Theoretical Population Biology 132 (2020) 47–59

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Inferring trait-specific similarity among individuals from molecular markers and phenotypes with Bayesian regression

Daniel Gianola [a,b,c,d,e,*], Rohan L. Fernando [c], Chris-Carolin Schön [d]

[a] Department of Animal Sciences, University of Wisconsin-Madison, USA
[b] Department of Dairy Science, University of Wisconsin-Madison, USA
[c] Department of Animal Science, Iowa State University, USA
[d] Department of Plant Sciences, Technical University of Munich, TUM School of Life Sciences, Germany
[e] Institut Pasteur de Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 18 April 2019
Available online 9 December 2019

Keywords:
Genomic relationship
Prediction
Complex traits
Genomic selection
Genomic prediction

ABSTRACT

Modeling covariance structure based on genetic similarity between pairs of relatives plays an important role in evolutionary, quantitative and statistical genetics. Historically, genetic similarity between individuals has been quantified from pedigrees via the probability that randomly chosen homologous alleles between individuals are identical by descent (IBD). At present, however, many genetic analyses rely on molecular markers, with realized measures of genomic similarity replacing IBD-based expected similarities. Animal and plant breeders, for example, now employ marker-based genomic relationship matrices between individuals in prediction models and in estimation of genome-based heritability coefficients. Phenotypes convey information about genetic similarity as well. For instance, if phenotypic values are at least partially the result of the action of quantitative trait loci, one would expect the former to inform about the latter, as in genome-wide association studies. Statistically, a non-trivial conditional distribution of unknown genetic similarities, given phenotypes, is to be expected. A Bayesian formalism is presented here that applies to whole-genome regression methods where some genetic similarity matrix, e.g., a genomic relationship matrix, can be defined. Our Bayesian approach, based on phenotypes and markers, converts prior (markers only) expected similarity into trait-specific posterior similarity. A simulation illustrates situations under which effective Bayesian learning from phenotypes occurs. *Pinus* and wheat data sets were used to demonstrate applicability of the concept in practice. The methodology applies to a wide class of Bayesian linear regression models, it extends to the multiple-trait domain, and can also be used to develop phenotype-guided similarity kernels in prediction problems.

1. Introduction

Assessing genetic or genomic similarity among species or individuals is a central topic in evolutionary and quantitative genetics (Lynch and Walsh, 1998; Walsh and Lynch, 2018). Sethuraman (2018) reviewed areas where estimation of genetic relatedness is important, including paternity and maternity assignments, forensic, association and linkage studies, and inference and prediction in quantitative genetics.

Until recently, many quantitative-trait analyses such as estimation of genetic variances and covariances and prediction of unobservable genotypic values (e.g., breeding values in animal and plant breeding), have relied on modeling covariances based on pedigree-based genetic relatedness between relatives; such covariances enter into the dispersion structure of mixed effects and Bayesian models. Historically, agricultural breeders have quantified genetic similarity by the probability that randomly chosen homologous alleles are identical by descent (IBD) in a pair of individuals. These probabilities are used to form Sewall Wright's numerator relationship matrix (**A**), which is proportional to the covariance matrix between additive genetic values of individuals. Eventually, elements of **A** enter into quantitative genetic models via the notion of covariances between relatives (Kempthorne, 1954). The calculations of IBD probabilities are deterministic and rely on the notion of a conceptual ancestral population from which descendants evolve in the absence of selection and mutation, following some equilibrium laws.

The advent of massive genomic data, e.g., DNA sequences, has changed the classical paradigm. Realized measures of genomic similarity have been developed and applied to estimation of genomic variances (genomic heritability) and covariances (genomic correlations), and for predicting genome-derived breeding values

* Corresponding author at: Department of Animal Sciences, University of Wisconsin-Madison, USA.
E-mail address: gianola@ansci.wisc.edu (D. Gianola).

https://doi.org/10.1016/j.tpb.2019.11.008

(Bernardo, 1994; Nejati-Javaremi et al., 1997; Visscher et al., 2006; Van Raden, 2008; Hayes et al., 2009; de los Campos et al., 2011, 2013, 2015; Gianola et al., 2015; Lehermeier et al., 2017). For instance, a method known as "genomic best linear unbiased prediction" (GBLUP) uses a genome-based similarity matrix (the "genomic relationship matrix", **G**) among individuals to predict genomic breeding values (e.g., Van Raden, 2008. In GBLUP, **G** replaces the **A** used in pedigree-based prediction models, and the method has become a gold standard in dairy cattle breeding (García-Ruiz et al., 2016; Wiggans et al., 2017). Fernando et al. (2017) present a discussion of important differences between **A** and genome-based similarity matrices.

The most widely used **G** (Van Raden, 2008; $\mathbf{G}_{VR}$ hereinafter) is built from genotype codes (e.g., 0, 1, 2 for *bb*, *Bb* and *BB* individuals, respectively, with *B* being the reference allele) deviated from their means and observed at p single nucleotide polymorphism (SNP) loci in each of n individuals. Typically, the genotype scoring encodes additive genetic effects. With **X** denoting an observed matrix of mean-deviated marker scores, then the $n \times n$ matrix $\mathbf{G}_{VR}$ is proportional to $\mathbf{XX}'$. Under Hardy–Weinberg equilibrium and in the absence of selection or mutation, the expected value of $\mathbf{G}_{VR}$ (after some normalization) is **A**, the pedigree-based similarity matrix (Habier et al., 2007; Gianola et al., 2009). $\mathbf{G}_{VR}$ weights all SNPs alike and does not exploit linkage disequilibrium (LD) information beyond what is conveyed by colinearity among columns of **X**. To see this point, note that the i, jth element of $\mathbf{G}_{VR}$ is proportional to $\sum_{k=1}^{p} x_{ik}x_{jk}$, where x_{ik} and x_{jk} are the scores at locus k for individuals i and j. An alternative similarity matrix could take the more general form $\mathbf{XWX}'$ for some $p \times p$ weight matrix **W** constructed such that LD or effect-size based differential weights enter into **G**; in $\mathbf{G}_{VR}$, **W** is actually an $n \times n$ identity matrix. Speed et al. (2012) proposed a genomic relationship matrix that exploits LD information .

Importantly, $\mathbf{G}_{VR}$ also ignores information that phenotypes may convey about genetic similarities. A central dogma of quantitative genetics is that genetic similarity generates covariance among relatives, so phenotypic similarities (after accounting for environmental sources of variation) are expected to reflect similarity due to allele sharing at quantitative trait loci (QTL) affecting a trait in question. Conversely, phenotypic similarity should be expected to inform about genetic (genomic) similarity. Further, if distinct QTL affect different traits, similarities on genotypic values for feed or nutrient consumption, say, should differ from genotypic similarities due to resistance to disease. Use of trait-specific similarity may be useful for enhancing prediction of complex traits in animal and plant breeding and in personalized medicine (de los Campos et al., 2011). Hence, $\mathbf{G}_{VR}$ is not necessarily the best prescription for all traits.

Thompson (1975) is representative of approaches that employ likelihood-based inference of kinship relationships from genetic data. Bravington et al. (2016) and Wang et al. (2017) pointed out that existing measures of kinship did not exploit information such as order along a chromosome (linkage or LD), and addressed estimation of genetic similarities using formal statistical procedures such as the Day-Williams method (Day-Williams et al., 2011). Making use of estimation theory is clearly a step forward towards ascertaining molecular kinship, as properties of estimators can be examined in a well-defined theoretical framework. However, neither Bravington et al. (2016) nor Wang et al. (2017) incorporated phenotypic information into their models, perhaps because in many areas (e.g., evolutionary biology) the relevant phenotypes are not easily available. On the other hand, animal breeders have attempted to incorporate phenotypic values into calculations of trait-specific genomic similarity. Zhang et al. (2010) used a two-stage approach to obtain a trait-specific genomic relationship matrix. In the first stage, Bayesian multiple-regression methods produce estimates of substitution effects at the SNP loci. In the second stage, weights computed from the effect size estimates are used to form a genomic relationship matrix that is a weighted average of identity-by-state matrices computed for each SNP locus. The actual weight assigned to a given marker was the "estimated genetic variance at that locus", obtained using estimates of allelic frequency and of substitution effects. The expressions used to represent "estimated genetic variance" cannot be always justified as formal metrics for such parameter. In particular, one representation assumed Hardy–Weinberg and linkage equilibrium among markers; the latter assumption is manifestly violated in plant and animal breeding data.

