Received: 22 October 2018 | Revised: 3 December 2018 | Accepted: 6 December 2018

DOI: 10.1111/jbg.12378

ORIGINAL ARTICLE

WILEY Journal of Animal Breeding and Genetics

A certain invariance property of BLUE in a whole-genome regression context

Daniel Gianola[1,2] | Rohan L. Fernando[2] | Dorian J. Garrick[3]

[1]Department of Animal Science, Iowa State University, Ames, Iowa

[2]Departments of Animal Sciences and Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin

[3]AL Rae Centre of Genetics and Breeding, Massey University, Palmerston North, New Zealand

Correspondence
Daniel Gianola, Department of Animal Sciences, University of Wisconsin, Madison, WI.
Email: gianola@ansci.wisc.edu

Abstract

A curious result from mixed linear models applied to genome-wide association studies was expanded. In particular, a model in which one or more markers are considered as fixed but are allowed to contribute to the covariance structure by treating such markers as random as well was examined. The best linear unbiased estimator of marker effects is invariant with respect to whether those markers are employed in constructing a genomic relationship matrix or are ignored, provided marker effects are uncorrelated with those not being tested. Also, the implications of regarding some marker effects as fixed when, in fact, these possess a non-trivial covariance structure with those declared as random were examined.

1 | INTRODUCTION

In genome-wide association (GWAS) studies (e.g., Manolio et al., 2009), an objective is to find statistical connections between molecular markers and genomic regions affecting some complex trait. A linear regression model including marker genotype codes as covariates is used, and the simplest version fits a single marker regression via ordinary least squares (OLS).

If aggregation or clustering due to familial or molecular similarity exists in the data, a better estimation approach is generalized least squares (GLS), as it poses a more general covariance structure than OLS (Aulchenko, de Köning, & Haley, 2007; Gianola, Fariello, Naya, & Schön, 2016; Hoffman, 2013; Kang et al., 2008; Listgarten et al., 2012; Yang, Zaitlen, Goddard, Visscher, & Price, 2014). One such structure, for example, results from declaring all or a subset of marker effects as random variables, for example, assuming that $\beta_j \sim N\left(0,\sigma^2_\beta\right), j = 1,2, \ldots, p$, with all markers in the set taken as independently and identically distributed random variables. A random effects specification induces marker-based measures of similarity among individuals called molecular "relationship" or "kinship" matrices (**G**). The marker (s) evaluated for association is (are) treated as a fixed effect (s), and a test of nullity of effects on a trait is based on well-established procedures.

Should the marker (s) being tested be included or excluded when building **G**? A priori, if a marker is declared random it could not be fixed and vice versa. Including the contribution of a marker to **G** while declaring it as fixed constitutes a form of "double counting." When the number of markers (p) is very large and a single marker regression is used, the impact on **G** of including or removing the marker is tiny. Listgarten et al. (2012) suggest that markers being tested should be removed from **G**, followed by a concomitant re-estimation of necessary variance components at each instance of testing. This approach is computationally taxing, especially when p is huge, as it is the case with DNA sequence data. In many situations, it may be reasonable to assume that variance component estimates are affected only mildly by including or excluding the tested marker in **G**. For many complex traits in animal and plant breeding, each of the numerous markers in a chip has a small effect on both mean and variance of the data distribution.

Gianola et al. (2016) showed that the best linear unbiased estimator (BLUE, also GLS) of the fixed effect of a marker (or sets of markers) examined in GWAS is invariant with respect to whether or not the marker (s) tested for association is (are) included in the construction of **G**, provided that variance components are assumed constant.

This short communication expands on the preceding, as follows. First, we provide an expression that gives the variance–covariance matrix of the BLUE of each of the marker effects being tested using a simple adjustment. Second, it is shown that the best linear unbiased predictor (BLUP) of effects treated both as fixed and random is exactly zero, provided that no covariance exists between such effects and other marker effects treated as random in the model. Third, it is shown that if such covariance is not null, the fixed effects of a set of markers affect phenotypes through direct and indirect paths, and over and above the impact of linkage disequilibrium captured by columns of the matrix of genotype codes.

