsilon^2\operatorname{tr}(\mathbb{1}_{p\times p})\\ &= \frac{p}{n-1}\hat{\sigma}_\varepsilon^2.\end{aligned}$$

It is well-known in OLS theory that an unbiased estimator for the residual variance σ_ε^2 in the case of $(p+1)$ variables (including the intercept) is given by

$$\hat{\sigma}_\varepsilon^2 := \frac{1}{n-(p+1)}y^\top(1-\mathbf{H})y,$$

where $\mathbf{H} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the so-called hat-matrix.

For mean-centered phenotypes ($\sum_{i=1}^n y_i/n = 0$) we express the nearly unbiased estimator $\hat{V}_{\text{f}}$ (8) in OLS as

$$\begin{aligned}\hat{V}_{\text{f}} &= \hat{\beta}^\top\hat{\Sigma}_X\hat{\beta} - \operatorname{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\right)\\ &= y^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y - \frac{p}{n-1}\hat{\sigma}_\varepsilon^2\\ &= \frac{1}{n-1}y^\top\mathbf{H}y - \frac{p}{n-1}\frac{1}{n-(p+1)}y^\top(1-\mathbf{H})y\\ &= \frac{1}{n-1}y^\top\mathbf{H}y + \left(\frac{1}{n-1} - \frac{1}{n-(p+1)}\right)y^\top(1-\mathbf{H})y\\ &= \frac{1}{n-1}y^\top y - \hat{\sigma}_\varepsilon^2\\ &= \hat{\sigma}_y^2 - \hat{\sigma}_\varepsilon^2.\end{aligned}$$

We obtain the exact empirical variance decomposition

$$\hat{\sigma}_y^2 = \hat{V}_{\text{f}} + \hat{\sigma}_\varepsilon^2. \tag{9}$$

in the OLS model which resembles the theoretical variance decomposition (3) in model (1).

BRM

In Section Bayesian Regression Model (BRM) we assume β in the stochastic model

$$Y = \mu + X\beta + \varepsilon, \tag{1}$$

to be distributed according to the prior distribution $p(\beta)$ with finite prior expectation $\mu_\beta := \mathbb{E}[\beta]$ and finite prior variance-covariance matrix $\Sigma_\beta := \operatorname{Cov}(\beta)$. We leave the specific form of the distribution $p(\beta)$ unspecified in this general approach. We calculate the genomic variance

$$\begin{aligned}V_{\text{b}} :&= \operatorname{Var}(X\beta)\\ &= \operatorname{Var}_\beta(\mathbb{E}[X\beta|\beta]) + \mathbb{E}_\beta[\operatorname{Var}(X\beta|\beta)]\\ &= \operatorname{Var}_\beta(\mathbb{E}[X]\beta) + \mathbb{E}_\beta[\beta^\top\Sigma_X\beta]\\ &= \mathbb{E}[X]\Sigma_\beta\mathbb{E}[X]^\top + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\left[\sigma_{ij}^X\beta_i\beta_j\right]\\ &\overset{\mathbb{E}[X]=0}{=} \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}^X\left(\sigma_{ij}^\beta + \mathbb{E}[\beta_i]\mathbb{E}[\beta_j]\right)\\ &= \operatorname{tr}(\Sigma_X\Sigma_\beta) + \mu_\beta^\top\Sigma_X\mu_\beta.\end{aligned}\tag{10}$$

We can arbitrarily strongly influence the genomic variance V_{b} (10) by the choice of the prior first and second moment of β. As a solution, we define

$$W := \mathrm{Var}(X\beta|\beta) = \beta^\top \Sigma_X \beta = \mathrm{tr}\left(\Sigma_X \beta\beta^\top\right), \tag{11}$$

as the variance of the genomic value $X\beta$ conditional on β. This random quantity has (prior) expectation

$$\begin{aligned} \mathbb{E}[W] &= \mathbb{E}\left[\mathrm{tr}\left(\Sigma_X\beta\beta^\top\right)\right] \\ &= \mathrm{tr}\left(\Sigma_X \mathbb{E}\left[\beta\beta^\top\right]\right) \\ &= \mathrm{tr}\left(\Sigma_X\left(\mathrm{Cov}(\beta)+\mathbb{E}\left[\beta\right]\mathbb{E}\left[\beta^\top\right]\right)\right)\\ &= \mathrm{tr}(\Sigma_X\Sigma_\beta)+\mu_\beta^\top\Sigma_X\mu_\beta \\ &= V_{\mathrm{b}}.\end{aligned}$$