In another approach (Wang et al., 2012), the weights of Zhang et al. (2010) were applied iteratively. In each iteration, GBLUP was used to estimate genomic breeding values, with estimates of SNP effects obtained indirectly as in Strandén and Garrick (2009). In a subsequent iteration, an idea suggested by Van Raden et al. (2009) was used to form a weighted **G** matrix where the contribution from each locus was weighted by the "estimated variance at that locus". Sun et al. (2012) described a method that is identical to that of Wang et al. (2012) except in the weights used for computing **G**. Sun et al. (2012) derived weights from an EM algorithm (Dempster et al., 1977) that presumably converged to produce the joint posterior mode of SNP effects under the Bayes A model of Meuwissen et al. (2001). In the context of genome-wide association studies, Liu et al. (2016) described an iteration between fixed and random effects models which would lead to some phenotype-informed similarity matrix. Karaman et al. (2018) proposed an ad-hoc multiple-trait method that used a Bayes A-type (Bayes AS) procedure hybridized with a GBLUP approach. All these approaches, although intuitively appealing, are heuristic so it is difficult to characterize their properties from a formal statistical perspective.

We present a Bayesian single-stage method for inferring genomic similarities among individuals that uses both marker and phenotypic information. Since the Bayesian model can be solved by sampling from the joint posterior distribution of a similarity matrix, uncertainty can be characterized fully and precisely in the framework of a well-established theory. Our method is adapted to several different Bayesian regression models and is illustrated employing simulation and with analyses of *Pinus* and wheat data sets. It is also shown how the concept can be adapted to prediction in a training–testing setting, and multiple-trait extensions are suggested.

2. Genomic similarity (relationship) matrices

2.1. Linear regression connecting markers to phenotypes

Let $\mathbf{X} = \{x_{ij}\}$ be an $n \times p$ (n = number of individuals; p = number of markers) observed matrix of centered molecular scores used as covariates in the linear regression model

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

where $\mathbf{y}$ (phenotypes) and $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$ (model residuals, where σ_e^2 is a variance parameter) are n-dimensional vectors, and β is a $p \times 1$ vector of regression coefficients; typically $n < p$. Assume that nuisance location effects have been eliminated. One way of dealing with the lack of likelihood identification caused by $n < p$ is to adopt a Bayesian position and assign some prior distribution to the vector of regression coefficients (Meuwissen et al., 2001). For instance, Bayes A uses a $t-$distribution as prior, the Bayesian Lasso (BL) adopts a conditional double exponential prior, and Bayes R assigns a mixture of four normal distributions with known variance. These methods were reviewed and discussed by, e.g., de los Campos et al. (2013) and Gianola (2009, 2013).

2.2. Definition of genomic similarity matrices

2.2.1. General considerations

The notion of a genetic similarity matrix is motivated here. In a pedigree-based model, if $\mathbf{g}$ is a vector of additive effects of n individuals, their covariance matrix (Henderson, 1976) is

$$Var\,(\mathbf{g}|\mathbf{A}) = \mathbf{A}\sigma_a^2, \tag{2}$$

where $\mathbf{A}$, defined earlier, is a similarity matrix satisfying

$$\mathbf{A} = \frac{Var\,(\mathbf{g}|\mathbf{A})}{\sigma_a^2}, \tag{3}$$

and σ_a^2 is the additive genetic variance.

The most widely used form of genome-based similarity matrix $(n \times n)$ is denoted generically as $\mathbf{G} \propto \mathbf{XX}'$. Such $\mathbf{G}$ appears naturally if regression coefficients in (1) are treated as random variables. Assume $\boldsymbol{\beta} \sim F(\mathbf{0}, \mathbf{B}\sigma_\beta^2)$, where F is some prior distribution with covariance matrix $\mathbf{B}\sigma_\beta^2$ and σ_β^2 is a variance parameter. From (1), with $\mathbf{g} = \mathbf{X}\beta$ and $\mathbf{B} = \mathbf{I}$ it follows that

$$Var\,(\mathbf{g}|\mathbf{X}) = \mathbf{XX}'\sigma_\beta^2, \tag{4}$$

so consistently with (3) an *a priori* genomic similarity matrix can be defined as

$$\mathbf{G} = \frac{\mathbf{XX}'\sigma_\beta^2}{\sigma_g^2}, \tag{5}$$

where σ_g^2 is the genomic variance encoded by the model. The distribution of $\mathbf{g}$ depends on F.

As an example, suppose the method of inference is best linear unbiased prediction (BLUP). Here, σ_β^2 is a known variance parameter representing the variability of marker effects over conceptual repeated sampling. The estimates of β obtained with BLUP are numerically identical (at the same level of regularization) to those from ridge-regression so the procedure is often called "ridge-regression BLUP" (RRBLUP). The two needed variance components in BLUP, σ_β^2 and σ_e^2, can be inferred using Bayesian or likelihood-based methods, and then kept fixed in the BLUP computations as if they were true values. Van Raden (2008) defined as "genomic relationship" matrix

$$\mathbf{G}_{VR} = \frac{\mathbf{XX}'}{\sum_{j=1}^{p} 2q_j(1-q_j)} = \frac{\mathbf{XX}'}{pH}, \tag{6}$$

where q_j is the frequency of a reference allele at marker locus j and H is average heterozygosity, taken over markers. Observe that only $\mathbf{X}$ informs about similarity in $\mathbf{G}_{VR}$. The connection between σ_β^2 and σ_g^2 in BLUP can be deduced readily. The regression model in scalar form is $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$. If it is assumed that all x_{ij} (marker genotypes) and β_j (regression coefficients) are independently distributed random variables with null means, unconditionally with respect to the $x_{ij}'s$ and $\beta_j's$, the genomic variance in GBLUP can be defined as

$$\begin{aligned}\sigma_g^2 &= Var\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) = \sum_{j=1}^{p} E\left(x_{ij}^2\beta_j^2\right) \\ &= \sum_{j=1}^{p} E\left(x_{ij}^2\right) E\left(\beta_j^2\right) = \sum_{j=1}^{p} 2q_j(1-q_j)\sigma_\beta^2 = pH\sigma_\beta^2.\end{aligned} \tag{7}$$

assuming Hardy–Weinberg equilibrium holds at each locus; independence of x_{ij} from any $x_{ij'}$ implies linkage equilibrium. This representation of genomic variance holds for any regression model, but the specific form of σ_β^2 depends on the prior adopted for marker effects. Here, $Var(\mathbf{g}|\mathbf{X}) = \mathbf{XX}'\sigma_\beta^2 = \mathbf{G}_{VR}\sigma_g^2$.