2 | MODEL

The linear regression model (assume that nuisance location effects have been eliminated somehow) used in GWAS is often posed as

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypes, $\mathbf{X}$ is an $n \times p$ matrix of marker genotype codes, $\boldsymbol{\beta}$ is a vector of p allelic substitution effects and $\mathbf{e} \sim N\left(\mathbf{0},\mathbf{I}\sigma_e^2\right)$ is a residual vector where σ_e^2 is the variance of the distribution of model residuals. Let $\mathbf{X}=[\mathbf{X}_1 \quad \mathbf{X}_2]$, where $\mathbf{X}_1$ is $n \times p_1$ (without loss of generality assume that $\mathbf{X}_1$ has full column rank), and $\mathbf{X}_2$ is $n \times p_2$ where p_2 may be much larger than n.

An equivalent representation of 1 is

$$\mathbf{y}=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\mathbf{e}, \tag{2}$$

and consider two alternative covariance structures for the phenotypes. The first structure results from treating $\boldsymbol{\beta}_1$ as a fixed vector and assuming $\boldsymbol{\beta}_2 \sim N\left(\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$:

$$\mathbf{V}_2=\mathbf{X}_2\mathbf{X}_2'\sigma_\beta^2+\mathbf{I}\sigma_e^2. \tag{3}$$

The GLS estimate of the fixed effect $\boldsymbol{\beta}_1$ under $\mathbf{V}_2$ is

$$\hat{\boldsymbol{\beta}}_1=(\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{y}, \tag{4}$$

The second covariance structure stems from treating $\boldsymbol{\beta}_1$ as random, with the $\boldsymbol{\beta} \sim N\left(\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$ assumption assigned to all marker effects, but then $\boldsymbol{\beta}_1$ is estimated as if it were fixed. Here, the phenotypic covariance matrix is

$$\mathbf{V}_{12}=\mathbf{X}_1\mathbf{X}_1'\sigma_\beta^2+\mathbf{X}_2\mathbf{X}_2'\sigma_\beta^2+\mathbf{I}\sigma_e^2, \tag{5}$$

with the GLS estimator of $\boldsymbol{\alpha}$, the fixed effect corresponding to $\boldsymbol{\beta}_1$ being

$$\hat{\boldsymbol{\alpha}}=(\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{y}. \tag{6}$$

Note that $\mathbf{V}_2=\mathbf{V}_{12}-\mathbf{X}_1\mathbf{X}_1'\sigma_\beta^2$. Arrays $\mathbf{X}_2\mathbf{X}_2'\sigma_\beta^2=\mathbf{S}_2$ and $\mathbf{X}_1\mathbf{X}_1'\sigma_\beta^2+\mathbf{X}_2\mathbf{X}_2'\sigma_\beta^2=\mathbf{S}_{12}$ can be referred to as “similarity” matrices, as in Listgarten et al. (2012).

3 | INVARIANCE PROPERTIES

3.1 | Best linear unbiased estimation

Gianola et al. (2016) showed that 4 and 6 are identical; the proof is presented in more detail here. To show this, we employ a model representation where the effects of one or more loci with genotypes in $\mathbf{X}_1$ are regarded as possessing both fixed and random effects:

$$\mathbf{y}=\mathbf{X}_1\left(\boldsymbol{\alpha}+\boldsymbol{\beta}_1\right)+\mathbf{X}_2\boldsymbol{\beta}_2+\mathbf{e}=X_1\boldsymbol{\alpha}+\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\mathbf{e}, \tag{7}$$

where $\boldsymbol{\alpha}$ are the fixed effects of markers in $\mathbf{X}_1$, with $\boldsymbol{\beta}_1 \sim N\left(\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$ and $\boldsymbol{\beta}_2 \sim N\left(\mathbf{0},\mathbf{I}\sigma_\beta^2\right)$, as before.

Using the Sherman–Morrison–Woodbury identity (Seber & Lee, 2003),

$$\mathbf{V}_2^{-1}=\mathbf{V}_{12}^{-1}+\sigma_\beta^2\mathbf{V}_{12}^{-1}\mathbf{X}_1(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}. \tag{8}$$

Using 8, $\mathbf{X}_1'\mathbf{V}_2^{-1}$ can be written as:

$$\begin{aligned}\mathbf{X}_1'\mathbf{V}_2^{-1}&=\mathbf{X}_1'\mathbf{V}_{12}^{-1}+\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\\&=[\mathbf{I}+\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}]\mathbf{X}_1'\mathbf{V}_{12}^{-1}\\&=[\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1+\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1](\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\\&=(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}.\end{aligned} \tag{9}$$

