Best Prediction of the Additive Genomic Variance in Random-Effects Models

Nicholas Schreck[*,1], **Hans-Peter Piepho**[**] **and Martin Schlather**[*,†]

[*]Chair of Stochastics and Its Applications, University of Mannheim, B6, 26, 68159 Mannheim, Germany, [**]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, 70593 Stuttgart, 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 in linear models with random marker effects can be defined as a random variable that is in accordance with classical quantitative genetics theory. Common approaches to estimate the genomic variance in random-effects linear models based on genomic marker data can be regarded as the unconditional (or prior) expectation of this random additive genomic variance, and result in a negligence of the contribution of linkage disequilibrium.

We introduce a novel best prediction (BP) approach for the additive genomic variance in both the current and the base population in the framework of genomic prediction using the gBLUP-method. The resulting best predictor is the conditional (or posterior) expectation of the additive genomic variance when using the additional information given by the phenotypic data, and is structurally in accordance with the genomic equivalent of the classical additive genetic variance in random-effects models. In particular, the best predictor includes the contribution of (marker) linkage disequilibrium to the additive genomic variance and eliminates the missing contribution of LD that is caused by the assumptions of statistical frameworks such as the random-effects model. We derive an empirical best predictor (eBP) and compare its performance with common approaches to estimate the additive genomic variance in random-effects models on commonly used genomic datasets.

KEYWORDS best prediction; genetic variance; quantitative genetics; genomic variance; random-effects models; BLUP; whole-genome regression

Introduction

The additive genetic variance is defined as the variance of the breeding value (BV) and is the most important determinant of the response of a population to selection (Falconer and Mackay 1996). 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 quantities of interest 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 genetic variation exists, most of the genetic variation is additive, such that it is usually sufficient to investigate the additive genetic variance (Hill *et al.* 2008). More specifically, epistasis is only important on the gene level but not for the genetic variance (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 additive genomic 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" genetic variance can be captured by regression on influential markers. Initially, researchers have used genome-wide association studies (GWAS) in order to find quantitative trait loci (QTL) by single-marker 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"

[1]Corresponding author: Chair of Stochastics and Its Applications, University of Mannheim, B6, 26, 68159 Mannheim, Germany. Email: nschreck@mail.uni-mannheim.de

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 (markers) 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, Bernardo (1994) decided to fit all (RFLP-) markers in maize jointly using genomic best linear unbiased prediction (gBLUP), where he assumes the marker effect vector to be random. In animal breeding, Meuwissen *et al.* (2001) used Bayesian approaches (BayesA and BayesB) to fit all markers jointly in order to predict breeding values. Then, Yang *et al.* (2010) estimated the genomic variance in an approach that they termed genome-wide complex trait analysis genomic restricted maximum likelihood (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, also termed "chip heritability" (Zhou *et al.* 2013), of about 45%. They concluded that the main reason for the remaining missing heritability was incomplete LD of causal variants with the genotyped SNPs, which refers to the general difference between the genetic variance and the genomic variance (Powell *et al.* 2010; de los Campos *et al.* 2015). However, the GCTA-GREML approach can be biased upwards as well as downwards (Wolc *et al.* 2013; de los Campos *et al.* 2015; Lehermeier *et al.* 2017; Fernando *et al.* 2017a). 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 the marker genotypes (Bulmer 1971). Some authors argue that estimators similar to GCTA-GREML lack the contribution of LD (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 values 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 additive genomic variance. In a study on the model plant *Arabidopsis thaliana* (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) argue that the reason for the better performance of the Bayesian estimator for the additive genomic variance (already mentioned in Sorensen *et al.* (2000); Zhou *et al.* (2013); Fernando and Garrick (2013); Fernando *et al.* (2017b)) is the explicit inclusion of linkage disequilibrium.

We show that the additive genomic variance in linear models with random marker effects (REM) can be defined as a random variable. Based on this premise, we propose a novel predictor of the additive genomic variance and place existing estimators in a joint framework permitting comparison with the new predictor. We contribute to the solution of many of the above mentioned controversies by reviewing common approaches to estimate the additive genomic variance, e.g. GCTA-GREML, and show that they estimate the unconditional (or prior) expectation of the random additive genomic variance. Combined with the assumptions on the unconditional distribution of the marker effects in the gBLUP-method this leads to an insufficient adaptation to the data and a negligence of the contribution of LD.
We introduce a novel best prediction approach for the additive genomic variance in both the current and the base population, i.e. we use the conditional (or posterior) expectation of the random additive genomic variance given the additional information by the phenotypic values for an improved adaptation to the data. We decompose the best predictor into the GCTA-GREML estimator and a function for the contribution of marker LD which determines whether GCTA-GREML is biased up- or downwards. The best predictor is structurally in accordance with the genomic equivalent of the additive genetic variance from classical quantitative genetics, i.e. it explicitly includes the contribution of LD. We propose an empirical best predictor (eBP) and illustrate our theoretical results on several commonly used genomic datasets.

Material and Methods

Linear Models

The connection of the n-vector y of phenotypic values and the mean-centered n-vector g of genomic values is given by

$$y = \mu \mathbb{1}_n + g + \varepsilon, \tag{1}$$

where μ denotes a fixed intercept, $\mathbb{1}_n := (1, ..., 1)^\top$ is a n-row-vector containing 1's, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{1}_{n\times n})$ denotes environmental deviations, and $\mathbb{1}_{n\times n}$ is the identity matrix of dimension n. For simplicity, we restrict the sample mean of the genotypic values to be 0 ($\bar{g} := \frac{1}{n}\mathbb{1}_n^\top g = 0$).
In the following, we assume that the genome is mapped with $p \in \mathbb{N}$ markers and we denote by $\mathbf{X}$ the $n \times p$ design matrix coding the genotypes of the markers. Then, the genomic values can be separated into the coded genotypes of the single markers and their corresponding p-vector β of marker effects:

$$g := \mathbf{X}\beta = \left[\sum_{j=1}^{p} \left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\cdot j}\right)\beta_j\right]_{i=1,...,n}, \tag{2}$$

where $\mathbf{P} := \mathbb{1}_{n\times n} - \mathbb{1}_n\mathbb{1}_n^\top/n$ is the idempotent $n \times n$-matrix used for column-wise mean-centering and $\bar{\mathbf{x}}_{\cdot j} := \sum_{i=1}^{n} \mathbf{x}_{ij}/n$ for $j = 1, ..., p$. The restriction of the column-means of the marker genotype matrix to be 0 guarantees that the sample mean of the genomic values in (1) equals 0. This ensures the uniqueness of the definition of the vector g of genomic values in (2). Otherwise, different coding of the marker genotypes lead to different genomic values g.
Model (1) is called linear equivalent model (Henderson 1984) to the "standard" additive linear regression model

$$y = \mu \mathbb{1}_n + \mathbf{X}\beta + \varepsilon. \tag{3}$$

Model (3) allows for marker specific investigations and inferences on the genomic contribution to the phenotypic

values, whereas estimation of parameters in model (1) has computational advantages.

Model (3) is a realization of n draws from the underlying data-generating process of the (mean-centered) marker genotypes $(X_1, ..., X_p)$ (Bühlmann and van de Geer 2011). This distribution as well as the corresponding genomic values in (2) relate to the current population of individuals.
When we are interested in the genomic values in the corresponding consistent base population, we should take the relationship (correlation) between the individuals into account (Powell *et al.* 2010; Legarra 2015). Assume that we have given a $n \times n$ relationship matrix $\mathbf{R}$. Instead of the genomic values g or $\mathbf{P}\mathbf{X}\beta$ we investigate the uncorrelated genomic values defined by

$$g^* := \mathbf{R}^{-0.5} g = \mathbf{R}^{-0.5}\mathbf{P}\mathbf{X}\beta =: \mathbf{X}^*\beta. \tag{4}$$

These are realizations of n draws from the underlying data-generating process $(X_1^*, ..., X_p^*)$ of marker genotypes in the base population. The sample mean of the genomic values in the base population, $\frac{1}{n}\mathbb{1}_n^\top g^*$, is usually different from 0.

The random-effects model is the statistical model that is probably most popular in genomic applications. Inferences on quantities based on genomic data in model (1) and (3) are often performed with the genomic best-linear-unbiased-prediction (gBLUP) method (Bernardo 1994). In this framework, we consider the single p components of the marker effect vector β in set-up (3) as independent normal random variables:

$$\beta_j \sim \mathcal{N}\left(0, \sigma_\beta^2\right), \quad j = 1, ..., p, \tag{5}$$

which implies that the effects are drawn at random from a common fixed normal distribution for each marker genotype.
In order to maintain the equivalence of models (1) and (3) we have to ensure the following equality in distribution:

$$g \overset{\mathrm{d}}{=} \mathbf{P}\mathbf{X}\beta.$$

This can, for instance, be achieved by setting

$$g \sim \mathcal{N}\left(0, \sigma_g^2 \mathbf{G}\right),$$

where $\sigma_g^2 := c\sigma_\beta^2$ and

$$\mathbf{G} := \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}/c \tag{6}$$

for some arbitrary $c > 0$. The $n \times n$-matrix $\mathbf{G}$ is called genomic relationship matrix (GRM) and often $c := 2\sum p_j(1 - p_j)$, where p_j is the frequency of the minor allele at marker j (VanRaden 2008). For additional information on the gBLUP-method we refer to Appendix Genomic Best Linear Unbiased Prediction.

Definitions of the Genomic Variance

We give an overview of the different approaches to define a genomic variance in the framework of the linear models (1) and (3).