It is instructive to contrast (7) with the classical formula for additive genetic variance (V_A) in a model with a finite $(K$, say) number of QTL under linkage equilibrium. Here

$$V_A = \sum_{k=1}^{K} 2q_k(1-q_k)\alpha_k^2, \tag{8}$$

where q_k is allelic frequency at QTL k and α_k^2 is a fixed substitution effect, for $k = 1, 2, \ldots, K$ (Falconer and Mackay, 1996; Lynch and Walsh, 1998). There are important conceptual differences between σ_g^2 and V_A (e.g., de los Campos et al., 2015). In particular, additive genetic variance stems from randomness of genotypes and fixed effects at QTL, whereas σ_β^2 (variance of marker effects distribution, with perhaps even none of the markers being a QTL) and σ_g^2 (genomic variance) stem from a model where marker substitution effects are random in frequentist or Bayesian senses. In the latter case, σ_β^2 is a metric for prior Bayesian uncertainty (Gianola et al., 2009; Gianola, 2013) that appears as a hyperparameter in a GBLUP model. Note that $\mathbf{G}_{VR}$ derives from a single realization of $\mathbf{X}$, an observable matrix. On the other hand, in classical quantitative genetics $\mathbf{X}$ is viewed as varying at random according to a distribution reflecting equilibrium or disequilibrium laws, typically the latter in artificial breeding. Hence, $\mathbf{G}_{VR}$ would be a random matrix as well; in the absence of selection and mutation and under Hardy–Weinberg equilibrium, it can be shown (e.g., Habier et al., 2007) that $E\,(\mathbf{G}_{VR}) = \mathbf{A}$.

2.2.2. Trait-specific (effects based) similarity matrices

Write model (1) as $\mathbf{g} = \sum_{j=1}^{p} \mathbf{X}_j\beta_j$ where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. The notion of a trait-specific similarity matrix is introduced by defining

$$\mathbf{G}\,(\beta) = \sum_{j=1}^{p} \mathbf{X}_j \left(\frac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}\right) \mathbf{X}_j' = \mathbf{XD}\,(\beta)\,\mathbf{X}', \tag{9}$$

where $\mathbf{D}\,(\beta)$ is an unknown $p \times p$ diagonal matrix with typical element $\dfrac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}$; $j = 1, 2, \ldots, p$, whose values range between 0 and 1. If $\beta_j = 0$, locus j does not contribute to trait-specific similarity. Further, loci with stronger (positive or negative) effects contribute more towards similarity than loci with weak effects. Note that, if marked effects are independent and identically distributed, for any prior distribution

$$E\left[\frac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}|\mathbf{X}\right] \approx \frac{E\left(\beta_j^2\right)}{E\left(\sum_{j=1}^{p}\beta_j^2\right)} = \frac{1}{p}. \tag{10}$$

As it will be seen throughout the paper, this definition produces an expected prior similarity that is proportional to common, marker derived, measures of similarity. Subsequently, Bayesian learning incorporates phenotypic information, producing measures of posterior similarity.

Once β is assigned a prior distribution F, a prior distribution is automatically induced for $\mathbf{G}\,(\beta)$. If F is such that marker effects are independent and identically distributed with variance σ_β^2, the prior expectation of $\mathbf{G}\,(\beta)$ is, approximately

$$E\left[\mathbf{G}\,(\beta)\,|\mathbf{X}\right] \approx \mathbf{X}diag\left\{\frac{\sigma_\beta^2}{p\sigma_\beta^2}\right\}\mathbf{X}' = \frac{\mathbf{XX}'}{p} = H\mathbf{G}_{VR}, \tag{11}$$

where H is average heterozygosity, as defined earlier. Hence, prior similarity is proportional to the similarity conveyed by Van Raden's genomic relationship matrix, i.e., *a priori* all markers contribute to expected similarity.

If Bayesian learning for β takes place, it must also occur for $\mathbf{G}\,(\beta)$, producing a posterior distribution of similarities

$[\mathbf{G}(\beta)|\mathbf{X},\mathbf{y}]$. Using definition (9), the posterior expectation of the similarity matrix is

$$E[\mathbf{G}(\beta)|\mathbf{X},\mathbf{y}] = \sum_{j=1}^{p}\mathbf{X}_j E\left(\frac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}|\mathbf{y}\right)\mathbf{X}_j' = \mathbf{X}E[\mathbf{D}(\beta)|\mathbf{y}]\mathbf{X}'. \quad (12)$$

Approximately,

$$E\left(\frac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}|\mathbf{y}\right) \approx \frac{E\left(\beta_j^2|\mathbf{y}\right)}{\sum_{j=1}^{p}E\left(\beta_j^2|\mathbf{y}\right)} = \frac{E^2\left(\beta_j|\mathbf{y}\right)+Var\left(\beta_j|\mathbf{y}\right)}{\sum_{j=1}^{p}\left[E^2\left(\beta_j|\mathbf{y}\right)+Var\left(\beta_j|\mathbf{y}\right)\right]};\ j=1,2,\ldots,p. \quad (13)$$

Provided the model holds and assuming likelihood-identified parameters, standard asymptotic Bayesian theory (Sorensen and Gianola, 2002) indicates that as $n\to\infty$, then $E^2\left(\beta_j|\mathbf{y}\right)\to\beta_j^2$ and $Var\left(\beta_j|\mathbf{y}\right)\to 0$; in the limit the jth diagonal element of $\mathbf{D}(\beta)$ goes to $d_j = \frac{\beta_j^2}{\sum_{j=1}^{p}\beta_j^2}$, which is the "true" contribution of locus j towards similarity. In finite samples, if two markers are inferred at the same level of precision, the one with stronger effects will be assigned more weight.

The extent of Bayesian learning stemming from markers and phenotypes, relative to the information conveyed from markers only, can be evaluated by computing Frobenius distances (*dist*) between random draws from the prior and posterior distributions of $\mathbf{G}(\beta)$ or $\mathbf{D}(\beta)$. For instance, the Frobenius distance between pairs of samples of $\mathbf{G}(\beta)$ is

$$dist^{(s)} = \sqrt{tr\left[\mathbf{X}\mathbf{D}\left(\beta^{(s,prior)}\right)\mathbf{X}' - \mathbf{X}\mathbf{D}\left(\beta^{(s,post)}\right)\mathbf{X}'\right]^2};\quad s=1,2,\ldots,S. \quad (14)$$

The S samples can be used to estimate the distribution of *dist*.

The effects-based notion of similarity proposed holds for any member of the Bayesian alphabet (Gianola et al., 2009; Gianola, 2013), i.e., irrespective of the prior adopted. Further, the use of MCMC samples permits a full characterization of the posterior distribution of *dist* since any of its features can be estimated directly from the draws.

2.3. Implementation-specific similarity matrices

In contrast to the generic effect-size matrix defined in (9), measures of similarity can be constructed based on specific features of members of the Bayesian alphabet referred to previously. Here, we present the main ideas focusing on Bayes A (Meuwissen et al., 2001); specific details pertaining to some other members of the Bayesian alphabet (Gianola et al., 2009) are given in the Appendix

2.3.1. Bayes A

Bayes A is a linear regression model with a two-stage hierarchical prior assigned to marker effects. The model assumes: (1) $\beta_j|\sigma^2_{\beta_j}\sim N\left(0,\sigma^2_{\beta j}\right)$ and (2) $\sigma^2_{\beta j}|\nu_\beta,S^2_\beta\sim IID(\nu_\beta S^2_\beta\chi_\beta^{-2})$ for each $j=1,2,\ldots,p$; *IID* means "independent and identically distributed". The first stage poses a marker-specific variance $\sigma^2_{\beta_j}$ and the second stage assumes that all such variances follow the same scale inverted chi-square distribution on ν_β degrees of freedom and with scale parameter S^2_β; both ν_β and S^2_β are hyper-parameters. Gianola et al. (2009) pointed out that the unconditional prior distribution assigned to each of the marker effects is actually $\beta_j|\nu_\beta,S^2_\beta\sim t_{\nu_\beta}(0,S^2_\beta)$, $j=1,2,\ldots,p$; that is, all markers are assumed, *a priori*, to follow the same $t-$distribution with variance $\sigma^2_\beta = S^2_\beta\frac{\nu_\beta}{\nu_\beta-2}$. Meuwissen et al. (2001) interpreted the meaning of $\sigma^2_{\beta_j}$ incorrectly and attempted to connect such parameter to some region-specific genetic variance. In fact, markers are homoscedastic, as the same t-distribution is assigned throughout; the incorrect interpretation reappears in the literature (e.g., Zhang et al., 2010).