We define

$$\begin{aligned} W_{\mathrm{b}} :&= \mathbb{E}[W|y] \\ &\overset{(11)}{=} \mathrm{tr}\left(\Sigma_X \mathbb{E}\left[\beta\beta^\top|y\right]\right) \\ &= \mathrm{tr}\left(\Sigma_X\left(\mathbb{E}\left[\beta|y\right]\mathbb{E}[\beta^\top|y] + \mathrm{Cov}(\beta|y)\right)\right) \\ &= \mu_{\beta|y}^\top\Sigma_X\mu_{\beta|y} + \mathrm{tr}\left(\Sigma_X\Sigma_{\beta|y}\right) \end{aligned} \tag{12}$$

as the corresponding posterior mean of the genomic variance W (11), where

$$\mathbb{E}[W_{\mathrm{b}}] = \mathbb{E}\left[\mathbb{E}[W|y]\right] = \mathbb{E}[W] = V_{\mathrm{b}}.$$

In the conditional model

$$y_i = \mu + (\mathbf{X}\beta)_i + \varepsilon_i := \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, ..., n, \tag{4}$$

we adapt to the data and define the estimator $\widehat{W}_{\mathrm{b}}$ for the expectation of the posterior genomic variance W_{b} (12):

$$\widehat{W}_{\mathrm{b}} := \underbrace{\hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \mathrm{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\mu}_{\beta|y}}\right)}_{\widehat{W}_{\mathrm{b}}^{(1)}} + \underbrace{\mathrm{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}\right)}_{\widehat{W}_{\mathrm{b}}^{(2)}}, \tag{15}$$

where we correct for the bias of the purely plug-in estimator by subtracting $\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\mu}_{\beta|y}})$ equivalently to the nearly unbiased estimator $\hat{V}_{\mathrm{f}}$ (8) in Subsection FEM. The first part of expression (15), $\widehat{W}_{\mathrm{b}}^{(1)}$, is similar to $\hat{V}_{\mathrm{f}}$ (8) such that we calculate as in (37):

$$\begin{aligned} \mathbb{E}_{\beta|y}\left[\widehat{W}_{\mathrm{b}}^{(1)}\right] &= \mu_{\beta|y}^\top\Sigma_X\mu_{\beta|y} \\ &\quad + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j - \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}}\right). \end{aligned}$$

We derive the expectation of the second part of expression (15), $\widehat{W}_{\mathrm{b}}^{(2)}$, as

$$\begin{aligned} \mathbb{E}_{\beta|y}\left[\widehat{W}_{\mathrm{b}}^{(2)}\right] &= \mathbb{E}\left[\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y})\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\hat{\sigma}_{ij}^X\hat{\sigma}_{ij}^{\beta|y}\right] \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^X, \hat{\sigma}_{ij}^{\beta|y}\right) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\left[\hat{\sigma}_{ij}^X\right]\mathbb{E}\left[\hat{\sigma}_{ij}^{\beta|y}\right] \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^X, \hat{\sigma}_{ij}^{\beta|y}\right) + \mathrm{tr}(\Sigma_X\Sigma_{\beta|y}). \end{aligned}$$

Combining these results, we find

$$\begin{aligned} \mathbb{E}\left[\widehat{W}_{\mathrm{b}}\right] &= \mathbb{E}\left[\widehat{W}_{\mathrm{b}}^{(1)}\right] + \mathbb{E}\left[\widehat{W}_{\mathrm{b}}^{(2)}\right] \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j - \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}}\right) \\ &\quad + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\left(\hat{\sigma}_{ij}^X, \hat{\sigma}_{ij}^{\beta|y}\right). \end{aligned}$$

The remaining bias of the estimator $\widehat{W}_{\mathrm{b}}$ vanishes if the estimators $\hat{\sigma}_{ij}^{X|y}, \hat{\sigma}_{ij}^{\beta|y}, \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}}$ and $(\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j$ themselves are pairwise uncorrelated for all $i = 1, ..., n$ and $j = 1, ..., p$.