Without further assumptions on the nature of the genome, we can define the sample variance

$$s_g^2 := \frac{1}{n-1} g^\top g = \frac{1}{n-1}\beta^\top\mathbf{X}^\top\mathbf{P}\mathbf{X}\beta = \beta^\top\hat{\Sigma}_X\beta \tag{7}$$

of the mean-centered n-vector of genomic values $g = \mathbf{P}\mathbf{X}\beta$, see (2), in the current population (Ould Estaghvirou *et al.* 2013). Here, $\hat{\Sigma}_X := \mathbf{X}^\top\mathbf{P}\mathbf{X}/(n-1)$ defines the sample variance-covariance matrix of the marker genotypes in the current population.
In the base population, we define the sample variance

$$s_{g^*}^2 := \frac{1}{n-1}(g^*)^\top\mathbf{P}g^* = \frac{1}{n-1}\beta^\top(\mathbf{X}^*)^\top\mathbf{P}\mathbf{X}^*\beta = \beta^\top\hat{\Sigma}_{X^*}\beta \tag{8}$$

of the uncorrelated genomic values $g^* = \mathbf{X}^*\beta$, see (4). Here, $\hat{\Sigma}_{X^*} := (\mathbf{X}^*)^\top\mathbf{P}\mathbf{X}^*/(n-1)$ defines the sample variance-covariance matrix of the marker genotypes in the base population.

Alternatively, we can define the theoretical variance of the genomic values directly in the REM. The linear model is generated by drawing from the data-generating process of the marker genotypes (representative individual), and the model assumptions in the REM dictate that marker effects are random variables. This gives rise to three different sources of variance of the genomic values in the REM (marker genotypes random, marker effects random, or both random).
The additive genomic variance of a randomly sampled (representative) individual (Gianola *et al.* 2009; de los Campos *et al.* 2015; Fernando *et al.* 2017b) equals

$$\mathrm{Var}(X\beta \,|\, \beta) = \beta^\top\Sigma_X\beta, \tag{9}$$

where Σ_X denotes the variance-covariance matrix of the marker genotypes.
The variance of a randomly sampled (representative) individual with random marker effects is given by

$$\mathrm{Var}(X\beta) = \sigma_\beta^2\mathrm{tr}(\Sigma_X). \tag{10}$$

This is not the additive genomic variance (Gianola *et al.* 2009; de los Campos *et al.* 2015).
The variance of a randomly sampled trait averaged over individuals with fixed genotypes $\mathbf{X}$ equals

$$\frac{1}{n}\mathrm{tr}(\mathrm{Cov}(X\beta \,|\, X)) = \sigma_\beta^2\frac{n-1}{n}\mathrm{tr}(\hat{\Sigma}_X) \approx \sigma_\beta^2\mathrm{tr}(\hat{\Sigma}_X), \tag{11}$$

and does not equal the additive genomic variance.
We derive the equalities in (9), (10) and (11) in more detail in the Appendix Theoretical Variances of the Genomic Values in the REM. These quantities refer to the genotypes in the current population. We can apply the same definitions in the base population by considering the data-generating process of the genotypes in the base population (exchange X by X^*).
In Table 1 we give an overview of the different possibilities to define the variance of the genomic values in the REM.

In actual applications, we have to replace Σ_X in (9) by its estimator $\hat{\Sigma}_X$. Consequently, the sample variance (7) as well as the theoretical (9) effectively represent the additive genomic variance, the genomic equivalent of the additive genetic variance (Bulmer 1971; Falconer and Mackay 1996), in the current population. In the following, we do not explicitly distinguish between the sample or the theoretical version of the variance, and will speak only of the additive genomic variance.
In the following, we focus on the estimation of the additive genomic variance in the general form

$$s_{g,\mathbf{B}}^2 := \frac{1}{n-1} g^\top\mathbf{B}g = \frac{1}{n-1}\beta^\top\mathbf{X}^\top\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{X}\beta, \tag{12}$$

which is a non-negative quadratic form of the genomic values. By specifying the positive semi-definite and symmetric

$n \times n$-matrix $\mathbf{B}$ we determine whether the genomic variance refers to the current population ($\mathbf{B} = \mathbb{1}_{n\times n}$), see (7), or the base population ($\mathbf{B} = \mathbf{R}^{-0.5}\mathbf{P}\mathbf{R}^{-0.5}$), see (8). Because the randomness of the marker genotypes is not explicitly necessary to derive (12), we can easily express all results in the terminology of the genomic values g defined in the equivalent model.

In the framework of the REM, the marker effects β in model (3) and the genomic values g in model (1) are random variables. Consequently, the additive genomic variance in (12) is also a random variable, and has to be predicted in an optimal way before finally being estimated.
First, we will show that estimators for the unconditional expectation of (12), like GCTA-GREML, are of the form (10) and (11), and therefore do not estimate the additive genomic variance. Then, we introduce the (frequentist) best predictor for the additive genomic variance $s^2_{g,\mathbf{B}}$ and show that this approach maintains the structure of the additive genomic variance in (12), the genomic equivalent of the additive genetic variance.

Table 1 Overview of different definitions of the variance of the genomic values in the current population and their expression in the random-effects model. Analogous quantities for the base population can be obtained by exchanging X by X^*. The sample variance s^2_g and the theoretical variance $\mathrm{Var}(X\beta|\beta)$ define the sample and theoretical version of the additive genomic variance.

Variance of Genomic Values			
Sample Variance	Theoretical Variance		
s^2_g	$\mathrm{Var}(X\beta)$	$\mathrm{tr}(\mathrm{Cov}(X\beta\|X))$	$\mathrm{Var}(X\beta\|\beta)$
$\beta^\top\hat{\Sigma}_X\beta$	$\sigma^2_\beta\mathrm{tr}(\Sigma_X)$	$n\sigma^2_\beta\mathrm{tr}(\hat{\Sigma}_X)$	$\beta^\top\Sigma_X\beta$

The Expectation of the Additive Genomic Variance

The expectation of the random variable $s^2_{g,\mathbf{B}}$ in (12) minimizes the quadratic form

$$\mathbb{E}\left[\left(s^2_{g,\mathbf{B}} - a\right)^2\right],$$

with respect to all real numbers a, i.e. $\tilde{a} := \mathbb{E}\left[s^2_{g,\mathbf{B}}\right]$ is the best approximation of $s^2_{g,\mathbf{B}}$ in the absence of additional information (van der Vaart 2007). The unconditional (or prior) expectation of $s^2_{g,\mathbf{B}}$ equals

$$\begin{aligned}\mathbb{E}\left[s^2_{g,\mathbf{B}}\right] &= \mathbb{E}\left[\frac{1}{n-1}\beta^\top\mathbf{X}^\top\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{X}\beta\right]\\ &= \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{X}\mathbb{E}\left[\beta\beta^\top\right]\right)\\ &\overset{(5)}{=} \frac{1}{n-1}\sigma^2_\beta\mathrm{tr}\left(\mathbf{X}^\top\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{X}\right)\\ &\overset{(6)}{=} \frac{1}{n-1}\sigma^2_g\mathrm{tr}\left(\mathbf{B}\mathbf{G}\right)\end{aligned} \tag{13}$$

because of the properties of the trace.

For the additive genomic variance in the current population, s^2_g, we choose $\mathbf{B} = \mathbb{1}_{n\times n}$ in (13) and obtain

$$V := \mathbb{E}\left[s^2_g\right] = \sigma^2_\beta\mathrm{tr}\left(\hat{\Sigma}_X\right) \tag{14}$$

in model (3) or

$$V = \frac{1}{n-1}\sigma^2_g\mathrm{tr}\left(\mathbf{G}\right) \tag{15}$$

in the equivalent model (1). Unconditional expectations of the form V for the additive genomic variance are considered in Ould Estaghvirou *et al.* (2013), for example.

For the additive genomic variance in the base population, $s^2_{g^*}$, we choose $\mathbf{B} = \mathbf{R}^{-0.5}\mathbf{P}\mathbf{R}^{-0.5}$ in (13) and obtain

$$V^* := \mathbb{E}\left[s^2_{g^*}\right] = \sigma^2_\beta\mathrm{tr}\left(\hat{\Sigma}_{X^*}\right) \tag{16}$$

in model (3) or

$$V^* = \frac{1}{n-1}\sigma^2_g\mathrm{tr}\left(\mathbf{P}\mathbf{R}^{-0.5}\mathbf{G}\mathbf{R}^{-0.5}\right) \tag{17}$$

in the equivalent model. Often (VanRaden 2008; Yang *et al.* 2010, 2011; Speed *et al.* 2012; Vinkhuyzen *et al.* 2014; Legarra 2015), the matrix $\mathbf{R}$ used for the transformation to the base population is assumed to be the GRM $\mathbf{G}$ defined in (6). Then, the unconditional expectation V^* of the additive genomic variance simplifies to

$$V^*_s := \frac{1}{n-1}\sigma^2_g\mathrm{tr}\left(\mathbf{P}\right) = \sigma^2_g, \tag{18}$$

and the variance component σ^2_g from the gBLUP-method is considered as the (unconditional expectation of the) additive genomic variance in the base population. We recommend caution when using this simplification because the GRM $\mathbf{G}$ is in general singular (because $\mathbf{P}$ is singular), and therefore $\mathbf{G}^{-1}$ is not well defined.

We emphasize that only the diagonal elements of the sample variance-covariance matrix ($\hat{\Sigma}_X$ or $\hat{\Sigma}_{X^*}$) of the marker genotypes influence the unconditional expectations V and V^* of the additive genomic variance. The model assumptions in the REM dictate the matrix $\mathbb{E}\left[\beta\beta^\top\right]$ to be diagonal which leads to a negligence of the off-diagonal elements of $\hat{\Sigma}_X$ or $\hat{\Sigma}_{X^*}$ in (13). The covariances (LD) between the marker genotypes are not included and V, $V*$ and V^*_s are of the same form as $\mathrm{Var}(X\beta)$ in (10) and $\frac{1}{n}\mathrm{tr}(\mathrm{Cov}(X\beta \mid X))$ in (11). This implies that the unconditional expectation of the random additive genomic variance $s^2_{g,\mathbf{B}}$ is structurally not fully in accordance with the additive genomic variance.