In Bayes A the covariance matrix of marker effects is $Var\left(\beta|\nu_\beta,S^2_\beta\right) = \mathbf{I}S^2_\beta\frac{\nu_\beta}{\nu_\beta-2}$, where $\mathbf{I}$ is a $p\times p$ identity matrix. Hence, the vector of genomic breeding values $\mathbf{g}=\mathbf{X}\beta$ has as mean vector and covariance matrix of the prior distribution

$$\mathbf{g}|\nu_\beta,S^2_\beta\sim\left(\mathbf{0},\mathbf{X}\mathbf{X}'S^2_\beta\frac{\nu_\beta}{\nu_\beta-2}\right). \quad (15)$$

The unknown prior distribution of each element of $\mathbf{g}$ is that of a linear combination of independent t-random variables (Fisher, 1935; Sukhatme, 1938; Walker and Saw, 1978). It follows that the "prior genomic relationship matrix" in Bayes A has the same form as (6), but the genomic variance here is

$$\sigma^2_g = \sum_{j=1}^{p}2q_j(1-q_j)S^2_\beta\frac{\nu_\beta}{\nu_\beta-2}. \quad (16)$$

Interpretation of σ^2_g in Bayes A cannot be disassociated from the hyper-parameters intervening in the distribution of marker effects; given σ^2_g, an increase in ν_β must be compensated by a decrease in S^2_β.

Since $\beta_j|\sigma^2_{\beta_j}\sim N\left(0,\sigma^2_{\beta j}\right)$, then

$$\beta|\mathbf{V}_{BA}\sim N(\mathbf{0},\mathbf{V}_{BA}), \quad (17)$$

where $\mathbf{V}_{BA} = diag\left\{\sigma^2_{\beta_j}\right\}$ is an unknown matrix. Hence, $\mathbf{g}=\mathbf{X}\beta|X,V_{BA}\sim N\left(\mathbf{0},\mathbf{X}\mathbf{V}_{BA}\mathbf{X}'\right)$. An unobserved similarity matrix in Bayes A, can be defined following (4) and (6) as

$$\mathbf{G}_{BA} = \mathbf{X}\mathbf{D}_{BA}\mathbf{X}', \quad (18)$$

where

$$\mathbf{D}_{BA} = diag\left\{\frac{\sigma^2_{\beta_j}}{\sum_{j=1}^{p}\sigma^2_{\beta_j}}\right\}. \quad (19)$$

A priori, $E\left(\sigma^2_{\beta_j}|\nu_\beta,S^2_\beta\right) = \frac{\nu_\beta S^2_\beta}{\nu_\beta-2}$ for any j, so

$$E\left(\mathbf{G}_{BA}|\nu_\beta,S^2_\beta\right)\approx\frac{H\mathbf{X}\mathbf{X}'}{pH} = H\mathbf{G}_{VR}, \quad (20)$$

i.e., the prior expectation is the same as that of the effect-size based similarity given in (11).

A posterior distribution can be defined for $\mathbf{G}_{BA}$ as well. The Bayes A method allows for some (but limited) Bayesian learning about the $\sigma^2_{\beta_j}$ parameters and, therefore, about $\mathbf{D}_{BA}$ and $\mathbf{G}_{BA}$. Bayes A is implemented with MCMC using a Gibbs sampler (e.g., Meuwissen et al., 2001; Pérez and de los Campos, 2014). In the sampler, draws are made from the conditional posterior distributions $\left[\sigma^2_{\beta_j}|ELSE\right]$ where *ELSE* denotes all other parameters, hyper-parameters and data. For Bayes A, $\sigma^2_{\beta_j}|ELSE\sim\left(\beta_j^2+\nu_\beta S^2_\beta\right)\chi^{-2}_{1+\nu_\beta}$, $j=1,2,\ldots,p$. Note that only 1 degree of freedom is gained over the prior, thus posing an upper limit to the learning that can be attained for each $\sigma^2_{\beta_j}$, with the same holding for $\mathbf{D}_{BA}$ (Gianola et al., 2009). If S draws are available from the

posterior distributions of each of the $\sigma^2_{\beta_j}$ parameters, the posterior distribution of $\mathbf{G}_{BA}$ can be estimated from the MCMC samples

$$\mathbf{G}^{(s)}_{BA} = \mathbf{X}\mathbf{D}^{(s)}_{BA}\mathbf{X}'; s = 1, 2, \ldots, S. \tag{21}$$

The posterior expectation of the distribution is estimated as $\overline{\mathbf{G}}_{BA} = \sum_{s=1}^{S}\mathbf{G}^{(s)}_{BA}/S$, and the posterior variance of the similarity between individuals i and j can be assessed as $\sum_{s=1}^{S}\left(G^{(s)}_{BA,ij} - \overline{G}_{BA,ij}\right)^2/S$. Frobenius distances between draws from the prior and posterior distributions of $\mathbf{D}_{BA}$ can be used to quantify the extent of Bayesian learning about similarity stemming from use of phenotypes over and above what is learned from markers only.

3. Learning similarity from multiple-trait models

Similarity matrices can also be learned from multiple-trait analyses. For example, consider a two-trait Bayes A regression, e.g., a joint analysis of stature (trait 1) and body weight (trait 2) in a group of subjects, and assume that both traits are measured in each of n individuals. A bivariate Bayes A model poses the hierarchy

$$(1)\ \begin{bmatrix}\beta_{j1}\\ \beta_{j2}\end{bmatrix}|\mathbf{B}_j \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \mathbf{B}_j = \begin{bmatrix}B_{j11} & B_{j12}\\ B_{j21} & B_{j22}\end{bmatrix}\right); j = 1, 2, \ldots, p, \tag{22}$$

and

$$(2)\ \mathbf{B}_j|\Omega, \nu \sim IW(\Omega, \nu); j = 1, 2, \ldots, p, \tag{23}$$

Above, $\mathbf{B}_j$ is a 2×2 marker-specific covariance matrix and IW denotes an inverted Wishart distribution with known $\Omega = \{\Omega_{ij}\}$ (scale matrix) and ν (degrees of freedom) parameters. All $\mathbf{B}_j$ matrices are independent and identically distributed, *a priori,* so the bivariate Bayes A model assigns the same IW distribution to each of the p matrices $\mathbf{B}_j$. The marginal prior distribution of the vector of marker effects is bivariate t, with null mean vector and covariance matrix $\Omega\frac{\nu}{\nu-2}$. Sorting individuals within traits, let $\mathbf{g}_1 = \mathbf{X}\beta_1$ and $\mathbf{g}_2 = \mathbf{X}\beta_2$ where β_1 is the $p \times 1$ vector of marker effects on trait 1, and so on. The model is

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix} = (\mathbf{I}\otimes\mathbf{X})\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}, \tag{24}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix, $\otimes$ is the Kronecker product and the $\mathbf{e}'s$ are model residuals. The genetic variance–covariance matrix of genomic values is therefore

$$\begin{aligned} Var\left(\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix}|\Omega, \nu\right) &= (\mathbf{I}\otimes\mathbf{X})\, Var\left(\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}|\Omega, \nu\right)(\mathbf{I}\otimes\mathbf{X}')\\ &= (\mathbf{I}\otimes\mathbf{X})\left(\Omega\frac{\nu}{\nu-2}\otimes\mathbf{I}\right)(\mathbf{I}\otimes\mathbf{X}')\\ &= \frac{\nu}{\nu-2}\begin{bmatrix}\Omega_{11}\mathbf{X}\mathbf{X}' & \Omega_{12}\mathbf{X}\mathbf{X}'\\ \Omega_{21}\mathbf{X}\mathbf{X}' & \Omega_{22}\mathbf{X}\mathbf{X}'\end{bmatrix}. \end{aligned} \tag{25}$$