In applications, we obtain the Markov chain sequence of p-vectors $\left(\hat{\beta}^{(m)}\right)_{m=1,...,M}$ after discarding the burn-in iterations and after thinning the chain. We use the empirical mean

$$\hat{\mu}_{\beta|y} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)} \tag{13}$$

as an unbiased estimator for the posterior expectation $\mu_{\beta|y}$ and the empirical variance

$$\begin{aligned} \hat{\Sigma}_{\beta|y} &= \frac{1}{M-1}\sum_{m=1}^{M}\hat{\beta}^{(m)}\left(\hat{\beta}^{(m)}\right)^\top \\ &\quad - \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}\left(\hat{\beta}^{(k)}\right)^\top \end{aligned} \tag{14}$$

as an estimator for the posterior covariance $\Sigma_{\beta|y}$. In order to calculate $\widehat{W}_{\mathrm{b}}$ (15) we still need an empirical expression for the covariance $\Sigma_{\hat{\mu}_{\beta|y}}$ of the estimated effects.

It holds for all $k, m \in \{1, ..., M\}, k \neq m$, that

$$\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right) \approx 0, \tag{39}$$

because we have thinned the MCMC sample in order to obtain an approximately independent chain. We find

$$\begin{aligned} \Sigma_{\hat{\mu}_{\beta|y}} &:= \mathrm{Cov}(\hat{\mu}_{\beta|y}) \\ &\overset{(13)}{=} \mathrm{Cov}\left(\frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}, \frac{1}{M}\sum_{k=1}^{M}\hat{\beta}^{(k)}\right) \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{k=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right) \\ &= \frac{1}{M^2}\left[\sum_{m=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}\right) + \sum_{m=1}^{M}\sum_{\substack{k=1\\k\neq m}}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right)\right] \\ &\overset{(39)}{\approx} \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}\right) \\ &= \frac{1}{M}\Sigma_{\beta|y}, \end{aligned} \tag{40}$$

where the last equation is due to the fact that all samples $\hat{\beta}^{(m)}, m = 1, ..., M$ left in the chain are representative of the posterior distribution. Thus,

$$\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\mu}_{\beta|y}}) \overset{(40)}{\approx} \frac{1}{M}\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}), \tag{41}$$

where the approximation in (41) is the more precise, the closer the chain is to independence.

We express the nearly unbiased estimator $\widehat{W}_{\mathrm{b}}$ on the basis of MCMC realizations as

$$\widehat{W}_{\mathrm{b}} \overset{(15),(41)}{\approx} \hat{\mu}_{\beta|y}^{\top}\hat{\Sigma}_X\hat{\mu}_{\beta|y} + \Big(1 - \frac{1}{M}\Big)\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}). \tag{42}$$

Plugging $\hat{\mu}_{\beta|y}$ (13) and $\hat{\Sigma}_{\beta|y}$ (14) into (42), we obtain:

$$\begin{aligned}\widehat{W}_{\mathrm{b}} \approx & \hat{\mu}_{\beta|y}^{\top}\hat{\Sigma}_X\hat{\mu}_{\beta|y} + (1 - \frac{1}{M})\mathrm{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}) \\ \overset{(13),(14)}{=} & \Big(\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)})^{\top}\Big)\hat{\Sigma}_X\Big(\frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}\Big) \\ & + \frac{M-1}{M}\Big[\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}^{(m)})^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)} \\ & - \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{m=1}^{M}(\hat{\beta}^{(k)})^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)}\Big] \\ = & \frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)})^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)}. \end{aligned} \tag{16}$$

We conclude that the estimator $\widehat{W}_{\mathrm{post}}$ (16) approximates the nearly unbiased estimator $\widehat{W}_{\mathrm{b}}$ (15) for the mean of the posterior genomic variance W_{b} in BRM.

REM

In Section Random Effect Model (REM) we assume that β in the stochastic model

$$Y = \mu + X\beta + \varepsilon, \tag{1}$$

is normally distributed with mean $\mu_\beta = 0$ and finite marginal variance-covariance matrix $\sigma_\beta^2\mathbb{1}_{p\times p}$:

$$\beta \sim \mathcal{N}(0, \sigma_\beta^2\mathbb{1}_{p\times p}). \tag{43}$$

We calculate the genomic variance V (2) as

$$\begin{aligned} V_{\mathrm{r}} &= \mathrm{Var}(X\beta) \\ &= \mathrm{Var}_\beta(\mathbb{E}[X\beta|\beta]) + \mathbb{E}_\beta[\mathrm{Var}(X\beta|\beta)] \\ &= \mathrm{Var}_\beta(\mathbb{E}[X]\beta) + \mathbb{E}_\beta[\beta^{\top}\Sigma_X\beta] \\ &= \sigma_\beta^2\mathbb{E}[X]^{\top}\mathbb{E}[X] + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\Big[\sigma_{ij}^X\beta_i\beta_j\Big] \\ &\overset{\mathbb{E}[X]=0}{=} \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}^X\Big(\sigma_{ij}^\beta + \mathbb{E}[\beta_i]\mathbb{E}[\beta_j]\Big) \\ &= \mathrm{tr}(\Sigma_X\Sigma_\beta) + \mathbb{E}[\beta]^{\top}\Sigma_X\mathbb{E}[\beta] \\ &\overset{(43)}{=} \sigma_\beta^2\mathrm{tr}(\Sigma_X) = \sigma_\beta^2\sum_{j=1}^{p}\mathrm{Var}(X_j). \end{aligned} \tag{18}$$

After obtaining an unbiased estimator $\hat{\sigma}_\beta^2$ for the variance component σ_β^2 (e.g. REML) the marginal genomic variance V_{r} (18) can be estimated by

$$\hat{V}_{\mathrm{r}} = \hat{\sigma}_\beta^2\mathrm{tr}(\hat{\Sigma}_X). \tag{19}$$

We calculate

$$\begin{aligned} \mathbb{E}\big[\hat{V}_{\mathrm{r}}\big] &= \mathbb{E}\Big[\hat{\sigma}_\beta^2\mathrm{tr}(\hat{\Sigma}_X)\Big] \\ &= \mathrm{Cov}\Big(\hat{\sigma}_\beta^2, \mathrm{tr}(\hat{\Sigma}_X)\Big) + \mathbb{E}\Big[\hat{\sigma}_\beta^2\Big]\mathbb{E}\Big[\mathrm{tr}(\hat{\Sigma}_X)\Big] \\ &= \sigma_\beta^2\mathrm{tr}(\Sigma_X) + \mathrm{Cov}\Big(\hat{\sigma}_\beta^2, \mathrm{tr}(\hat{\Sigma}_X)\Big). \end{aligned}$$

We conclude that $\hat{V}_{\mathrm{r}}$ (19) is a nearly unbiased estimator for the marginal genomic variance V_{r} (18) and is bias-free if the estimators $\hat{\sigma}_\beta^2$ and $\hat{\Sigma}_X$ are uncorrelated. The estimators $\hat{V}_{\mathrm{r}}^{\mathrm{real}}$ (21) and $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ (25) are nearly unbiased estimators for the marginal genomic variance V_{r} (18) using the same reasoning.
In order to explicitly include LD into the expression of the genomic variance we condition on the effect vector β and define

$$W := \mathrm{Var}(X\beta|\beta) = \beta^{\top}\Sigma_X\beta = \mathrm{tr}(\Sigma_X\beta\beta^{\top}), \tag{11}$$

which is a quadratic form in the normally distributed β. This random variable has expectation

$$\begin{aligned} \mathbb{E}[W] &= \mathbb{E}\Big[\mathrm{tr}\Big(\Sigma_X\beta\beta^{\top}\Big)\Big] \\ &= \mathrm{tr}\Big(\Sigma_X\mathbb{E}\Big[\beta\beta^{\top}\Big]\Big) \\ &= \mathrm{tr}\Big(\Sigma_X\Big(\mathbb{E}\Big[\beta\Big]\mathbb{E}\Big[\beta^{\top}\Big] + \mathrm{Cov}(\beta)\Big)\Big) \\ &= \mathrm{tr}(\Sigma_X\sigma_\beta^2\mathbb{1}_{p\times p}) \\ &= \sigma_\beta^2\sum_{j=1}^{p}\mathrm{Var}(X_j) = V_{\mathrm{r}}. \end{aligned}$$