Explicit formulae for the estimation of the unconditional expectations V, V^* and V^*_s will be given in the Appendix Estimation of the Additive Genomic Variance in the REM.

Best Prediction of the Additive Genomic Variance

The unconditional expectation of $s^2_{g,\mathbf{B}}$ in (13) is strongly influenced by the model assumption on the marginal distribution of the marker effects and does not use additional information given by the phenotypic values y in model equations (1) and (3). By contrast, the conditional expectation, given the phenotypic values y, can make use of the information in y.
Generally, the conditional (or posterior) expectation of a random variable Z (in our case $Z = s^2_{g,\mathbf{B}}$) given the knowledge of the random vector Y is defined as the "best prediction" (Searle *et al.*

1992; van der Vaart 2007) of the random variable Z. The best predictor

$$\mathrm{BP}\left(Z\right) := \mathbb{E}\left[Z \,|\, Y\right] \tag{19}$$

is the unique function $g_0(Y)$ that minimizes the mean square error of prediction

$$\mathbb{E}\left[\left(Z - g(Y)\right)^2\right]$$

over all functions in Y, i.e. the conditional expectation is the projection (closest element in a given set of functions) of Z onto the linear space of all functions in Y (Searle *et al.* 1992; van der Vaart 2007).
The best predictor in (19) is by definition an unbiased predictor for the random variable Z and $g_0(Y)$ maximizes the correlation $\mathrm{Cor}\big(Z, g(Y)\big)$, i.e. we can replace the target random variable Z by the best predictor defined in (19) in an optimal way (Searle *et al.* 1992). Instead of inferring the unobservable target random variable, we conduct inferences on the best predictor. Because the best predictor has realized in a given dataset ($Y = y$), it is estimable (Searle *et al.* 1992).

In the following, we introduce a novel approach of considering the frequentist best predictor instead of the unconditional expectation for the random additive genomic variance $s^2_{g,\mathbf{B}}$ in (12). We proceed according to (19) and define

$$\begin{aligned}\mathrm{BP}\left(s^2_{g,\mathbf{B}}\right) := \mathbb{E}\left[s^2_{g,\mathbf{B}} \,|\, y\right] &= \mathbb{E}\left[\frac{1}{n-1}\beta^\top \mathbf{X}^\top \mathbf{PBPX}\beta \,|\, y\right] \\ &= \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{X}^\top \mathbf{PBPX}\mathbb{E}\left[\beta\beta^\top \,|\, y\right]\Big)\end{aligned}$$

for the given phenotypic values y, because $\mathbf{X}$ is constant and therefore independent of y. The matrix of conditional second moments of the marker effects β is usually non-diagonal (contrary to $\mathbb{E}\left[\beta\beta^\top\right]$) and can be expressed as

$$\mathbb{E}\left[\beta\beta^\top \,|\, y\right] = \mu_{\beta|y}\mu_{\beta|y}^\top + \Sigma_{\beta|y}$$

using the BLUP $\mu_{\beta|y} := \mathbb{E}\left[\beta \,|\, y\right]$ of the random vector β and the variance-covariance matrix $\Sigma_{\beta|y} := \mathrm{Cov}\left(\beta \,|\, y\right)$ of β given the data y. Then, the best predictor equals

$$\begin{aligned}\mathrm{BP}\left(s^2_{g,\mathbf{B}}\right) &= \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{X}^\top \mathbf{PBPX}\Big[\mu_{\beta|y}\mu_{\beta|y}^\top + \Sigma_{\beta|y}\Big]\Big) \\ &= \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{B}\Big[\mu_{g|y}\mu_{g|y}^\top + \Sigma_{g|y}\Big]\Big)\end{aligned} \tag{20}$$

where the last equality holds because of the connection

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

of the BLUPs and the conditional variance-covariance matrices

$$\Sigma_{g|y} := \mathrm{Cov}\left(g \,|\, y\right) = \mathrm{Cov}\left(\mathbf{X}\beta \,|\, y\right) = \mathbf{X}\Sigma_{\beta|y}\mathbf{X}^\top\mathbf{P}.$$

in models (1) and (3), see also Appendix Genomic Best Linear Unbiased Prediction.

For the best predictor of the additive genomic variance in the current population we set $\mathbf{B} = \mathbb{1}_{n\times n}$ in (20) and obtain

$$W := \mathrm{BP}\left(s^2_g\right) = \mathrm{tr}\Big(\hat{\Sigma}_X\Big[\mu_{\beta|y}\mu_{\beta|y}^\top + \Sigma_{\beta|y}\Big]\Big) \tag{21}$$

in model (3) or

$$W = \frac{1}{n-1}\mathrm{tr}\Big(\mu_{g|y}\mu_{g|y}^\top + \Sigma_{g|y}\Big) \tag{22}$$

in the terminology of the equivalent model (1).

For the best predictor of the additive genomic variance in the base population we set $\mathbf{B} = \mathbf{R}^{-0.5}\mathbf{PR}^{-0.5}$ in (20) and obtain

$$W^* := \mathrm{BP}\left(s^2_{g^*}\right) = \mathrm{tr}\Big(\hat{\Sigma}_{X^*}\Big[\mu_{\beta|y}\mu_{\beta|y}^\top + \Sigma_{\beta|y}\Big]\Big) \tag{23}$$

in model (3) or

$$W^* = \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{PR}^{-0.5}\Big[\mu_{g|y}\mu_{g|y}^\top + \Sigma_{g|y}\Big]\mathbf{R}^{-0.5}\Big) \tag{24}$$

in the terminology of the equivalent model (1).

We emphasize that the best predictor of the additive genomic variance in the current population (W) as well as in the base population (W^*) includes the contribution of all elements of the sample variance-covariance matrix of marker genotypes ($\hat{\Sigma}_X$ or $\hat{\Sigma}_{X^*}$), and hence comprise LD information, contrary to the unconditional expectations V, V^*, and V_s^* of the additive genomic variance from the previous section.

Explicit formulae for the empirical best predictors (eBP) of the additive genomic variance as well as a formula for W_s^* (approximate approach using the GRM $\mathbf{G}$ for transformation to the base population) will be given in the Appendix Estimation of the Additive Genomic Variance in the REM. We compare the use of the unconditional expectation and the best predictor for the prediction of the random additive genomic variance in the REM in Tables 3 and 4 in the Appendix.

Statistical Analysis (Genomic Data)

For an illustration of the theoretical results of the previous sections we used the mice dataset that comes with the *R*-package BGLR (Perez and de los Campos 2014). The data originally stem from an experiment by Valdar *et al.* (2006a,b) in a mice population. The dataset contains the matrix $\mathbf{X}$ with values in $\{0, 1, 2\}$ of $p = 10346$ polymorphic marker genotypes that were measured in $n = 1814$ mice. The trait (n-vector y) under consideration was body length (BL). The relationship of the mice is recorded in the $n \times n$ pedigree matrix $\mathbf{R}$ and is used for the transformation to the base population.
Additionally, we used the publicly available historical wheat dataset that also comes with the *R*-package "BGLR" (Perez and de los Campos 2014). The data originally stems from CIMMYT's Global Wheat Program and consists of $n = 599$ lines of wheat where the trait under consideration was average grain yield. The phenotypes are divided up into four basic target sets of environments designated as Wheat I, Wheat II, Wheat III and Wheat IV where we only considered the first one. The dataset contains the matrix of marker genotypes for $p = 1279$ markers as well as a relationship matrix.
Moreover, we analyzed a population of $n = 1057$ fully sequenced Arabidopsis lines for which phenotypes and genotypes are publicly available by the effort of the Arabidopsis 1001 Genomes project (The 1001 Genomes Consortium 2016). The lines represent natural inbred lines and we examined the same trait, namely flowering time at 10°C (FT10), and the same $p = 193697$ SNP-markers that were used in Lehermeier *et al.*

(2017). For these data no relationship matrix was available.

For each dataset, we used the gBLUP-method in the equivalent version (computational advantages) implemented in the *R*-package "sommer" (Covarrubias-Pazaran 2017) to fit a REM. We worked with the option REML (restricted maximum likelihood) to obtain estimates ($\hat{\sigma}^2_g$ and $\hat{\sigma}^2_\varepsilon$) for the variance components. The method also returned estimates of the best predictor of the genomic effects $\mu_{g|y}$ and the their variance-covariance matrix $\Sigma_{\hat{\mu}_{g|y}}$.
We used this outcome for the estimation of the unconditional expecation V and the BP W of the additive genomic variance in the current population and the as well as the unconditional expectation V^* and the BP W^* for the additive genomic variance in the base population (except for the Arabidopsis dataset, where no relationship matrix was available). Although the GRM is not invertible, we will show in the Appendix Estimation of the Additive Genomic Variance in the REM how to use the GRM for a transformation to the base population, and to calculate the corresponding unconditional expectation V^*_s and the BP W^*_s for the additive genomic variance in the base population.

We conducted all calculations with the free software R (R Development Core Team 2017). Detailed information about the calculations as well as the programming code with its output is provided in the supplemental File S1.

Data Availability

The authors affirm that all data necessary for confirming the conclusions of this article are represented fully within the manuscript and the supplemental material that has been uploaded to figshare. Supplemental File S1 contains a detailed description of the estimation of the genomic variances for the gBLUP-method as well as the corresponding *R*-code and its output.

Results

In the first section of Table 2 we present the estimation results for the unconditional expectation V and the best predictor W for the additive genomic variance in the current population. In the mice and wheat dataset $\hat{V}$ exceeds $\widehat{W}$, whereas for the Arabidopsis data, the empirical best predictor is about double the size of the unconditional expectation.
The sample variance of the phenotypic values has been scaled to 1. The sum of $\hat{V}$ and the residual variance is larger than the phenotyic variance for the mice and wheat data but smaller for the Arabidopsis data. Technically, it is possible to define the heritability in two ways, namely with respect to the phenotypic variance and with respect to the sum of the additive genomic variance and the residual variance. The sum of the emprical best predictor $\widehat{W}$ and the residual variance, however, equals the scaled phenotypic variance of 1 remarkably exactly for all datasets considered.