Hence, trait-specific genomic variances in the bivariate Bayes A model are

$$\sigma^2_{g_t} = \Omega_{tt}\frac{\nu}{\nu-2}pH; t = 1, 2, \tag{26}$$

and the genomic covariance is

$$\sigma_{g_{tt'}} = \Omega_{tt'}\frac{\nu}{\nu-2}pH;\ t \neq t'. \tag{27}$$

As in (18), an unknown (normalized) similarity matrix can be defined for trait t as

$$\mathbf{G}_{BA_t} = \mathbf{X}\mathbf{D}_{BA_t}\mathbf{X}';\ t = 1, 2. \tag{28}$$

where $\mathbf{D}_{BA_t} = diag\left\{B_{jtt}/\sum_{j=1}^{p}B_{jtt}\right\}$, $t = 1, 2$, is a diagonal matrix of order p. Hence, *a priori*

$$E\left(\mathbf{G}_{BA_i}|Hyp\right) \approx \mathbf{X}\begin{bmatrix}\Omega_{11}\frac{\nu}{\nu-2} & \\ & p\Omega_{11}\frac{\nu}{\nu-2}\end{bmatrix}\mathbf{X}' = H\mathbf{G}_{VR} \tag{29}$$

In a Gibbs sampling context, it can be shown that $\mathbf{B}_j|ELSE$ is an inverse Wishart distribution. The sampler provides draws from the posterior distribution of $\mathbf{B}_j$ for each marker, thus leading to draws from the posterior distributions of $\mathbf{D}_{BA_t}$ and of $\mathbf{G}_{BAt}$. The two $n \times n$ blocks, $\mathbf{G}_{BA_1}$ and $\mathbf{G}_{BA_2}$ represent the similarity matrices for each of the two traits in the bivariate analysis. Given S samples from the posterior distributions of $\mathbf{G}_{BA_1}$ and $\mathbf{G}_{BA_2}$ the posterior distribution of the Frobenius distance between the two similarity matrices can be estimated as

$$dist^{(s)}_{1-2} = \sqrt{tr\left(\mathbf{G}^{[s]}_{BA_1} - \mathbf{G}^{[s]}_{BA_2}\right)^2}; s = 1, 2, \ldots, S. \tag{30}$$

If the distribution of the Frobenius distance between $\mathbf{G}_{BA_1}$ and $\mathbf{G}_{BA_2}$ is away from 0, such finding would support the view that trait-specific similarity matrices are required.

Bayes B and Bayes $C\pi$ (see Appendix for a description of the basics of these methods) can be adapted to the multiple trait case following Cheng et al. (2018). Likewise, a multiple-trait Bayesian Lasso (MBL) that assumes a conditional multivariate Laplace distribution for marker effects can be considered as well. As in Gianola and Fernando (2018), ordering the $p \times 1$ vectors of marker effects (β_i) within trait, the conditional prior for bivariate marker effects is

$$\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}|\Sigma, \mathbf{v}^2 \sim N(\mathbf{0}, \Sigma\otimes\mathbf{D}_v), \tag{31}$$

where $\mathbf{D}_v = \{v_j^2\}$ is a $p \times p$ diagonal matrix of unknown weights, each distributed as $Gamma\left(1, \frac{1}{8}\right)$, *a priori,* and Σ is the scale matrix of a bivariate Laplace distribution. Hence,

$$\mathbf{g} = \begin{bmatrix}\mathbf{X}\beta_1\\ \mathbf{X}\beta_2\end{bmatrix}|\Sigma, \mathbf{v}^2 \sim N\left(\mathbf{0}, \Sigma\otimes\mathbf{X}\mathbf{D}_v\mathbf{X}'\right). \tag{32}$$

As shown in Gianola and Fernando (2018) the diagonal elements of $\mathbf{D}_v$ have unrecognizable conditional posterior distributions and a Metropolis–Hastings algorithm was suggested by these authors for drawing samples. Gianola and Fernando (2018) give $Gamma(v_j^2|\frac{T+1}{2}, \frac{1}{8})$ as prior distribution for the weights; $T =$ number of traits. The similarity matrix in MBL can be defined as

$$\mathbf{G}_{BL} = \mathbf{X}\mathbf{D}_{MBL}\mathbf{X}', \tag{33}$$

where $\mathbf{D}_{MBL} = diag\left(v_j^2/\sum_{j=1}^{p}v_j^2\right)$. Note that there is a single similarity matrix for the MBL model, as v_j^2 takes the same value across traits. *A priori,* for $T = 2$

$$E(\mathbf{G}_{BL}|Hyp) \approx \frac{\mathbf{X}\mathbf{X}'}{p} = H\mathbf{G}_{VR}. \tag{34}$$

As before, the extent of Bayesian learning can be evaluated by comparing the prior and posterior distribution of $\mathbf{G}_{BL}$ in (33).

4. Simulation

4.1. Setting

A genome consisting of 10 chromosomes, each of length one Morgan and containing 2000 SNP loci per chromosome, was simulated using the XSim package developed in the Julia programming language environment (Cheng et al., 2015). Random

mating was practiced in a population of size $n = 100$ for one-hundred generations to generate linkage disequilibrium between loci. Thereafter, the population was expanded to $n = 500$, 2000 or 4000 to produce increasingly large training samples and to evaluate the extent of Bayesian learning as a function of size of the sample. One hundred loci were randomly drawn from across the genome and designated as QTL; substitution effects for the QTL were simulated from a standard normal distribution. Phenotypic values were generated as

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{e}, \tag{35}$$

where $\mathbf{Q}$ is the matrix of the QTL genotypes and $\boldsymbol{\alpha}$ is a fixed vector of "true" substitution effects. The residuals in vector $\mathbf{e}$ were sampled independently from a normal distribution with null mean and variance σ_e^2 chosen to generate a trait with a heritability of 0.5.

The 20,000 simulated SNP genotypes (including the 100 designated as QTL) and the phenotypes from the training samples were used in Bayes $C\pi$ analyses (Habier et al., 2011), with the proportion of loci with null substitution effects π treated as unknown and assigned a uniform prior. For the Bayes $C\pi$ model we constructed an effect-size based similarity matrix as defined in (9). The Julia package JWAS (Cheng et al., 2016) was used to draw 60,000 Gibbs samples of model unknowns for each Bayes$C\pi$ analysis, and every 20th posterior sample of the vector of effects, β, was saved for inference; hence, 3000 samples were employed for inference on genomic similarity.

Since genotypes and their effects are known in the simulation, true genomic similarity at the QTL level could be computed, and this was done for illustrative purposes for the first 100 individuals from each of the three population sizes. Using (9) the "true" genomic similarity matrix employed for these individuals was

$$\mathbf{G}_Q = \mathbf{Q}_c\mathbf{D}(\alpha)\mathbf{Q}_c', \tag{36}$$

where $\mathbf{Q}_c$ is the matrix of centered QTL genotype scores for the 100 individuals at the 100 QTL and α were the true (simulated) QTL effect sizes. For each of the 3000 MCMC posterior samples, a similarity matrix was computed from marker effects and marker genotypes of the 100 individuals as

$$\mathbf{G}\left[\beta^{(s)}\right] = \mathbf{X}\mathbf{D}_\beta^{(s)}\mathbf{X}';\ s = 1, 2, \ldots, 3000. \tag{37}$$

Similarly, 3000 samples from the prior distribution of β were used to produce draws from the prior distribution of $\mathbf{G}(\beta)$

The Frobenius distance between $\mathbf{G}_Q$ and each of the $\mathbf{G}\left[\beta^{(s)}\right]$ 3000 matrices drawn from either the prior or posterior distributions of $\mathbf{G}$ was calculated as

$$d^{(s)} = \sqrt{tr(\mathbf{G}_Q - \mathbf{G}\left[\beta^{(s)}\right])^2}. \tag{38}$$

Bayesian learning on similarity was evaluated by comparing the prior and posterior distributions of the Frobenius distances away from $\mathbf{G}_Q$. An overlap of distributions would indicate that phenotypes do not inform about similarity between individuals beyond what is conveyed by marker data only. Conversely, when phenotypes inform about similarity beyond markers, distances between $\mathbf{G}_Q$ and posterior similarities should be shorter than those between $\mathbf{G}_Q$ and prior similarities.