Now, using 9,

$$\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{X}_1=(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1, \tag{10}$$

and $\mathbf{X}'\mathbf{V}_2^{-1}\mathbf{y}$ is

$$\mathbf{X}'\mathbf{V}_2^{-1}\mathbf{y}=(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{y}. \tag{11}$$

From 10 and 11, the GLS estimator $\hat{\boldsymbol{\beta}}_1$ given in 4 is formed as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1&=(\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{y}\\&=\left[(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1\right]^{-1}(\mathbf{I}-\sigma_\beta^2\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{y}\\&=(\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{y}\\&=\hat{\boldsymbol{\alpha}},\end{aligned} \tag{12}$$

where $\hat{\boldsymbol{\alpha}}$, also given in 6, is the GLS estimator resulting from 7.

The preceding shows that the estimator obtained under covariance structure $\mathbf{V}_2$ is identical to the estimator resulting from structure $\mathbf{V}_{12}$. The practical implication is that the same similarity matrix, $\mathbf{S}_{12}$, can be used for conducting either single marker or sets of markers GWAS studies using linear regression models, provided that σ^2_β is assumed known and kept constant (as well as the residual variance) throughout.

Note that the sampling variance–covariance matrix of the estimates of the fixed effects must be taken under $\mathbf{V}_2$, that is, $Var\left(\hat{\boldsymbol{\beta}}_1\right) = (\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{X}_1)^{-1}$. Since $\mathbf{X}_1$ typically has one or a few columns in GWAS, advantage can be taken of 8 for computing $Var\hat{\boldsymbol{\beta}}_1$), as $\mathbf{V}_{12}^{-1}$ is obtained only once, whereas $\mathbf{V}_2^{-1}$ changes with the set of markers included in $\mathbf{X}_1$ and, therefore, in $\mathbf{X}_2$. Further, note from 9 that

$$\begin{aligned} Var\left(\hat{\boldsymbol{\beta}}_1\right) = (\mathbf{X}_1'\mathbf{V}_2^{-1}\mathbf{X}_1)^{-1} = \left[(\mathbf{I}-\sigma^2_\beta\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1\right]^{-1} \\ = \left(\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1\right)^{-1} - \mathbf{I}\sigma^2_\beta, \end{aligned} \tag{13}$$

so $\mathbf{V}_{12}^{-1}$ can be used throughout. The preceding representation illustrates the over-statement of uncertainty and "loss of power" incurred by use of $\left(\mathbf{X}_1'\mathbf{V}_{12}^{-1}\mathbf{X}_1\right)^{-1}$ instead of 13. Yang et al. (2014) present a related discussion and recommend that markers in close linkage disequilibrium with the target marker(s) be removed when building $\mathbf{G}$. Their approach requires re-estimation of variance components at every instance of testing. Our results do not apply under such strategy, as the variance–covariance structure would be expected to change over markers tested (Listgarten et al., 2012; Yang et al., 2014). An alternative could be to include the marker tested and some neighbours in close linkage disequilibrium in $\mathbf{X}_1$, provided that $p_1 < n$ and that no rank deficiency accrues, and then use all markers when building $\mathbf{G}$. A disadvantage of the alternative is the potentially strong collinearity in the set of markers in $\mathbf{X}_1$, producing unstable estimates with large sampling variances.

A caveat must be mentioned. In a genomic best linear unbiased prediction setting (e.g., Legarra, 2016; Van Raden, 2008), a similarity matrix under $\mathbf{V}_{12}$ can be constructed as

$$\begin{aligned} \mathbf{S}_{12} &= \frac{1}{p}\left(\mathbf{X}_1\mathbf{X}_1' + \mathbf{X}_2\mathbf{X}_2'\right)p\sigma^2_\beta \\ &= \mathbf{G}_{12}\sigma^2_g, \end{aligned} \tag{14}$$

where $\sigma^2_g = p\sigma^2_\beta$ is the "genomic variance" captured by all available markers; $\mathbf{G}_{12}$ is known as the genomic relationship matrix (Van Raden, 2008). Accordingly,

$$\begin{aligned} \mathbf{S}_2 &= \frac{1}{p_2}\left(\mathbf{X}_2\mathbf{X}_2'\right)p_2\sigma^2_\beta \\ &= \mathbf{G}_2\sigma^2_{g'}, \end{aligned} \tag{15}$$

where $\sigma^2_{g'} = p_2\sigma^2_\beta$ is the genomic variance marked by the variants included in $\mathbf{X}_2$. Clearly, two maximum-likelihood analyses of variance components, one with a set markers fixed and removed from the covariance structure, and the other one with all markers contributing to similarity, will produce different estimates and interpretations of genomic variance.