In the second section of Table 2 we first present the estimation results for the unconditional expectation V^* and the best predictor W^* for the additive genomic variance in the base population using the given relationship matrices for the transformation. For the mice data, $\hat{V}^*$ and $\widehat{W}^*$ are similar to their analogons $\hat{V}$ and $\widehat{W}$ in the current population. For the wheat data, however, the estimated unconditional expectation and empirical best predictor in the base population are about five times larger than those in the current population and exceed the sample phenotypic variance in the current population. By this approach, it is not possible to define a heritbility in the base population because both the estimate of the residual variance and the phenotypic sample variance refer to the current population.
The estimation results for the unconditional expectation V^*_s and the best predictor W^*_s for the additive genomic variance in the base population using the GRM for the transformation differ from those using the given relationship matrices by a considerable amount. In the mice and wheat data, V^*_s is larger than W^*_s, whereas for the Arabidopsis data the empirical best predictor exceeds the estimated unconditional expectation. This conforms to the behavior of V and W in the current population.

Discussion

We have shown that commonly used estimators for the additive genomic variance in the REM with genomic marker data are based on the unconditional expectation of the random additive genomic variance. We have introduced a novel best prediction approach for the random additive genomic variance in both the current and the base population. In the following, we discuss several important implications.

Current and Base population

Common ways of estimating the additive genomic variance focus on the base population. These approaches are independent of the actual current population and consequently valid even if the generations change.
If one aims at the response of a population to selection, however, it might be more meaningful to estimate the additive genomic variance in the actual given population. This implies that the estimation of the genomic variance has to be conducted again when the individuals change. A formal definition of the heritability is best possible in the current population, where the phenotypic and residual variance are estimable.
We have preferred to use given relationship matrices for the transformation of the genomic values to the base population. In the case that such a matrix is not available, we have shown how to use genomic relationship matrices for the transformation, although a formal inversion of GRM's is in general not possible.

In Table 2 we have illustrated that we can decompose the sample phenotypic variance into the sum of the empirical best predictor of the additive genomic variance in the current population and into the estimated residual variance. This is due to the orthogonal projection property of the conditional expectation which gives the best approximation of the random additive genomic variance. This enables a unique definition of the heritability in the current population. It is never possible, however, to transfer the residual variance to the base population. Consequently, a definition of the heritability in the base population is not straight-forward.

The gBLUP-method and the Bayesian approach

The frequentist gBLUP-method can also be set-up in the context of Bayesian regression models (with prior distribution for the effect vector as defined in (5) and uninformative priors for the variance components). Lehermeier *et al.* (2017) considered the additive genomic variance s^2_g in the current population, see (7),

Table 2 Estimation results for the unconditional expectation V and the best predictor W for the additive genomic variance in the current population for the mice, wheat, and Arabidopsis datasets. We also present the corresponding heritabilities with respect to the sample variance of the phenotypic values and with respect to the sum of the additive genomic and residual variance σ_ε^2. In addition, we depict the estimation results for the unconditional expectation V^* (V_s^* when using the GRM for the transformation) and the best predictor W^* (W_s^* when using the GRM for the transformation) for the additive genomic variance in the base population.

Genom. Var. / Heritab. \ Data	Population	Mice	Wheat	Arabidopsis
$\hat{V}$ $(=\hat{h}_V^2)$ [a]	Current	0.3737749	0.6039708	0.47333803
$\hat{V}+\hat{\sigma}_\varepsilon^2$ [b]		1.0754963	1.1449704	0.54832029
$\tilde{h}_V^2 := \hat{V}/(\hat{V}+\hat{\sigma}_\varepsilon^2)$ [c]		0.3475371	0.5274990	0.86325098
$\widehat{W}$ $(=\hat{h}_W^2)$ [a]		0.2982787	0.4590001	0.92501779
$\widehat{W}+\hat{\sigma}_\varepsilon^2$ [b]		1.0000002	0.9999998	1.00000005
$\tilde{h}_W^2 := \widehat{W}/(\widehat{W}+\hat{\sigma}_\varepsilon^2)$ [c]		0.2982787	0.4590002	0.92501774
$\hat{V}^*$	Base	0.3704021	3.0621134	–
$\widehat{W}^*$		0.3089758	2.0095836	–
$\hat{V}_s^*$		0.3639248	1.3158006	0.80762011
$\widehat{W}_s^*$		0.3577692	1.2300300	1.30240520

[a] Heritability with respect to phenotypic sample variance $\hat{\sigma}_y^2$ which has been scaled to 1.
[b] Alternative definition of the phenotypic variance that depends on the estimate of the genomic variance.
[c] Alternative definition of the heritability that depends on the alternative definition of the phenotypic variance.

and used Bayesian ridge regression to estimate

$$M_2 := \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)}\right)^\top \hat{\Sigma}_X \hat{\beta}^{(m)},$$

where $\left(\hat{\beta}^{(m)}\right)_{m=1,\ldots,M}$ denotes MCMC samples from the posterior distribution of β. In that approach, Lehermeier *et al.* (2017) have estimated the posterior mean of the additive genomic variance s_g^2 in the current population. This approach is the Bayesian equivalent of the (frequentist) empirical version of the best predictor of s_g^2 in (12) in the current population. M_2 does not describe the genomic variance in the base population and should not directly be compared with approaches introduced e.g. in Yang *et al.* (2010, 2011). Analogously to the best predictor W^*, see (23), for the genomic variance in the base population, one can consider

$$M_2^* := \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)}\right)^\top \hat{\Sigma}_{X^*} \hat{\beta}^{(m)}$$

as the posterior mean of the genomic variance in the base population in Bayesian regression models.

The frequentist gBLUP-method provides a more formal approach to the prediction of the random additive genomic variance in linear models with random effects than the Bayesian approach. It enables the derivation of explicit formulas for the predictors (unconditional expectation and best predictor) of the random additive genomic variance using the standard output of the gBLUP-method which goes hand in hand with a fast implementation of the empirical version of the predictors. The connection between the BLUP $\mu_{\beta|y}$ and its covariance for the random marker effects with the additive genomic variance are clearly visible. This enables us, for instance, to derive the decomposition of the best predictor of the random additive genomic variance into the unconditional expectation and a function for the marker LD in the following section.

Influence of Linkage Disequilibrium

In Section Definitions of the Genomic Variance we have seen that the (random) additive genomic variance equals

$$\begin{aligned} s_g^2 &= \beta^\top \hat{\Sigma}_X \beta \\ &= \sum_{j=1}^{p} \beta_j^2 (\hat{\Sigma}_X)_{jj} + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p} \beta_i \beta_j (\hat{\Sigma}_X)_{ij} \end{aligned}$$

in the current population, and

$$\begin{aligned} s_{g^*}^2 &= \beta^\top \hat{\Sigma}_{X^*} \beta \\ &= \sum_{j=1}^{p} \beta_j^2 (\hat{\Sigma}_{X^*})_{jj} + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p} \beta_i \beta_j (\hat{\Sigma}_{X^*})_{ij} \end{aligned}$$

in the base population. We emphasize that the variance-covariance matrix of the marker genotypes (marker LD) plays a decisive part in the determination of the additive genomic variances in both the current and the base population. The variances s_g^2 and $s_{g^*}^2$ are structurally in accordance with the classical additive genetic variance (Bulmer 1971; Falconer and Mackay 1996), which is caused by the genotypes whereas the genotypic effects are fixed.

In the REM, however, the marker effects are random with unconditional expectation 0 and unconditional diagonal variance-covariance matrix with equal variances σ_β^2. As a consequence,

the unconditional expectation of the additive genomic variance

$$\mathbb{E}\left[s_g^2\right] = \sigma_\beta^2 \text{tr}\left(\hat{\Sigma}_X\right)$$

in the current population and

$$\mathbb{E}\left[s_{g^*}^2\right] = \sigma_\beta^2 \text{tr}\left(\hat{\Sigma}_{X^*}\right)$$

in the base population contain only the variances of the marker genotypes in the corresponding population. In addition, the unconditional expectation resembles both the variance of a randomly sampled trait for a randomly sampled individual and the variance of a randomly sampled trait for individual with fixed genotypes, see Table 1 for an overview.

We show in the Appendix Estimation of the Additive Genomic Variance in the REM that we can partition the best predictor of the random additive genomic variance $s_{g,\mathbf{B}}^2$ in the following way:

$$\text{BP}\left(s_{g,\mathbf{B}}^2\right) = \mathbb{E}\left[s_{g,\mathbf{B}}^2\right] + Z(y),$$

where

$$\begin{aligned} Z(y) = &\sum_{j=1}^{p} \left[\mu_{\beta|y}\mu_{\beta|y}^\top - \Sigma_{\mu_{\beta|y}}\right]_{jj} \left(\hat{\Sigma}_X\right)_{jj} \\ &+ \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p} \left[\mu_{\beta|y}\mu_{\beta|y}^\top - \Sigma_{\mu_{\beta|y}}\right]_{ij} \left(\hat{\Sigma}_X\right)_{ij} \end{aligned}$$

in the current population and

$$\begin{aligned} Z(y) = &\sum_{j=1}^{p} \left[\mu_{\beta|y}\mu_{\beta|y}^\top - \Sigma_{\mu_{\beta|y}}\right]_{jj} \left(\hat{\Sigma}_{X^*}\right)_{jj} \\ &+ \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p} \left[\mu_{\beta|y}\mu_{\beta|y}^\top - \Sigma_{\mu_{\beta|y}}\right]_{ij} \left(\hat{\Sigma}_{X^*}\right)_{ij} \end{aligned}$$

in the base population ($\Sigma_{\mu_{\beta|y}} = \text{Cov}(\mu_{\beta|y})$).
The best predictor, therefore, consists of the unconditional expectation of the additive genomic variance (no contribution of LD) and a function that explicitly contains the weighted contribution of marker LD. This function determines whether estimators like GCTA-GREML (unconditional expectation of the random genomic variance in the base population) are biased upwards or downwards, i.e. it determines the direction and the magnitude of the bias of GCTA-GREML (this method is based on the assumption that the function Z constantly equals 0). In addition, we notice that this bias does not depend only on the sign of the covariance between the marker genotypes, but on the sign and the magnitude of the weighted covariances.