4.2. Results

Fig. 1 shows overlap between posterior and prior distributions of Frobenius distances when training data set size was 500, but draws from the posterior distribution were closer to $\mathbf{G}_Q$ than prior draws. As the size of the training data set increased to 2000 (Fig. 2) and 4000 (Fig. 3) the overlap between posterior and prior distributions disappeared. Frobenius distances based on posterior

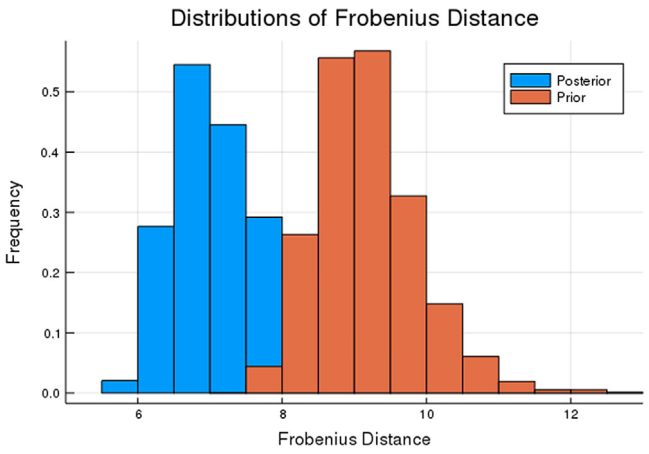

Fig. 1. Distribution of Frobenius distances away from true relatedness at the QTL level of similarity matrices drawn from prior and posterior distributions. Simulated data; $n = 500$.

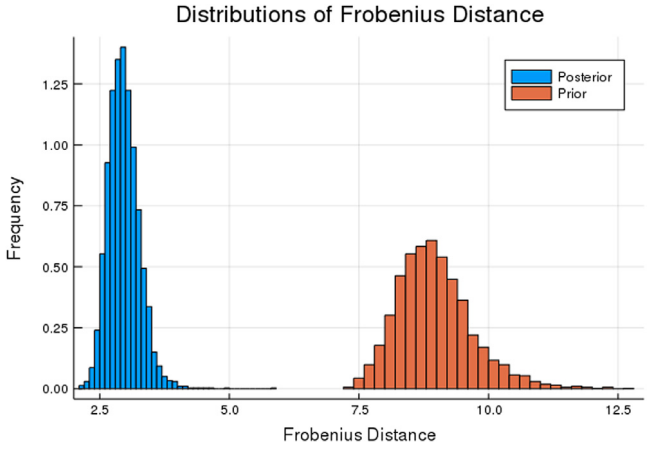

Fig. 2. Distribution of Frobenius distances away from true relatedness at the QTL level of similarity matrices drawn from prior and posterior distributions. Simulated data; $n = 2000$.

draws were closer to similarity at the QTL level than those based on samples from the prior; further, the posterior distribution of distances was sharper than the prior distribution. Note, however, that even though the QTL were in the marker panel, learning was still imperfect, as 0 was not assigned appreciable probability. In the limit, as n tends to infinity, the posterior distribution of $\mathbf{G}(\beta)$ is expected to converge to $\mathbf{G}_Q$, the genetic similarity at the QTL level, since QTL genotypes are included in the marker panel. The results corroborate, at least for Bayes $C\pi$, that when a training data set has sufficient size, the phenotypic data provide information beyond markers on trait-specific genetic similarities between individuals.

5. *Pinus taeda* data

The Loblolly (*Pinus taeda*) data described in Cheng et al. (2018) was employed to test if Bayesian learning could be attained in a real data set. After edits, there were $n = 807$ individuals with $p = 4828$ SNP markers with phenotypic measurements on Rust bin scores (presence or absence) and Rust gall volume, two disease traits; see Cheng et al. (2018) for additional information on these two disease traits.

A bivariate Bayes $C\pi$ method was used as in Cheng et al. (2018); these authors assumed bivariate normality for the sampling model. The MCMC scheme had a burn-in period of 10,000

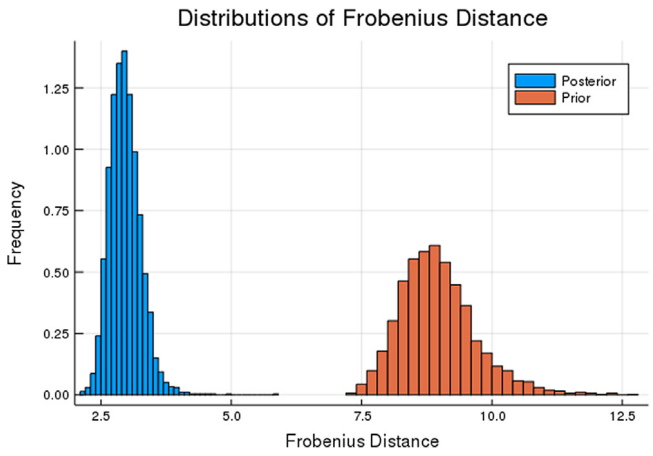

Fig. 3. Distribution of Frobenius distances away from true relatedness at the QTL level of similarity matrices drawn from prior and posterior distributions. Simulated data; $n = 4000$.

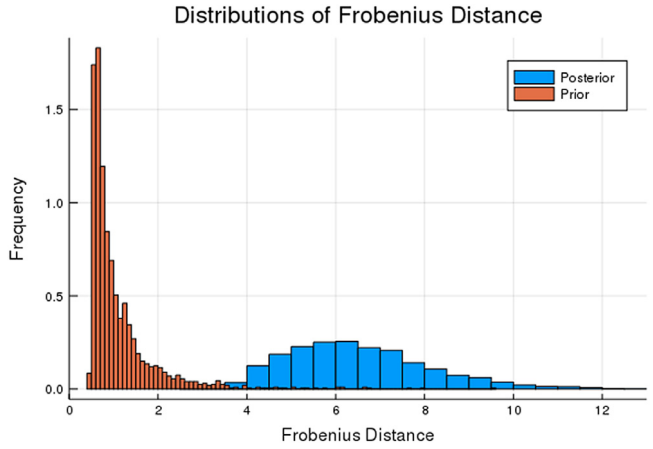

Fig. 4. Distribution of Frobenius distances between similarity matrices drawn from prior and posterior distributions. *Pinus* data; gall volumes.

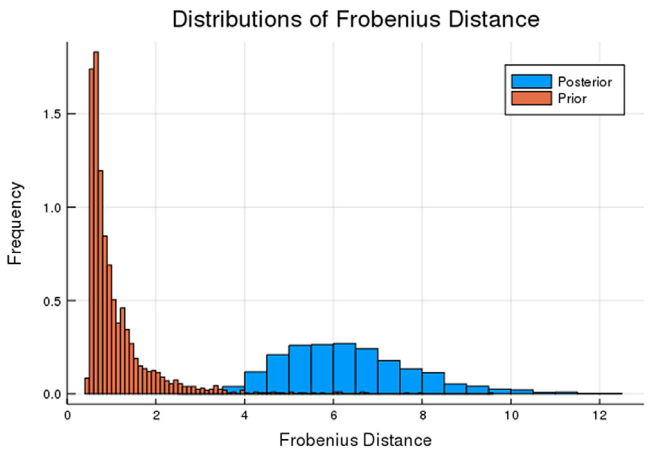

Fig. 5. Distribution of Frobenius distances between similarity matrices drawn from prior and posterior distributions. *Pinus* data; rust bin.