By strictly conditiong on the data we define the unbiased predictor W_{r}

$$W_{\mathrm{r}} := \mathbb{E}[W|y] = \mathrm{tr}\Big(\Sigma_X\mathbb{E}\Big[\beta\beta^{\top}|y\Big]\Big) = \mu_{\beta|y}^{\top}\Sigma_X\mu_{\beta|y} + \mathrm{tr}(\Sigma_X\Sigma_{\beta|y}), \tag{26}$$

for the random genomic variance W (11), where $\mu_{\beta|y} := \mathbb{E}[\beta|y]$ is the BLUP of β and $\Sigma_{\beta|y} := \mathrm{Cov}(\beta|y)$.
The predictor W_{r} is by definition unbiased for the random variable W, if $\mathbb{E}[W_{\mathrm{r}}] = \mathbb{E}[W]$. We calculate

$$\mathbb{E}[W_{\mathrm{r}}] = \mathbb{E}\big[\mathbb{E}[W|y]\big] = \mathbb{E}[W] = V_{\mathrm{b}}$$

and conclude that W_{r} (26) is an unbiased predictor for W (11).
In the conditional model

$$y_i = \mu + (\mathbf{X}\beta)_i + \varepsilon_i := \mu + \sum_{j=1}^{p}x_{ij}\beta_j + \varepsilon_i, \quad i = 1, ..., n, \tag{4}$$

it holds that

$$y \sim \mathcal{N}\Big(\mu, \underbrace{\mathbf{X}\mathbf{X}^{\top}\sigma_\beta^2 + \sigma_\varepsilon^2\mathbb{1}_{n\times n}}_{:=\tilde{\Sigma}^{-1}}\Big),$$

and we investigate the joint distribution of y and β as

$$\begin{pmatrix} y \\ \beta \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{\Sigma}^{-1} & \sigma_\beta^2\mathbf{X} \\ \sigma_\beta^2\mathbf{X}^{\top} & \sigma_\beta^2\mathbb{1}_{p\times p} \end{pmatrix}\right].$$

Because of the joint normal distribution we obtain

$$\beta|y \sim \mathcal{N}(\sigma_\beta^2\mathbf{X}^{\top}\tilde{\Sigma}(y-\mu), \sigma_\beta^2\mathbb{1}_{p\times p} - \sigma_\beta^2\mathbf{X}^{\top}\tilde{\Sigma}\mathbf{X}\sigma_\beta^2).$$

The BLUP for β is defined as $\mu_{\beta|y} := \mathbb{E}[\beta|y]$ (Searle *et al.* 1992) such that we obtain

$$\mu_{\beta|y} = \mathbb{E}[\beta|y] = \sigma_\beta^2\mathbf{X}^{\top}\tilde{\Sigma}(y-\mu), \tag{44}$$

as well the variance-covariance matrix $\Sigma_{\beta|y}$ of the conditional distribution of β

$$\Sigma_{\beta|y} := \mathrm{Cov}(\beta|y) = \sigma_\beta^2\mathbb{1}_{p\times p} - \sigma_\beta^2\mathbf{X}^{\top}\tilde{\Sigma}\mathbf{X}\sigma_\beta^2. \tag{27}$$

and

$$\begin{aligned} \mathrm{Cov}(\mu_{\beta|y}) &= \mathrm{Var}(\beta) - \mathbb{E}\left[\mathrm{Cov}(\beta|y)\right] \\ &= \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2 \end{aligned} \tag{45}$$

We plug (27) into W_r (26) and obtain

$$\begin{aligned} W_\mathrm{r} &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \mathrm{tr}(\Sigma_X \Sigma_{\beta|y}) \\ &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \sigma_\beta^2 \mathrm{tr}(\Sigma_X) - \mathrm{tr}(\Sigma_X \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2) \\ &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + V_\mathrm{r} - \mathrm{tr}(\Sigma_X \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2). \end{aligned} \tag{46}$$

We replace the variance components σ_β^2 and σ_ε^2 in (44) and (27) by unbiased estimators (e.g. REML) and plug them into W_r (26):

$$\begin{aligned} \widehat{W}_\mathrm{r} &= \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} + \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\beta|y}), \qquad (28) \\ &= \hat{V}_r + \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \hat{\sigma}_\beta^4 \mathrm{tr}(\hat{\Sigma}_X \mathbf{X}^\top \hat{\hat{\Sigma}} \mathbf{X}). \end{aligned}$$