We emphasize that, contrary to the unconditional expectation, the best predictor maintains the structure of the additive genomic variance s_g^2 and $s_{g^*}^2$, because the function Z can be decomposed into the weighted sample variances and covariances of the marker genotypes. Instead of the marker effects, the components of the matrix $\mu_{\beta|y}\mu_{\beta|y}^\top - \Sigma_{\mu_{\beta|y}}$, which is typically non-zero and non-diagonal, take the part of the weighting factors of the elements of $\hat{\Sigma}_X$ and $\hat{\Sigma}_{X^*}$. The best predictor maintains the structure of the additive genomic variance in both the current (s_g^2) and the base population ($s_{g^*}^2$) and thus conforms to the classical genetic variance (Bulmer 1971; Falconer and Mackay 1996).
The difference between the estimators V and W (V^* and W^*, V_s^* and W_s^*) is given by the estimated $Z(y)$ and can be obtained from Table 2. We notice that the weighted contribution of marker LD is large and positive in the case of the Arabidopsis data, whereas in the mice and wheat data the weighted contribution of marker LD is slightly negative.

To sum up, the application of the unconditional expectation of the additive genomic variance combined with the model assumptions on the marker effects in random effect models cause, at least partially, the missing contribution of LD to the estimated additive genomic variance. This goes hand in hand with the critique expressed in Kumar *et al.* (2015, 2016). It is, however, less important when estimating the additive genomic variance in the base population where the individuals are uncorrelated and less LD persists (although the marker genotypes need not be uncorrelated).
The best prediction approach eliminates the problem of the missing contribution of LD to the additive genomic variance that is caused by mathematical modeling (e.g. the assumptions in the random-effects model).

Concluding Remarks

The variability in the genomic values and with it the additive genomic variance, is induced by the marker genotypes. The main task in investigating the random additive genomic variance in the REM is to treat the additional randomness of the genomic variance that is induced by the randomness of the marker effects. We have shown that commonly used estimators use the unconditional expectation to handle this randomness. However, we recommend the use of the best prediction approach (conditional expectation) that uses the additional information given by the genomic data, minimizes the mean square error of prediction, includes the contribution of LD, and maintains the structure of the genomic equivalent of the classical additive genetic variance.

Acknowledgements

NS and MS were 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'. HPP was supported by the DFG project PI 377/18-1.
We would like to express our gratitude to Chris-Carolin Schön and Leo Dempfle for important remarks and advice on the paper. The remarks of anonymous reviewers helped to greatly improve the quality of this manuscript. We are also grateful to Henner Simianer, Torsten Pook and Jonas Brehmer for their comments on earlier versions of the manuscript.

Literature Cited

Bernardo, R., 1994 Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Science **34**: 20–25.
Bühlmann, P. and S. van de Geer, 2011 *Statistics for High-Dimensional Data*. Springer Series in Statistics.
Bulmer, M., 1971 The effect of selection on genetic variability. American Naturalist **105**: 201–211.

Corbeil, R. R. and S. R. Searle, 1976 Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Techometrics **18**: 31–38.

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

Das, K., J. Jiang, and J. Rao, 2004 Mean squared error of empirical predictor. The Annals of Statistics **32**: 818–840.

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, 2017a 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.

Fernando, R. and D. Garrick, 2013 *Genome-Wide Association Studies and Genomic Prediction*. Humana Press.

Fernando, R., A. Toosi, A. Wolc, D. Garrick, and J. Dekkers, 2017b Application of whole-genome prediction methods for genome-wide association studies: A bayesian approach. Journal of Agricultural, Biological and Environmental Statistics pp. 1–24.

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*. University of Guelph.

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**: 1–10.

Jiang, J., 1999 On unbiasedness of the empirical BLUE and BLUP. Statistics and Probability Letters **41**: 19–24.

Kackar, R. N. and D. A. Harville, 1984 Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the Americal Statistical Association **79**: 853–862.

Kotz, S., N. Balakrishnan, and N. L. Johnson, 2000 *Continuous Multivariate Distributions*. Wiley, Second edition.

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 Response to "commentary on limitations of GCTA as a solution to the missing heritability problem". bioRxiv: http://dx.doi.org/10.1101/039594.

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

Ould Estaghvirou, S. B., J. O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova, *et al.*, 2013 Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genomics **14**.

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

Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics .

R Development Core Team, 2017 *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.

Sorensen, D., R. Fernando, and D. Gianola, 2000 Inferring the trajectory of genetic variance in the course of artificial selection. Genetical Research **77**: 83–94.

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 1,135 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 complex 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 environmental effects on complex traits in mice. Genetics **174**: 959–984.

van der Vaart, A. W., 2007 *Asymptotic Statistics*. Cambridge University Press, 8th edition.

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

Vinkhuyzen, A. A. E., N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, 2014 Estimation and partitioning of heritability in human populations using whole genome analysis methods. Annual Review of Genetics **47**: 75–95.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, *et al.*, 2013 Analysis of egg production in layer chickens using a random regression model with genomic relationships. Poultry Science **92**: 1486–1491.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the 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.

Zhou, X., P. Carbonetto, and M. Stephens, 2013 Polygenic modeling with bayesian sparse linear mixed models. PLOS Genetics **9**.

Zhu, Z., A. Bakshi, A. A. E. 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

Genomic Best Linear Unbiased Prediction

In the REM ($\beta_j \sim \mathcal{N}(0, \sigma^2_\beta)$) for the model

$$y = \mu \mathbb{1}_n + \mathbf{P}\mathbf{X}\beta + \varepsilon \tag{3}$$

we have that

$$y \sim \mathcal{N}\Big(\mu, \underbrace{\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\sigma^2_\beta + \sigma^2_\varepsilon \mathbb{1}_{n\times n}}_{:=\tilde{\Sigma}^{-1}}\Big). \tag{25}$$

The marker effect vector β cannot be estimated because it is a random variable. Henderson (1984) introduced the concept of the prediction of β, which refers to the estimation of the realized values of the random effects. The best linear unbiased predictor (BLUP) $\mu_{\beta|y}$ for β is given by $\mu_{\beta|y} = \mathbb{E}[\beta \,|\, y]$ (Henderson 1984; Searle *et al.* 1992). The conditional expectation is the unique best predictor, i.e. it is unbiased and has minimal mean square error of prediction

$$\mathbb{E}\Big[(\beta - g(y))^\top (\beta - g(y))\Big]$$

within the whole set of functions g that depend on the data y (van der Vaart 2007).
The joint distribution of y and β equals

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

and we obtain

$$\beta|y \sim \mathcal{N}\Big(\sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}(y - \mu\mathbb{1}_n), \sigma^2_\beta \mathbb{1}_{p\times p} - \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta\Big), \tag{26}$$

see e.g. Kotz *et al.* (2000). Consequently, the BLUP equals

$$\mu_{\beta|y} = \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}(y - \mu\mathbb{1}_n) \tag{27}$$

and is linear in y. The conditional variance-covariance matrix of β equals

$$\Sigma_{\beta|y} := \mathrm{Cov}(\beta \,|\, y) \overset{(26)}{=} \sigma^2_\beta \mathbb{1}_{p\times p} - \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta, \tag{28}$$

and the variance-covariance matrix of the BLUP $\mu_{\beta|y}$ equals

$$\begin{aligned} \Sigma_{\mu_{\beta|y}} :&= \mathrm{Cov}\big(\mu_{\beta|y}\big) \\ &= \mathrm{Cov}\Big(\mathbb{E}\big[\beta \,|\, y\big]\Big) \\ &= \mathrm{Cov}\big(\beta\big) - \mathbb{E}\Big[\mathrm{Cov}\big(\beta \,|\, y\big)\Big] \\ &\overset{(28)}{=} \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta. \end{aligned} \tag{29}$$

The actual estimation of the parameters in model (3) with the BLUP-method is a two-stage procedure (Das *et al.* 2004). In the first stage, a BLUE for the fixed quantities and a BLUP for the random variables are derived. However, they involve the variance components σ^2_β and σ^2_ε as unknown parameters. In a second stage, these parameters are replaced by estimates, and the estimators for the BLUE and the BLUP are referred to as empirical BLUE (eBLUE) and empirical BLUP (eBLUP), see Kackar and Harville (1984); Jiang (1999). Investigations on the properties of the eBLUE and the eBLUP are very complex (Searle *et al.* 1992), and often only approximate results are obtained (Kackar and Harville 1984; Jiang 1999; Das *et al.* 2004).