Estimates of σ^2_β must always be interpreted with great care. In a standard random effects model, the variance among effects of levels of a random factor represents a population parameter, with maximum-likelihood estimates of variance components interpreted accordingly. For example, if the random factor is the effect of a paternal half-sib family (a situation known as a "sire" model in animal breeding), the variance among sires, σ^2_s, say, has the same interpretation irrespective of whether the number of families is 10 or 10,000. However, in a marker-based model with $n < p$, the meaning and estimates of σ^2_β depend crucially on p, as the variance component acts then as a regularization parameter. In the $n < p$ situation, it is typically the case that estimates of σ^2_β decrease as p increases, and the rate of decrease in σ^2_β is critical for interpretation of estimates of marker effects when $p \to \infty$ (Gianola, 2013; León-Novelo & Casella, 2012).

3.2 | Best linear unbiased prediction

The best linear unbiased predictor of $\boldsymbol{\beta}_1$ in model 7 is

$$\begin{aligned} \vec{\boldsymbol{\beta}}_1 &= Cov\left(\boldsymbol{\beta}_1,\mathbf{y}'\right)\mathbf{V}_{12}^{-1}\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\alpha}}\right) \\ &= \sigma^2_\beta\mathbf{X}_1'\mathbf{V}_{12}^{-1}\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\alpha}}\right) \\ &= \mathbf{0} \end{aligned} \tag{16}$$

with $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_1$ calculated as in 4 or 6. The previous result follows because $\mathbf{X}_1\mathbf{V}_{12}^{-1}\mathbf{X}_1\hat{\boldsymbol{\alpha}} = \mathbf{X}_1\mathbf{V}_{12}^{-1}\mathbf{y}$ are the GLS equations, implying that $\mathbf{X}_1'\mathbf{V}_{12}^{-1}\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\alpha}}\right) = 0$.

Henderson's mixed model equations (MME) can be employed to verify result 16. For model 7 the MME are as follows:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_1+\mathbf{I}\lambda & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2+\mathbf{I}\lambda \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \vec{\boldsymbol{\beta}}_1 \\ \vec{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_{12}'\mathbf{y} \end{bmatrix}, \tag{17}$$

where $\lambda = \frac{\sigma^2_e}{\sigma^2_\beta}$. Subtracting the equations for $\hat{\boldsymbol{\alpha}}$ from the equations for $\vec{\boldsymbol{\beta}}_1$ gives $(\mathbf{I}\lambda)\vec{\boldsymbol{\beta}}_1 = \mathbf{0}$, implying that $\vec{\boldsymbol{\beta}}_1 = \mathbf{0}$**,** which verifies 16. Thus, solutions for $\hat{\boldsymbol{\alpha}}$ and $\vec{\boldsymbol{\beta}}_2$ from 17 are identical to those from the MME corresponding to a model where $\mathbf{X}_1$ is excluded from forming a similarity matrix. Such model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}. \tag{18}$$

The result $\vec{\boldsymbol{\beta}}_1 = \mathbf{0}$ is easy to verify empirically (a reviewer pointed out that it is probably well known by scientists working in genetic evaluation computations) but, to our knowledge, has not been reported in the literature. A "mechanistic" explanation for 16 is that BLUP of random effects with null means depends on $\mathbf{y}$ through error contrasts that have a null mean vector. So, if a locus is included in a model as a fixed effect, the error contrasts used for BLUP do not possess any information on the effects at such locus. Thus, if any factor (e.g., a marker locus) is included in the model both as fixed and random, the BLUP of the random effect will depend entirely on the "prior" (Bayesian view), and, as shown above, it will be identically equal to 0.

3.3 | Interdependent sets of marker effects

All elements of $\boldsymbol{\beta}$ were assumed independent and identically distributed, but the result holds for more complex dependency structures. Markers included in $\mathbf{X}_1$ may be in linkage disequilibrium with those in $\mathbf{X}_2$, and the phenomenon is encoded by correlations between columns of $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$. Further, models have been suggested that include dependencies among marker effects (e.g., Gianola, Pérez-Enciso, & Toro, 2003).