We make the important note that the unbiasedness of the predictor $\widehat{W}_\mathrm{r}$ (28) can only be given conditional on the estimated variance components $\hat{\sigma}_\beta^2$ and $\hat{\sigma}_\varepsilon^2$ because of dependencies between these estimators and y. This problem is common in REM and also holds true for the BLUP $\hat{\mu}_{\beta|y} = \widehat{\mathbb{E}}[\beta|y] = \hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\hat{\Sigma}}(y - \hat{\mu})$ whose unbiasedness can only be asserted conditionally:

$$\begin{aligned} \mathbb{E}\left[\hat{\mu}_{\beta|y} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] &= \mathbb{E}\left[\hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\hat{\Sigma}}(y - \hat{\mu}) | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &= \hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\hat{\Sigma}} \mathbb{E}\left[y - \hat{\mu} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &= 0 = \mathbb{E}[\beta]. \end{aligned}$$

Similarly, we calculate

$$\begin{aligned} \mathbb{E}\left[\widehat{W}_\mathrm{r} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] &= \mathbb{E}\left[\hat{V}_r + \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \hat{\sigma}_\beta^4 \mathrm{tr}(\hat{\Sigma}_X \mathbf{X}^\top \hat{\hat{\Sigma}} \mathbf{X}) | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &= \hat{\sigma}_\beta^2 \mathbb{E}\left[\mathrm{tr}(\hat{\Sigma}_X) | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] + \mathbb{E}\left[\hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &\quad - \hat{\sigma}_\beta^4 \mathrm{tr}(\mathbb{E}\left[\hat{\Sigma}_X | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \mathbf{X}^\top \hat{\hat{\Sigma}} \mathbf{X}) \\ &\overset{(35)}{=} \hat{\sigma}_\beta^2 \mathrm{tr}(\Sigma_X) + \mathbb{E}\left[\hat{\mu}_{\beta|y} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right]^\top \Sigma_X \mathbb{E}\left[\hat{\mu}_{\beta|y} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &\quad + \mathrm{tr}\left(\Sigma_X \mathrm{Cov}\left(\hat{\mu}_{\beta|y} | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right)\right) \\ &\quad + \sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{Cov}\left(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right) \\ &\quad - \hat{\sigma}_\beta^4 \mathrm{tr}(\Sigma_X \mathbf{X}^\top \hat{\hat{\Sigma}} \mathbf{X}) \\ &\overset{(45)}{=} \hat{\sigma}_\beta^2 \mathrm{tr}(\Sigma_X) \\ &\quad + \sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{Cov}\left(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right) \\ &= \mathbb{E}\left[W | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right] \\ &\quad + \sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{Cov}\left(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j | \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\right), \end{aligned}$$

such that we can assert nearly (conditional) unbiasedness.
In the equivalent linear model

$$y = \mu + \mathbf{X}\beta + \varepsilon = \mu + g + \varepsilon, \tag{22}$$

it holds that

$$\mu_{g|y} := \mathbb{E}[g|y] = \mathbb{E}[\mathbf{X}\beta|y] = \mathbf{X}\mu_{\beta|y}$$

Consequently,

$$\frac{1}{n-1} \hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} = \frac{1}{n-1} \hat{\mu}_{\beta|y}^\top \mathbf{X}^\top \mathbf{X} \hat{\mu}_{\beta|y} = \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y}.$$

In addition to that, we calculate

$$\begin{aligned} \Sigma_{g|y} := \mathrm{Cov}(g|y) &= \mathbf{X}\mathrm{Cov}(\mu_{\beta|y})\mathbf{X}^\top \\ &= \mathbf{X}\Sigma_{\mu_{\beta|Y}}\mathbf{X}^\top \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n-1} \mathrm{tr}(\hat{\Sigma}_{g|y}) &= \frac{1}{n-1} \mathrm{tr}(\mathbf{X}\hat{\Sigma}_{\mu_{\beta|y}}\mathbf{X}^\top) \\ &= \frac{1}{n-1} \mathrm{tr}(\mathbf{X}^\top \mathbf{X} \hat{\Sigma}_{\mu_{\beta|y}}) \\ &= \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\mu_{\beta|y}}). \end{aligned}$$

This shows the equivalence of

$$\widehat{W}_\mathrm{r}^{\mathrm{equi}} := \frac{1}{n-1} \hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} + \frac{1}{n-1} \mathrm{tr}\left(\hat{\Sigma}_{g|y}\right). \tag{29}$$

to $\widehat{W}_\mathrm{r}$ (28) in the linear model (4).