Assume that we are provided with estimators for the variance components using e.g. restricted maximum likelihood (REML) (Patterson and Thompson 1971; Corbeil and Searle 1976; Searle *et al.* 1992). These estimated variance components are functions of the data y and consequently, the eBLUE

$$\hat{\mu} = \frac{\mathbb{1}_n^\top \Big(\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\hat{\sigma}^2_\beta + \hat{\sigma}^2_\varepsilon \mathbb{1}_{n\times n}\Big)^{-1} y}{\mathbb{1}_n^\top \Big(\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\hat{\sigma}^2_\beta + \hat{\sigma}^2_\varepsilon \mathbb{1}_{n\times n}\Big)^{-1} \mathbb{1}_n}$$

for the intercept and the eBLUP

$$\hat{\mu}_{\beta|y} = \hat{\sigma}^2_\beta \mathbf{X}^\top\mathbf{P}\Big(\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\hat{\sigma}^2_\beta + \hat{\sigma}^2_\varepsilon \mathbb{1}_{n\times n}\Big)^{-1}(y - \hat{\mu}), \tag{30}$$

for the marker effects β are not even linear in the data y anymore (despite their naming). The unbiasedness of the estimators eBLUE and the eBLUP can be asserted if the estimated variance components $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\varepsilon$ are non-negative, even functions in y, translation-invariant, and if the expectations of the eBLUE and eBLUP are finite (Kackar and Harville 1984). When using REML estimates for the variance components, these requirements are satisfied and the eBLUE $\hat{\mu}$ and the eBLUP $\hat{\mu}_{\beta|y}$ are bias-free estimators for μ and β (Jiang 1999).

Conditional on the estimation of the variance components (ignoring the randomness in the second stage of the estimation of the eBLUP), the variance-covariance matrix of the eBLUP $\hat{\mu}_{\beta|y}$ equals

$$\begin{aligned} \Sigma_{\hat{\mu}_{\beta|y}} :&= \mathrm{Cov}\Big(\hat{\mu}_{\beta|y} \,\Big|\, \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\Big) \\ &= \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathrm{Cov}\Big(y - \hat{\mu}\mathbb{1}_n \,|\, \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\Big)\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta \\ &= \sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta - \frac{\sigma^2_\beta \mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbb{1}_n\mathbb{1}_n^\top\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma^2_\beta}{\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n}, \end{aligned} \tag{31}$$

because

$$\begin{aligned} &\mathrm{Cov}\Big(y - \hat{\mu}\mathbb{1}_n \,\Big|\, \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\Big) = \\ &= \mathrm{Cov}\big(y\big) - 2\mathrm{Cov}\Big(y, \hat{\mu} \,\Big|\, \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\Big)\mathbb{1}_n^\top \\ &\quad + \mathbb{1}_n\mathrm{Cov}\Big(\hat{\mu} \,\Big|\, \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\Big)\mathbb{1}_n^\top \\ &= \tilde{\Sigma}^{-1} - 2\mathrm{Cov}\big(y\big)\frac{\tilde{\Sigma}\mathbb{1}_n}{\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n}\mathbb{1}_n^\top + \mathbb{1}_n\frac{\mathbb{1}_n^\top\tilde{\Sigma}\tilde{\Sigma}^{-1}\tilde{\Sigma}\mathbb{1}_n}{(\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n)^2}\mathbb{1}_n^\top \\ &= \tilde{\Sigma}^{-1} - \frac{\mathbb{1}_n\mathbb{1}_n^\top}{\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n} \end{aligned}$$

holds.

We can transfer the results derived in model (3) to model

$$y = \mu \mathbb{1}_n + g + \varepsilon \tag{1}$$

by using their equivalence in distribution. The genomic best linear unbiased predictor for g equals

$$\begin{aligned}
\mu_{g|y} := \mathbb{E}[g \mid y] &= \mathbb{E}\left[\mathbf{P}\mathbf{X}\beta \mid y\right] \\
&= \mathbf{P}\mathbf{X}\mu_{\beta|y} \\
&\overset{(27)}{=} \mathbf{P}\mathbf{X}\siga_\beta^2\mathbf{X}^\top\mathbf{P}\tilde{\Sigma}(y - \mu\mathbb{1}_n) \\
&\overset{(6)}{=} \sigma_g^2\mathbf{G}\left(\mathbf{G}\sigma_g^2 + \sigma_\varepsilon^2\mathbb{1}_{n\times n}\right)^{-1}(y - \mu\mathbb{1}_n). \qquad (32)
\end{aligned}$$

The conditional variance-covariance matrix of g is obtained as

$$\begin{aligned}
\Sigma_{g|y} &:= \mathrm{Cov}(g \mid y) \\
&= \mathbf{P}\mathbf{X}\mathrm{Cov}(\beta \mid y)\mathbf{X}^\top\mathbf{P} \\
&\overset{(28)}{=} \mathbf{P}\mathbf{X}\left(\sigma_\beta^2\mathbb{1}_{p\times p} - \sigma_\beta^2\mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma_\beta^2\right)\mathbf{X}^\top\mathbf{P} \\
&\overset{(6)}{=} \sigma_g^2\mathbf{G} - \sigma_g^2\mathbf{G}\left(\mathbf{G}\sigma_g^2 + \sigma_\varepsilon^2\mathbb{1}_{n\times n}\right)^{-1}\mathbf{G}\sigma_g^2, \qquad (33)
\end{aligned}$$

as well as the variance-covariance matrix of the BLUP $\mu_{g|y}$

$$\begin{aligned}
\Sigma_{\mu_{g|y}} &:= \mathrm{Cov}\left(\mu_{g|y}\right) \\
&= \mathbf{P}\mathbf{X}\mathrm{Cov}\left(\mu_{\beta|y}\right)\mathbf{X}^\top\mathbf{P} \\
&\overset{(29)}{=} \mathbf{P}\mathbf{X}\sigma_\beta^2\mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma_\beta^2\mathbf{X}^\top\mathbf{P} \\
&\overset{(6)}{=} \sigma_g^2\mathbf{G}\left(\mathbf{G}\sigma_g^2 + \sigma_\varepsilon^2\mathbb{1}_{n\times n}\right)^{-1}\mathbf{G}\sigma_g^2. \qquad (34)
\end{aligned}$$

The variance-covariance matrix of the eBLUP equals

$$\begin{aligned}
\Sigma_{\hat{\mu}_{g|y}} &:= \mathrm{Cov}\left(\hat{\mu}_{g|y} \,\middle|\, \sigma_g^2 = \hat{\sigma}_g^2, \sigma_\varepsilon^2 = \hat{\sigma}_\varepsilon^2\right) \\
&= \mathrm{Cov}\left(\mathbf{P}\mathbf{X}\hat{\mu}_{\beta|y} \,\middle|\, \sigma_\beta^2 = \hat{\sigma}_\beta^2, \sigma_\varepsilon^2 = \hat{\sigma}_\varepsilon^2\right) \\
&= \mathbf{P}\mathbf{X}\Sigma_{\hat{\mu}_{\beta|y}}\mathbf{X}^\top\mathbf{P} \\
&\overset{(31)}{=} \mathbf{P}\mathbf{X}\left[\sigma_\beta^2\mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma_\beta^2 - \frac{\sigma_\beta^2\mathbf{X}^\top\mathbf{P}\tilde{\Sigma}\mathbb{1}_n\mathbb{1}_n^\top\tilde{\Sigma}\mathbf{P}\mathbf{X}\sigma_\beta^2}{\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n}\right]\mathbf{X}^\top\mathbf{P} \\
&\overset{(6)}{=} \sigma_g^2\mathbf{G}\tilde{\Sigma}\mathbf{G}\sigma_g^2 - \frac{\sigma_g^2\mathbf{G}\tilde{\Sigma}\mathbb{1}_n\mathbb{1}_n^\top\tilde{\Sigma}\mathbf{G}\sigma_g^2}{\mathbb{1}_n^\top\tilde{\Sigma}\mathbb{1}_n}. \qquad (35)
\end{aligned}$$

Theoretical Variances of the Genomic Values in the REM

We review three different definitions of the theoretical variance of the genomic values in the REM (marker genotypes random, marker effects random, or both random). We focus the following analysis on the linear model (3) because of the explicit separation of marker genotypes and marker effects. For simplicity, we focus on the genomic variance in the current population. The results for the base population are obtained by replacing the data-generating process X with X^*.

Random Genotypes and Random Effects

If the marker genotypes as well as the marker effects are the source of genomic variation, we calculate the variance of the genomic value according to the law of total variance as:

$$\begin{aligned}
\mathrm{Var}(X\beta) &= \mathbb{E}\left[\mathrm{Var}(X\beta \mid \beta)\right] + \mathrm{Var}\left(\mathbb{E}[X\beta \mid \beta]\right) \\
&= \mathbb{E}\left[\beta^\top\Sigma_X\beta\right] + \mathrm{Var}\left(\mathbb{E}[X]\beta\right) \\
&= \mathrm{tr}\left(\Sigma_X\mathbb{E}\left[\beta\beta^\top\right]\right) + \mathbb{E}[X]\Sigma_\beta\mathbb{E}[X]^\top \\
&= \mathbb{E}[\beta]^\top\Sigma_X\mathbb{E}[\beta] + \mathrm{tr}(\Sigma_\beta\Sigma_X) + \mathbb{E}[X]\Sigma_\beta\mathbb{E}[X]^\top.
\end{aligned}$$

The unconditional expectation and the variance operator in the second line apply to the random marker effect vector β. Because of the model assumptions on the marker effects in (5) and the mean-centered marker genotypes ($\mathbb{E}[X] = 0$), we obtain

$$\mathrm{Var}(X\beta) = \sigma_\beta^2\mathrm{tr}(\Sigma_X) \tag{36}$$

with the interpretation as the variance of a randomly sampled (representative) individual for a trait with random effects.

Fixed Genotypes and Random Effects

If the genomic variation is caused by the marker effects only and the marker genotypes are fixed, then the n-vector of genomic values is normally distributed:

$$g = \mathbf{P}\mathbf{X}\beta \sim \mathcal{N}\left(0, \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\sigma_\beta^2\right). \tag{37}$$

In order to obtain an average theoretical variance of the individuals in the sample, we calculate the mean trace of the variance-covariance matrix of the genomic values:

$$\begin{aligned}
\frac{1}{n}\mathrm{tr}\left(\mathrm{Cov}\left(\mathbf{P}\mathbf{X}\beta\right)\right) &= \frac{1}{n}\sigma_\beta^2\mathrm{tr}\left(\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}\right) \\
&= \frac{n-1}{n}\sigma_\beta^2\mathrm{tr}\left(\hat{\Sigma}_X\right). \qquad (38)
\end{aligned}$$

This approximately equals the variance of a randomly sampled individual for a randomly sampled trait, see (36). Even when the marker genotypes are fixed, their sample variance-covariance matrix contributes to the theoretical variance of the genomic values when averaging over the individuals in the sample.