Suppose that $\boldsymbol{\beta} \sim (\mathbf{0,B})$ and $\mathbf{e} \sim (\mathbf{0,R})$ are independently distributed, that $\boldsymbol{\beta}_1 \sim (\mathbf{0,B}_{11})$, $\boldsymbol{\beta}_2 \sim (\mathbf{0,B}_{22})$ where $\mathbf{B}_{11}$ and $\mathbf{B}_{22}$ are non-singular and assume $Cov\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2'\right) = \mathbf{0}$. Here, the MME equations for the situation in which $\boldsymbol{\beta}_1$ is treated as both fixed and random take the form

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2 \\ \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1+\mathbf{B}^{11} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_2+\mathbf{B}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \vec{\boldsymbol{\beta}}_1 \\ \vec{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{19}$$

Subtracting the $\hat{\boldsymbol{\alpha}}$ from the $\vec{\boldsymbol{\beta}}_1$ equations gives $\mathbf{B}^{11}{\rightarrow}\boldsymbol{\beta}_1 = \mathbf{0}$, implying that the MME equations reduce to

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_2+\mathbf{B}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \vec{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \tag{20}$$

which provide $GLS(\boldsymbol{\alpha})$ and $BLUP(\boldsymbol{\beta}_2)$ under a model where $\boldsymbol{\beta}_1$ is fixed and $\boldsymbol{\beta}_2$ is random. Note that, within block, marker effects can be correlated or uncorrelated.

Allow now for a covariance structure between the two sets of random effects, and let $Cov\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2'\right) = \mathbf{B}_{12}$. The mixed model equations where $\boldsymbol{\beta}_1$ is treated as both fixed and random become

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2 \\ \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1+\mathbf{B}^{11} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2+\mathbf{B}^{12} \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1+\mathbf{B}^{21} & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_2+\mathbf{B}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \vec{\boldsymbol{\beta}}_1 \\ \vec{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{21}$$

Subtracting the $\boldsymbol{\alpha}$ from the $\boldsymbol{\beta}_1$ equations produces $\vec{\boldsymbol{\beta}}_1 = -\mathbf{B}^{12}\vec{\boldsymbol{\beta}}_2$, which is not null unless $\mathbf{B}^{12} = 0$, contradicting the model assumption. Hence, when $Cov\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2'\right) \neq 0$ use of $\mathbf{V}_2$ or $\mathbf{V}_{12}$ produces distinct sets of generalized least-squares solutions, so the result for the independence case does not hold here.

If two random vectors are not independent, fixing the value of one such vector (Listgarten et al., 2012, call this "conditioning") must alter the distribution of the other vector. Under multivariate normality, one can write $\boldsymbol{\beta}_2 = \mathbf{B}_{21}\boldsymbol{\beta}_1 + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \sim N\left(\mathbf{0,B}_{2.1} = \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\right)$. The model under fixed $\boldsymbol{\beta}_1$ becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\left(\mathbf{B}_{21}\boldsymbol{\beta}_1 + \boldsymbol{\delta}\right) + \mathbf{e} \\ &= \left(\mathbf{X}_1 + \mathbf{X}_2\mathbf{B}_{21}\right)\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\delta} + \mathbf{e} \\ &= \mathbf{X}_1^*\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\delta} + e, \end{aligned} \tag{22}$$

where $\mathbf{X}_1^* = \mathbf{X}_1 + \mathbf{X}_2\mathbf{B}_{21}$. Under this specification, the phenotypic covariance matrix is $\mathbf{V}_2^* = \mathbf{X}_2\mathbf{B}_{2.1}\mathbf{X}_2' + \mathbf{R}$ and $GLS(\boldsymbol{\beta}_1)$ should be computed as

$$\widetilde{\widetilde{\boldsymbol{\beta}}}_1 = \left(\mathbf{X}_1^{*\prime}\mathbf{V}_2^{*-1}\mathbf{X}_1^*\right)^{-1}\mathbf{X}_1^{*\prime}\mathbf{V}_2^{*-1}\mathbf{y}. \tag{23}$$

Likewise,

$$\begin{aligned} BLUP(\boldsymbol{\delta}) &= Cov\left(\boldsymbol{\delta},y'\right)\mathbf{V}_2^{*-1}\left(\mathbf{y} - \mathbf{X}_1^*\widetilde{\boldsymbol{\alpha}}\right) \\ &= \left(\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\right)\mathbf{X}_2'\mathbf{V}_2^{*-1}\left(\mathbf{y} - \mathbf{X}_1^*\widetilde{\boldsymbol{\alpha}}\right), \end{aligned} \tag{24}$$

will predict the effect of markers in $\mathbf{X}_2$ on phenotypes, conditionally on the effects of markers in $\mathbf{X}_1$, that is, in the absence of genetic variation at loci in marker set 1.