MEM

Up-to-now we have considered random effect models only. We extend model (1) by including a fixed effect Zf which results in a mixed effect model (MEM) of the form

$$Y = Zf + X\beta + \varepsilon, \tag{47}$$

where f is a k-vector of fixed effects as in section Fixed Effect Model (FEM), β is a p-vector of random effects as in section Random Effect Model (REM), Z is a random k row-vector and X is a random p row-vector. We assume that Zf and ε as well as $X\beta$ and ε are independent.
We calculate

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(Zf + X\beta + \varepsilon) \\ &= \mathrm{Var}(Zf) + \mathrm{Var}(X\beta) + 2\mathrm{Cov}(Zf, X\beta) + \sigma_\varepsilon^2. \end{aligned} \tag{48}$$

Inferences on the additive genomic variance of the fixed effect Zf can be done as in Section Fixed Effect Model (FEM) and inferences on the additive genomic variance of the random effect $X\beta$ can be done as in Section Random Effect Model (REM). If one is interested in the contribution of LD between fixed effects and random effects, e.g. when including single important markers as fixed effects in the MEM, we propose to predict the random conditional covariance

$$\mathrm{Cov}(Zf, X\beta|\beta) = f^\top \mathrm{Cov}(Z, X)\beta. \tag{49}$$

We propose

$$\hat{f}\hat{\Sigma}_{XZ}\hat{\mu}_{\beta|y} \tag{50}$$

as a predictor for (49), where

$$\hat{f} = (\mathbf{Z}^\top \hat{\hat{\Sigma}} \mathbf{Z})^{-1} \mathbf{Z}^\top \hat{\hat{\Sigma}} y \tag{51}$$

is the BLUE of f with conditional covariance

$$\begin{aligned}\mathrm{Cov}\left(\hat{f}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right) &= \mathrm{Cov}\left((\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\hat{\Sigma}}y|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\\ &= (\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathrm{Cov}\left(y|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\hat{\hat{\Sigma}}\mathbf{Z}(\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\\ &= (\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}. \end{aligned} \tag{52}$$

We calculate:

$$\begin{aligned}\mathbb{E}\left[\hat{f}\hat{\Sigma}_{XZ}\hat{\mu}_{\beta|y}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right] &\overset{(35)}{=} \mathrm{tr}\left(\Sigma_{XZ}\mathrm{Cov}\left(\hat{f},\hat{\mu}_{\beta|y}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\right)\\ &\quad+\sum_{i=1}^p\sum_{j=1}^p\mathrm{Cov}\left(\hat{\sigma}_{ij}^{XZ},\hat{f}_i(\hat{\mu}_{\beta|y})_j|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\\ &= \sum_{i=1}^p\sum_{j=1}^p\mathrm{Cov}\left(\hat{\sigma}_{ij}^{XZ},\hat{f}_i(\hat{\mu}_{\beta|y})_j|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right),\end{aligned}$$

because

$$\begin{aligned}\mathrm{Cov}\left(\hat{f},\hat{\mu}_{\beta|y}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right) &= \mathrm{Cov}\left(\hat{f},\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\hat{\Sigma}}y|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\\ &\quad-\mathrm{Cov}\left(\hat{f},\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\hat{\Sigma}}\mathbf{Z}\hat{f}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\\ &\overset{(51)}{=} \mathrm{Cov}\left((\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\hat{\Sigma}}y,\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\hat{\Sigma}}y|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\\ &\quad-\hat{\sigma}_\beta^2\mathrm{Cov}\left(\hat{f}|\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\right)\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{X}\\ &\overset{(52)}{=} \hat{\sigma}_\beta^2(\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\hat{\Sigma}}\hat{\hat{\Sigma}}^{-1}\hat{\hat{\Sigma}}\mathbf{X}\\ &\quad-\hat{\sigma}_\beta^2(\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\hat{\Sigma}}\mathbf{X}\\ &= 0.\end{aligned}$$

Because the covariance (49) has expectation 0, the predictor (50) is nearly unbiased given estimators $\hat{\sigma}_\beta^2$ and $\hat{\sigma}_\varepsilon^2$.