Random Genotypes and Fixed Effects

Probably the most common assumption on the nature of the genome is that the marker genotypes are random, whereas the marker effects are fixed (Falconer and Mackay 1996). In order to translate these assumption to the variance of the genomic values in the REM, we have to condition on the marker effects (i.e. we fix them). Then, the theoretical variance of the genomic values of a individual with random marker genotypes (representative individual) and with fixed marker effects equals

$$\begin{aligned}
\mathrm{Var}(X\beta \mid \beta) &= \beta^\top\Sigma_X\beta \\
&= \sum_{j=1}^{p}\beta_j\mathrm{Var}(X_j) + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ j\neq i}}^{p}\beta_i\beta_j\mathrm{Cov}(X_i, X_j), \qquad (39)
\end{aligned}$$

and describes the genomic equivalent of the definition of the additive genetic variance (Bulmer 1971; Falconer and Mackay 1996).

Estimation of the Additive Genomic Variance in the REM

In sections The Expectation of the Additive Genomic Variance and Best Prediction of the Additive Genomic Variance we have introduced ways to predict the random additive genomic variance

$$s^2_{g,\mathbf{B}} := \frac{1}{n-1} g^\top \mathbf{B} g = \frac{1}{n-1} \beta^\top \mathbf{X}^\top \mathbf{PBPX} \beta \tag{12}$$

in the REM, namely by using the unconditional expectation

$$\begin{aligned} \mathbb{E}\left[s^2_{g,\mathbf{B}}\right] &= \frac{1}{n-1}\sigma^2_\beta \mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\right) \\ &= \frac{1}{n-1}\sigma^2_g \mathrm{tr}\left(\mathbf{BG}\right) \end{aligned} \tag{13}$$

and the best predictor

$$\begin{aligned} \mathrm{BP}\left(s^2_{g,\mathbf{B}}\right) &= \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\mu_{\beta|y}\mu^\top_{\beta|y} + \Sigma_{\beta|y}\right]\right) \\ &= \frac{1}{n-1}\mathrm{tr}\left(\mathbf{B}\left[\mu_{g|y}\mu^\top_{g|y} + \Sigma_{g|y}\right]\right). \end{aligned} \tag{20}$$

In the following, we introduce estimators for these quantities and investigate their properties.

Estimation of the Unconditional Expectation

For any given positive semi-definite matrix $\mathbf{B}$, the unconditional expectation

$$\frac{1}{n-1}\sigma^2_\beta \mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\right)$$

in model (3) can be estimated by

$$\frac{1}{n-1}\hat{\sigma}^2_\beta \mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\right)$$

after having obtained an estimate $\hat{\sigma}^2_\beta$ of the variance component σ^2_β. In the equivalent model (1), we can estimate

$$\frac{1}{n-1}\sigma^2_g \mathrm{tr}\left(\mathbf{BG}\right)$$

by using

$$\frac{1}{n-1}\hat{\sigma}^2_g \mathrm{tr}\left(\mathbf{BG}\right)$$

for any positive semi-definite matrix $\mathbf{B}$.

The specification of $\mathbf{B}$ as in Section The Expectation of the Additive Genomic Variance leads to the explicit form of the estimators

$$\hat{V} = \hat{\sigma}^2_\beta \mathrm{tr}\left(\hat{\Sigma}_X\right) = \frac{1}{n-1}\hat{\sigma}^2_g \mathrm{tr}(\mathbf{G}),$$

$$\hat{V}^* = \hat{\sigma}^2_\beta \mathrm{tr}\left(\hat{\Sigma}_{X^*}\right) = \frac{1}{n-1}\hat{\sigma}^2_g \mathrm{tr}\left(\mathbf{PR}^{-0.5}\mathbf{GR}^{-0.5}\right)$$

and

$$\hat{V}^*_{\mathrm{s}} = c\hat{\sigma}^2_\beta = \hat{\sigma}^2_g. \tag{40}$$

Empirical Best Prediction (eBP)

Because of equalities (28) and (29) and the variance-covariance matrix $\Sigma_\beta = \sigma^2_\beta \mathbb{1}_{p\times p}$ of β we have that

$$\Sigma_{\beta|y} = \Sigma_\beta - \Sigma_{\mu_{\beta|y}} = \sigma^2_\beta \mathbb{1}_{p\times p} - \Sigma_{\mu_{\beta|y}}.$$

Consequently, the best predictor of $s^2_{g,\mathbf{B}}$ defined in (20) equals

$$\begin{aligned} \mathrm{BP}\left(s^2_{g,\mathbf{B}}\right) =& \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\mu_{\beta|y}\mu^\top_{\beta|y} + \Sigma_{\beta|y}\right]\right) \\ =& \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\mu_{\beta|y}\mu^\top_{\beta|y} + \Sigma_\beta - \Sigma_{\mu_{\beta|y}}\right]\right) \\ =& \frac{1}{n-1}\sigma^2_\beta \mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\right) \\ &+ \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\mu_{\beta|y}\mu^\top_{\beta|y} - \Sigma_{\mu_{\beta|y}}\right]\right) \\ =& \mathbb{E}\left[s^2_{g,\mathbf{B}}\right] + Z(y), \end{aligned} \tag{41}$$

where

$$\begin{aligned} Z(y) :=& \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\mu_{\beta|y}\mu^\top_{\beta|y} - \Sigma_{\mu_{\beta|y}}\right]\right) \\ =& \frac{1}{n-1}\mathrm{tr}\left(\mathbf{B}\left[\mu_{g|y}\mu^\top_{g|y} - \Sigma_{\mu_{g|y}}\right]\right). \end{aligned} \tag{42}$$

We have partitioned the best predictor of $s^2_{g,\mathbf{B}}$ into the unconditional expectation of $s^2_{g,\mathbf{B}}$ and the random variable Z which is realized in the phenotypic data y. The random variable Z specifies the adaption of the best predictor to the data and incorporates the contribution of (marker) LD. The expectation of Z over all possible data y is 0 because

$$\begin{aligned} \mathbb{E}\left[\mu_{\beta|y}\mu^\top_{\beta|y} - \Sigma_{\mu_{\beta|y}}\right] &= \mathrm{Cov}\left(\mu_{\beta|y}\right) + \mathbb{E}\left[\mu_{\beta|y}\right]\mathbb{E}\left[\mu_{\beta|y}\right]^\top - \Sigma_{\mu_{\beta|y}} \\ &= 0. \end{aligned}$$

The sign of the realization of Z determines whether the best predictor is larger (positive weighted LD) or smaller (negative weighted LD) than the unconditional expectation.

The task of finding an eBP for $s^2_{g,\mathbf{B}}$ is reduced to estimating the realized values of Z because of the connection derived in (41). We replace the BLUP and their variance-covariance matrix in equation (42) by the eBLUP and its estimated variance-covariance matrix:

$$\hat{Z}(y) := \frac{1}{n-1}\mathrm{tr}\left(\mathbf{X}^\top \mathbf{PBPX}\left[\hat{\mu}_{\beta|y}\hat{\mu}^\top_{\beta|y} - \hat{\Sigma}_{\hat{\mu}_{\beta|y}}\right]\right).$$

We assume that we are provided with REML-estimators $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\varepsilon$ for the variance components. Then, we find

$$\begin{aligned} \mathbb{E}\left[\hat{\mu}_{\beta|y}\hat{\mu}^\top_{\beta|y}\right] &= \mathrm{Cov}\left(\hat{\mu}_{\beta|y}\right) + \mathbb{E}\left[\hat{\mu}_{\beta|y}\right]\mathbb{E}\left[\hat{\mu}_{\beta|y}\right]^\top \\ &\overset{(31)}{=} \Sigma_{\hat{\mu}_{\beta|y}}, \end{aligned}$$

because the eBLUP is unbiased for β, i.e. $\mathbb{E}[\hat{\mu}_{\beta|y}] = \mathbb{E}[\beta] = 0$ (Jiang 1999). Unfortunately, the unbiasedness of the estimated variance-covariance matrix of the eBLUP can only be asserted in a trivial way by conditioning on the estimated variance components:

$$\mathbb{E}\left[\hat{\Sigma}_{\hat{\mu}_{\beta|y}} \mid \sigma^2_\beta = \hat{\sigma}^2_\beta, \sigma^2_\varepsilon = \hat{\sigma}^2_\varepsilon\right] \overset{(31)}{=} \Sigma_{\hat{\mu}_{\beta|y}}.$$

Therefore, the expectation of $\hat{Z}(y)$ (conditionally on the variance components) equals 0.
The same holds true for

$$\hat{Z}(y) = \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{B}\Big[\hat{\mu}_{g|y}\hat{\mu}_{g|y}^\top - \hat{\Sigma}_{\hat{\mu}_{g|y}}\Big]\Big)$$

in the equivalent model, because the quantities $\hat{\mu}_{g|y}$ and $\hat{\Sigma}_{\hat{\mu}_{g|y}}$ are linear combinations of $\hat{\mu}_{\beta|y}$ and $\hat{\Sigma}_{\hat{\mu}_{g|y}}$, see (32) and (35).