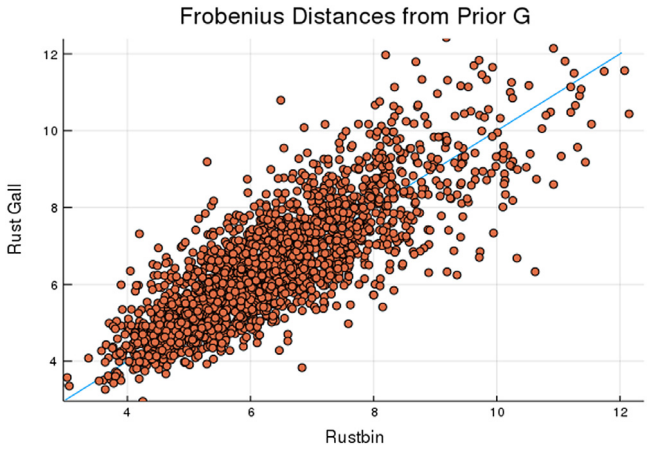

Fig. 6. Scatter plot of Frobenius distances away from the prior of posterior similarity matrices for rust bin vs. gall volume: *Pinus* data.

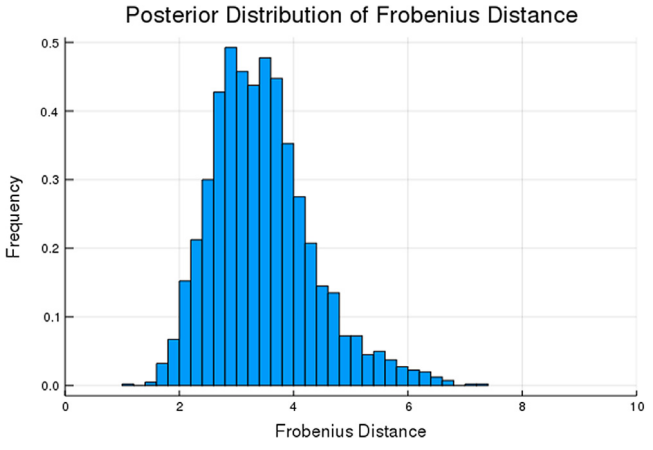

Fig. 7. Distribution of difference in Frobenius distances between rust bin and gall volume specific similarity matrices: *Pinus* data.

iterations, with an additional 100,000 samples drawn from the posterior distribution, and sub-sampled every 50th round such that sample size for inference was 2000. Further, 2000 samples were drawn from the prior distribution of the similarity matrices. Frobenius distances between $\mathbf{G}_{VR}$ and prior or posterior draws of the similarity matrices were calculated, leading to the distributions shown in Figs. 4 and 5, for Rust gall volume and Rust bin, respectively. The distributions had practically no overlap, indicating that use of phenotypic information did modify knowledge about similarity from what was conveyed by markers only. Note that the posterior distributions of the Frobenius distance away from $\mathbf{G}_{VR}$ were similar between traits. As shown in Fig. 6, Frobenius distances varied concomitantly for the two traits. Fig. 7 displays the distribution of the Frobenius distance between the similarity matrices for the two traits. Since there was no appreciable probability mass near zero, the analysis supports the view that trait-specific similarities differed. Hence, $\mathbf{G}_{VR}$ is not necessarily a suitable prescription for all traits.

6. *Wheat* data

A wheat yield data set in the R package BGLR (Pérez and de los Campos, 2014) was used to evaluate specific aspects of our methodology. Crossa et al. (2010), Gianola et al. (2011), Long et al. (2011) and Gianola et al. (2016) have characterized this data set extensively in several studies of genome-enabled prediction procedures. Briefly, there are 599 wheat inbred lines, each genotyped with 1279 DArT (Diversity Array Technology) markers and each planted in four environments. Sample size is $n = 599$ and $p = 1279$ is the number of markers. The DArT markers are binary $(0, 1)$ and denote presence or absence of an allele at a

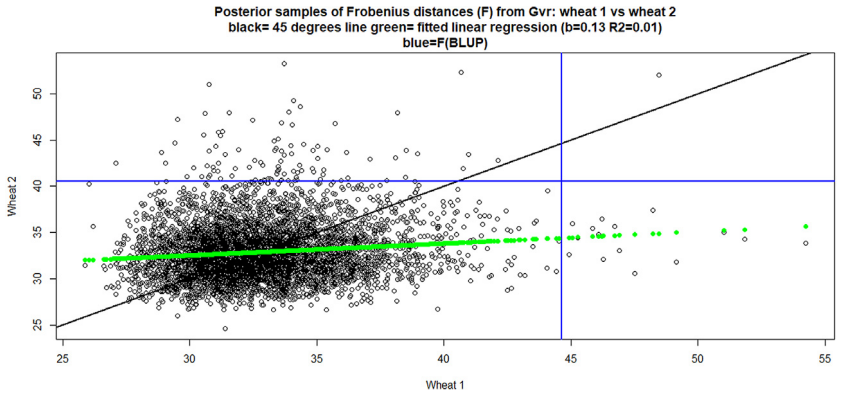

Fig. 8. Scatter plot of samples of Frobenius distances of $\mathbf{G}(\boldsymbol{\beta})$ away from the genomic relationship matrix $\mathbf{G}_{VR}$ for a bivariate Bayesian Lasso of wheat grain yield in environments 1 (X-axis) and 2 (Y-axis). The green line is the fitted regression line of Lasso samples for yield 2 on yield 1. Vertical and horizontal blue lines give the corresponding Frobenius distances for a bivariate GBLUP analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

marker locus in a given line. Grain yields in environments 1 and 2 were employed to compare outcomes between analyses based on bivariate GBLUP and the bivariate Bayesian Lasso (Gianola and Fernando, 2018). In the bivariate model, yields in the two environments are treated as distinct traits, conceptually. This type of data structure arises in dairy cattle-breeding (milk production of daughters of bulls in different countries treated as different traits) and in multi-environment situations in plant breeding.

Using the wheat data, we constructed similarity matrices derived from bivariate GBLUP and bivariate Lasso analyses. For GBLUP, the estimates of variance and covariance components were obtained via a maximum likelihood procedure; the bivariate Lasso model was implemented with a Gibbs/Metropolis–Hastings procedure based on six parallel chains leading to 12,000 samples post-convergence to the joint posterior distribution (Gianola and Fernando, 2018). The picture was similar to that obtained with the *Pinus* data set. Fig. 8 displays a scatter plot of 5000 randomly chosen posterior samples of Frobenius distances away from $\mathbf{G}_{VR}$ for wheat yield 1 and wheat yield 2. Clearly, the posterior samples of distances for the two traits were very weakly correlated; the green straight line depicts the fitted values of a linear regression of distances for trait 2 on trait 1 with a slope of 0.13 and $R^2 = 0.01$. The horizontal and vertical lines give the Frobenius distances between $\mathbf{G}\left(\widehat{\beta}\right)$ and $\mathbf{G}_{VR}$, where $\widehat{\beta}$ is the BLUP of marker effects for either trait 1 or trait 2. The location of the bivariate distribution of Lasso samples for the two traits was far from the centroid defined by the intersection of the distances between $\mathbf{G}\left(\widehat{\beta}_1\right)$ and $\mathbf{G}_{VR}$, and $\mathbf{G}\left(\widehat{\beta}_2\right)$ and $\mathbf{G}_{VR}$, where the subscript denotes the trait. Hence, Fig. 8 corroborates that phenotypes inform about trait-specific similarity.