Notes on the mean-centering of X

In model (1) in Section Linear Models and the Genomic Variance we consider X to be a random row vector with expectation 0. If we depart from that assumption and consider $\tilde{X}$ with $\mathbb{E}[\tilde{X}] \neq 0$ and $\mathrm{Cov}(\tilde{X}) = \Sigma_X$ instead of X, we reformulate model (1) based on $\tilde{X}$ as

$$\begin{aligned}Y = \mu + \tilde{X}\beta + \varepsilon &= \mu + (\tilde{X} - \mathbb{E}[\tilde{X}])\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon\\ &\overset{\mathrm{d}}{=} \mu + X\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon.\end{aligned}$$

In the FEM (β deterministic) the fixed term $\mathbb{E}[X]\beta$ is absorbed by the intercept such that

$$Y = \tilde{\mu} + X\beta + \varepsilon$$

with $\tilde{\mu} = \mu + \mathbb{E}[\tilde{X}]\beta$ and we obtain linear model (1) with mean-centered data and but different (fixed) intercept. Consequently, the genomic variance in the FEM V_f (6) is unchanged whether we consider mean-centered allele content X or not ($\tilde{X}$):

$$\mathrm{Var}(X\beta) = \beta^\top\Sigma_X\beta = \mathrm{Var}(\tilde{X}\beta).$$

In BRM and REM, where $\beta \sim (\mu_\beta, \Sigma_\beta)$ is a random variable, the term $\mathbb{E}[\tilde{X}]\beta$ is a random variable itself and is absorbed by the residual instead of the intercept:

$$\begin{aligned}Y &= \mu + \tilde{X}\beta + \varepsilon\\ &= \mu + (\tilde{X} - \mathbb{E}[\tilde{X}])\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon\\ &= \mu + X\beta + \tilde{\varepsilon}\end{aligned}$$

where $\tilde{\varepsilon} \sim (0, \sigma_\varepsilon^2 + \mathbb{E}[X]\Sigma_\beta\mathbb{E}[X]^\top)$.
For the genomic variance V (2) in BRM and REM it makes a difference whether we consider the mean-centered X or $\tilde{X}$ because:

$$\begin{aligned}\mathrm{Var}(\tilde{X}\beta) &= \mathrm{Var}_\beta(\mathbb{E}[\tilde{X}\beta|\beta]) + \mathbb{E}_\beta[\mathrm{Var}(\tilde{X}\beta|\beta)]\\ &= \mathrm{Var}_\beta(\mathbb{E}[\tilde{X}]\beta) + \mathbb{E}_\beta[\beta^\top\Sigma_X\beta]\\ &= \mathbb{E}[\tilde{X}]^\top\Sigma_\beta\mathbb{E}[\tilde{X}] + \sum_{i=1}^p\sum_{j=1}^p\mathbb{E}\left[\sigma_{ij}^X\beta_i\beta_j\right]\\ &= \mathbb{E}[\tilde{X}]\Sigma_\beta\mathbb{E}[\tilde{X}]^\top + \mathrm{tr}(\Sigma_X\Sigma_\beta) + \mu_\beta^\top\Sigma_X\mu_\beta\\ &\neq \mathrm{tr}(\Sigma_X\Sigma_\beta) + \mu_\beta^\top\Sigma_X\mu_\beta = \mathrm{Var}(X\beta).\end{aligned}$$

This is consistent with the approach in Section Random Effect Model (REM) in the realized model (4) where the genomic variance in REM is estimated based on $\mathbf{X}$:

$$\mathrm{Cov}(\mathbf{X}\beta) = \mathbf{X}\mathbf{X}^\top\sigma_\beta^2 \tag{20}$$

or when using GRM's in the equivalent model (22)

$$\sigma_\beta^2\mathbf{X}\mathbf{X}^\top = \frac{1}{p}\mathbf{X}\mathbf{X}^\top(p\sigma_\beta^2) =: \mathbf{G}\sigma_g^2.$$

The genomic variance in these models clearly depends on whether using mean-centered matrices $\mathbf{X}$ or not, especially in the equivalent model transformations of $\mathbf{X}$ change the variance-covariance matrix of g. The GRM's are generally based on mean-centered matrices which is the reason why we have based the main analysis in this paper on the mean-centered approach.
The random genomic variance W (11), however, does not depend on centering:

$$W := \mathrm{Var}(X\beta|\beta) = \beta^\top\Sigma_X\beta = \mathrm{tr}(\Sigma_X\beta\beta^\top) \tag{11}$$

and is therefore consistent with the genomic variance in FEM with respect to the independence to mean-centering.