Altogether, we can define the unbiased (conditionally on the estimated variance components) empirical best predictor

$$\begin{aligned}\mathrm{eBP}\left(s^2_{g,\mathbf{B}}\right) :=& \widehat{\mathbb{E}}\left[s^2_{g,\mathbf{B}}\right] + \hat{Z}(y)\\ =& \frac{1}{n-1}\hat{\sigma}^2_\beta \mathrm{tr}\Big(\mathbf{X}^\top\mathbf{PBPX}\Big)\\ &+ \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{X}^\top\mathbf{PBPX}\Big[\hat{\mu}_{\beta|y}\hat{\mu}_{\beta|y}^\top - \hat{\Sigma}_{\hat{\mu}_{\beta|y}}\Big]\Big)\\ =& \frac{1}{n-1}\hat{\sigma}^2_g \mathrm{tr}\Big(\mathbf{BG}\Big) + \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{B}\Big[\hat{\mu}_{g|y}\hat{\mu}_{g|y}^\top - \hat{\Sigma}_{\hat{\mu}_{g|y}}\Big]\Big)\end{aligned}$$

for the additive genomic variance $s^2_{g,\mathbf{B}}$.
The specification of $\mathbf{B}$ as in Section Best Prediction of the Additive Genomic Variance leads to the explicit form

$$\begin{aligned}\widehat{W} :=& \mathrm{eBP}\left(s^2_g\right) = \hat{V} + \mathrm{tr}\Big(\hat{\Sigma}_X\Big[\hat{\mu}_{\beta|y}\hat{\mu}_{\beta|y}^\top - \hat{\Sigma}_{\hat{\mu}_{\beta|y}}\Big]\Big)\\ &= \hat{V} + \frac{1}{n-1}\mathrm{tr}\Big(\Big[\hat{\mu}_{g|y}\hat{\mu}_{g|y}^\top - \hat{\Sigma}_{\hat{\mu}_{g|y}}\Big]\Big)\end{aligned}$$

of the eBP for the additive genomic variance in the current population, and to the eBP

$$\begin{aligned}\widehat{W}^* :=& \mathrm{eBP}\left(s^2_{g^*}\right)\\ =& \hat{V}^* + \mathrm{tr}\Big(\hat{\Sigma}_{X^*}\Big[\hat{\mu}_{\beta|y}\hat{\mu}_{\beta|y}^\top - \hat{\Sigma}_{\hat{\mu}_{\beta|y}}\Big]\Big)\\ =& \hat{V}^* + \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{PR}^{-0.5}\Big[\hat{\mu}_{g|y}\hat{\mu}_{g|y}^\top - \hat{\Sigma}_{\hat{\mu}_{g|y}}\Big]\mathbf{R}^{-0.5}\Big)\end{aligned}$$

for the additive genomic variance in the base population.

Using the GRM $\mathbf{G}$ for a transformation to the base population is not well-defined because $\mathbf{G}$ is singular. However, because $\hat{V}^*_\mathrm{s}$, see (40), is commonly used, we want to find an analogous formula for the empirical best predictor in this set-up.
Instead of calculating

$$\mathbf{G}^{-0.5}\hat{\mu}_{g|y} = \mathbf{G}^{-0.5}\hat{\sigma}^2_g\mathbf{G}\Big(\mathbf{G}\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon\mathbb{1}_{n\times n}\Big)^{-1}(y - \hat{\mu}\mathbb{1}_n)$$

and

$$\mathbf{G}^{-0.5}\hat{\Sigma}_{\hat{\mu}_{g|y}} = \mathbf{G}^{-0.5}\left[\hat{\sigma}^2_g\mathbf{G}\hat{\tilde{\Sigma}}\mathbf{G}\hat{\sigma}^2_g - \frac{\hat{\sigma}^2_g\mathbf{G}\hat{\tilde{\Sigma}}\mathbb{1}_n\mathbb{1}_n^\top\hat{\tilde{\Sigma}}\mathbf{G}\hat{\sigma}^2_g}{\mathbb{1}_n^\top\hat{\tilde{\Sigma}}\mathbb{1}_n}\right]\mathbf{G}^{-0.5},$$

we use

$$\hat{\mu}^*_{g|y} := \hat{\sigma}^2_g\mathbf{G}^{0.5}\Big(\mathbf{G}\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon\mathbb{1}_{n\times n}\Big)^{-1}(y - \hat{\mu}\mathbb{1}_n)$$

and

$$\hat{\Sigma}^*_{\hat{\mu}_{g|y}} := \hat{\sigma}^2_g\mathbf{G}^{0.5}\hat{\tilde{\Sigma}}\mathbf{G}^{0.5}\hat{\sigma}^2_g - \frac{\hat{\sigma}^2_g\mathbf{G}^{0.5}\hat{\tilde{\Sigma}}\mathbb{1}_n\mathbb{1}_n^\top\hat{\tilde{\Sigma}}\mathbf{G}^{0.5}\hat{\sigma}^2_g}{\mathbb{1}_n^\top\hat{\tilde{\Sigma}}\mathbb{1}_n}$$

as substitutes. Then, we define

$$\widehat{W}^*_\mathrm{s} = \hat{V}^*_\mathrm{s} + \frac{1}{n-1}\mathrm{tr}\Big(\mathbf{P}\Big[\hat{\mu}^*_{g|y}(\hat{\mu}^*_{g|y})^\top - \hat{\Sigma}^*_{\hat{\mu}_{g|y}}\Big]\Big)$$

as an approximation of the empirical best predictor of the additive genomic variance in the base population when using the GRM for the transformation.

Table 3 Overview of Prediction Approaches for the Random Additive Genomic Variance in the Random-Effects Model with the gBLUP-method ($s^2_{g,\mathbf{B}} = \frac{1}{n-1}\beta^\top \mathbf{X}^\top \mathbf{PBPX}\beta$). $\mathbf{X}$ is the matrix of marker genotypes, $\mathbf{P}$ the matrix for column-wise mean-centering, $\mathbf{B}$ a positive semi-definite matrix, σ^2_β the variance component of the marker effects β, $\mu_{\beta|y}$ the BLUP of β, $\Sigma_{\beta|y}$ the conditional covariance matrix of β given the phenotypic data y, $\mathbf{R}$ a relationship matrix, $\hat{\Sigma}_X$ the sample variance-covariance matrix of the marker genotypes in the current population, $\hat{\Sigma}_{X^*}$ the sample variance-covariance matrix of the marker genotypes in the base population.

	Unconditional Expectation	**Best Prediction**
General Formula	$\mathbb{E}[s^2_{g,\mathbf{B}}] = \frac{1}{n-1}\sigma^2_\beta \mathrm{tr}(\mathbf{X}^\top \mathbf{PBPX})$	$\mathrm{BP}(s^2_{g,\mathbf{B}}) = \frac{1}{n-1}\mathrm{tr}(\mathbf{X}^\top \mathbf{PBPX}[\mu_{\beta\|y}\mu^\top_{\beta\|y} + \Sigma_{\beta\|y}])$
Current population	$V = \sigma^2_\beta \mathrm{tr}(\hat{\Sigma}_X)$	$W = \mathrm{tr}(\hat{\Sigma}_X[\mu_{\beta\|y}\mu^\top_{\beta\|y} + \Sigma_{\beta\|y}])$
Base Population	$V^* = \sigma^2_\beta \mathrm{tr}(\hat{\Sigma}_{X^*})$	$W^* = \mathrm{tr}(\hat{\Sigma}_{X^*}[\mu_{\beta\|y}\mu^\top_{\beta\|y} + \Sigma_{\beta\|y}])$
Features	• Best approximation of the additive genomic variance in the absence of information • No inclusion of LD	• Best approximation of the additive genomic variance using additional information given by phenotypic values (adaptation to the data) • Explicit inclusion of LD • Orthogonal decomposition of the phenotypic variance in the current population (unique definition of the heritability) • Genomic equivalent of the additive genetic variance

Table 4 Overview of Prediction Approaches for the Random Additive Genomic Variance in the Random-Effects Model with the gBLUP-method in the equivalent version of the linear model ($s^2_{g,\mathbf{B}} = \frac{1}{n-1}g^\top \mathbf{B}g$). $\mathbf{G}$ is the genomic relationship matrix, $\mathbf{P}$ the matrix for column-wise mean-centering,$\mathbf{B}$ a positive semi-definite matrix, σ^2_g the variance component of the genomic values g, $\mu_{g|y}$ the BLUP of g, $\Sigma_{g|y}$ the conditional covariance matrix of g given the phenotypic data y, $\mathbf{R}$ a relationship matrix.

	Unconditional Expectation	**Best Prediction**
General Formula	$\mathbb{E}[s^2_{g,\mathbf{B}}] = \frac{1}{n-1}\sigma^2_g \mathrm{tr}(\mathbf{BG})$	$\mathrm{BP}(s^2_{g,\mathbf{B}}) = \frac{1}{n-1}\mathrm{tr}(\mathbf{B}[\mu_{g\|y}\mu^\top_{g\|y} + \Sigma_{g\|y}])$
Current population	$V = \frac{1}{n-1}\sigma^2_g \mathrm{tr}(\mathbf{G})$	$W = \frac{1}{n-1}\mathrm{tr}([\mu_{g\|y}\mu^\top_{g\|y} + \Sigma_{g\|y}])$
Base Population	$V^* = \frac{1}{n-1}\sigma^2_g \mathrm{tr}(\mathbf{PR}^{-0.5}\mathbf{GR}^{-0.5})$	$W^* = \frac{1}{n-1}\mathrm{tr}\left(\mathbf{PR}^{-0.5}[\mu_{g\|y}\mu^\top_{g\|y} + \Sigma_{g\|y}]\mathbf{R}^{-0.5}\right)$
Features	• Best approximation of the additive genomic variance in the absence of information • No inclusion of LD • Transformation with GRM: σ^2_g replaces V^*	• Best approximation of the additive genomic variance using additional information given by phenotypic values (adaptation to the data) • Explicit inclusion of marker LD • Orthogonal decomposition of the phenotypic variance in the current population (unique definition of the heritability) • Genomic equivalent of the additive genetic variance