Note in 22 that $\mathbf{X}_1^*\boldsymbol{\beta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\mathbf{B}_{21}\boldsymbol{\beta}_1$ so that the "total signal" on the trait contributed by $\boldsymbol{\beta}_1$ is decomposed into a "direct" component $\mathbf{X}_1\boldsymbol{\beta}_1$ and an indirect contribution $\mathbf{X}_2\mathbf{B}_{21}\boldsymbol{\beta}_1$ mediated through the covariance between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ ($\mathbf{B}_{12}$). This sort of phenomenon is well known in structural equation modelling and path analysis (Wright, 1921).

4 | CONCLUSION

When conducting a GWAS with either a single marker or a set of markers treated as fixed, it is unnecessary to reconstruct the phenotypic variance–covariance matrix at each specific instance of testing, provided that a BLUP model is used and that marker effects in sets regarded as both fixed and random are independent across sets but not necessarily within sets. BLUE is invariant with respect to whether the genotypes of markers being tested in GWAS are employed in the construction of a genetic similarity matrix. Likewise, BLUP of the effects of the sets treated as random is invariant as well. However, the variance–covariance matrix of the GLS estimator and the prediction error-covariance matrix must be taken with respect to the assumptions made in the model employed for analysis. If marker effects in the two subsets of genotypes have a between-set nontrivial dependency structure, the GWAS model requires modification.

The results presented in this paper, shown first by Gianola et al. (2016) just for GLS (BLUE), are seemingly unrecognized in the GWAS literature (e.g., Chen, Steibel, & Tempelman, 2017). An additional and probably useful result reported here is that represented by Equation 13: the variance of the estimate of any of the marker effects tested in

GWAS can be obtained via a simple adjustment of the variance obtained with all markers entering into the similarity matrix.

ACKNOWLEDGEMENTS

Work was done while Daniel Gianola was Visiting Professor at Iowa State University (ISU) and was supported by the J. L. Lush Chair in Animal Breeding and Genetics, Department of Animal Science, ISU. We thank two anonymous reviewers for useful comments.

ORCID

Daniel Gianola https://orcid.org/0000-0001-8217-2348
Rohan L. Fernando https://orcid.org/0000-0001-5821-099X

REFERENCES

Aulchenko, Y. S., de Köning, D. J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. *Genetics*, *177*, 577–585. https://doi.org/10.1534/genetics.107.075614

Chen, C., Steibel, J. P., & Tempelman, R. J. (2017). Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. *Genetics*, *206*, 1791–1806. https://doi.org/10.1534/genetics.117.202259

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, *194*, 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., Fariello, M. I., Naya, H., & Schön, C. C. (2016). Genomewide association studies with a genomic relationship matrix: a case study with wheat and *Arabidopsis*. *G3: Genes, Genomes, Genetics*, *6*, 3241–3256. https://doi.org/10.1534/g3.116.034256

Gianola, D., Pérez-Enciso, M., & Toro, M. A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, *163*, 347–365.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One*, *8*(10), e75707. https://doi.org/10.371/journal.pone.0075707

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*, 1709–1723. https://doi.org/10.1534/genetics.107.080101

Legarra, A. (2016). Comparing estimates of genetic variance across different relationshipmodels. *Theoretical Population Biology*, *107*, 26–30. https://doi.org/10.1016/j.tpb.2015.08.005Epub 2015 Sep 2.

León-Novelo, L., & Casella, G. (2012). Prior influence in linear regression when the number of covariates increases to infinity. *Statistics and Probability Letters*, *82*, 438–445. https://doi.org/10.1016/j.spl.2011.10.018

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for genomewide association studies. *Nature Methods*, *9*, 525–526. https://doi.org/10.1038/nmeth.2037

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753. https://doi.org/10.1038/nature08494

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis*. New York, NY: Wiley Blackwell. https://doi.org/10.1002/SERIES1345

Van Raden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, *46*(2), 100–106. https://doi.org/10.1038/ng.2876

How to cite this article: Gianola D, Fernando RL, Garrick DJ. A certain invariance property of BLUE in a whole-genome regression context. *J Anim Breed Genet*. 2018;00:1–5. https://doi.org/10.1111/jbg.12378