7. FBLUP

All members of the Bayesian alphabet are self-contained from a predictive point of view. For instance, if regression coefficients β in the standard Bayesian linear model $\mathbf{g} = \mathbf{X}_{Trn}\beta$ are unknown and inferred from a training set, then in a testing set with known genotypes $\mathbf{X}_{Tst}$ the mean of the predictive distribution is $\mathbf{X}_{Tst}\overline{\beta}$. Here, $\overline{\beta}$ is the posterior expectation in the training set. On the other hand, our procedures allow to learn similarity matrices in testing sets with genotypes but without phenotypes. Such trait-specific similarity could be exploited further in a GBLUP context while producing prediction machines that are different from GBLUP. To illustrate the basic concept, let $\mathbf{X} = \left[\mathbf{X}'_{Trn} \quad \mathbf{X}'_{Tst}\right]'$ and define

$$\mathbf{G}_{all} = \begin{bmatrix} \mathbf{G}_{Trn,Trn} & \mathbf{G}_{Trn,Tst} \\ \mathbf{G}_{Tst,Trn} & \mathbf{G}_{Tst,Tst} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{Trn}\mathbf{D}\mathbf{X}'_{Trn} & \mathbf{X}_{Trn}\mathbf{D}\mathbf{X}'_{Tst} \\ \mathbf{X}_{Tst}\mathbf{D}\mathbf{X}'_{Trn} & \mathbf{X}_{Tst}\mathbf{D}\mathbf{X}'_{Tst} \end{bmatrix}, \tag{39}$$

Above, $\mathbf{D}$ is $p \times p$ and learned from training data only and the training process assigns differential weights to SNPs; let $\overline{\mathbf{D}}$ be the posterior expectation of $\mathbf{D}$ obtained from the training process. A pseudo-BLUP (FBLUP, F stands for "fake") procedure could be used to produce predictions of genetic values in the testing set as

$$\widehat{\mathbf{g}}_{F,Tst} = \mathbf{X}_{Tst}\overline{\mathbf{D}}\mathbf{X}'_{Trn}\left(\mathbf{X}_{Trn}\overline{\mathbf{D}}\mathbf{X}'_{Trn}\right)^{-1}\overline{\mathbf{g}}_{Trn} \tag{40}$$

where $\overline{\mathbf{g}}_{Trn}$ is the posterior expectation obtained from the Bayesian analysis of the training set. Note that $\overline{\mathbf{g}}_{Trn}$ and $\overline{\mathbf{D}}$ derive from training set data only, e.g., from fitting Bayes Cπ, R or any other variable selection method, while $\widehat{\mathbf{g}}_{F,Tst}$ is based on a BLUP logic. FBLUP has not been evaluated empirically yet, but the idea is intriguing, as it represents a "hybrid" between some Bayesian approach and BLUP with, perhaps, predictive synergy. FBLUP is not a linear procedure since neither $\overline{\mathbf{D}}$ nor $\overline{\mathbf{g}}_{Trn}$ are linear functions of the phenotypes.

The focus here is introducing the methodology, as opposed to conducting an exhaustive evaluation of its behavior as a way of generating novel prediction machines. Nevertheless, we compared the predictive ability of univariate GBLUP with that of univariate FBLUP using the data for wheat yield 1 and a training–testing set layout reconstructed at random 1000 times. In each training–testing instance, the training and testing sets had sizes $n_{Trn} = 450$ and $n_{Tst} = 149$, respectively. In FBLUP, $\mathbf{G}_{VR}$ was replaced by $\mathbf{G}\left(\widehat{\beta}_1\right)$ where $\widehat{\beta}_1$ represents the univariate BLUP estimates of marker effects on wheat yield 1 obtained from the appropriate training set, then used to form predictions evaluated against the appropriate testing sets. Fig. 9 displays, training (for GBLUP) and testing (GBLUP and FBLUP) correlations and mean-squared errors. As usual, training set correlations (between fitted values and targets) were larger than the corresponding predictive correlations; likewise, training mean-squared errors were smaller than the mean squared errors of prediction. The commonly negative relationship between fitting and predictive abilities was observed as well: the better the model fitted to the training data, the worse its predictions were. No clear differences in favor of FBLUP over GBLUP were noted. For instance, the median predictive correlations for GBLUP were 0.523 and 0.491 for traits 1 and 2, respectively; for FBLUP the corresponding median correlations were 0.518 and 0.486, respectively. Differences in predictive mean square errors were small as well. Since FBLUP can be used in conjunction with any method in which markers are learned, the term actually defines a new family of prediction machines.

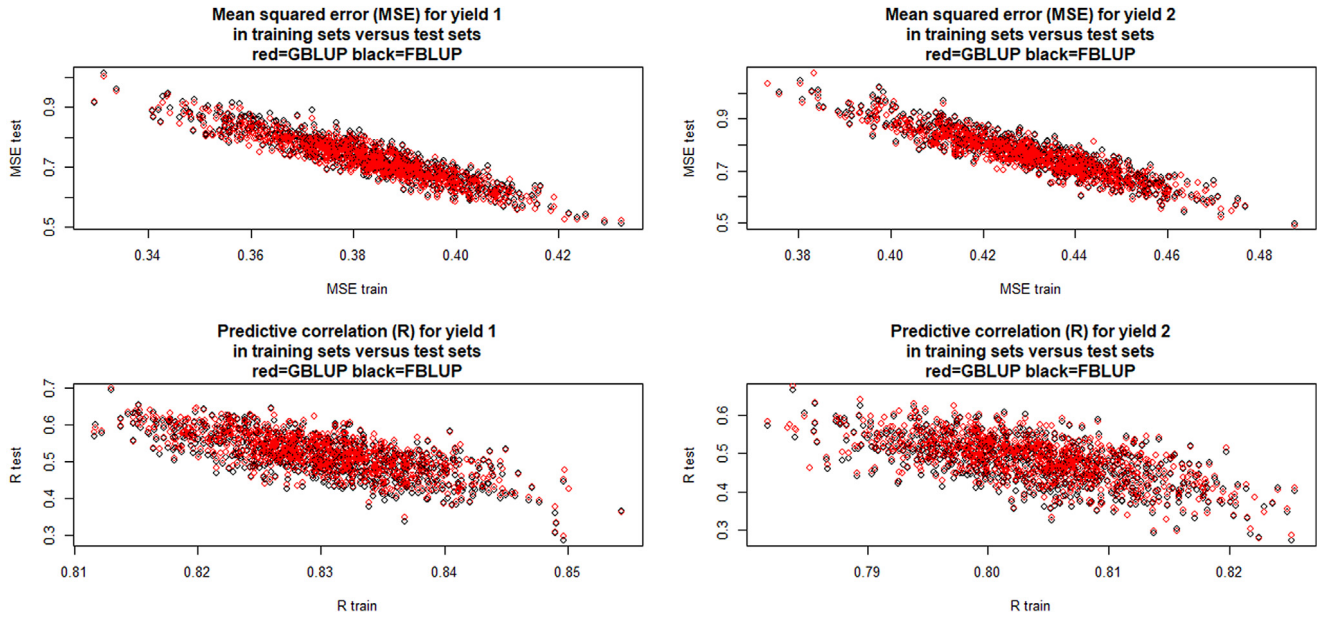

Fig. 9. Training and testing mean-squared errors (MSE) and predictive correlations (R) of GBLUP and FBLUP in 1000 random reconstructions of a training-testing layout in the what data set.

8. Discussion

Genomic relationship matrices (e.g., Van Raden, 2008) are a central part of GBLUP, arguably the prediction method most widely used in animal and plant improvement programs. In GBLUP, some $\mathbf{G}$ such as $\mathbf{G}_{VR}$ replaces the pedigree-based kinship matrix $\mathbf{A}$ used in classical BLUP. Van Raden's $\mathbf{G}_{VR}$ is based entirely on markers that are assigned the same weight, disregarding the possibility that some genomic region may be more important than others in trait determination. Also, linkage disequilibrium relationships among markers (beyond what is encoded by the columns of the genotype matrix $\mathbf{X}$) are not exploited in $\mathbf{G}_{VR}$. Bravington et al. (2016) assumed linkage equilibrium in kinship estimation in the context of mark-recapture studies. On the other hand, LD information was incorporated into a marker-based similarity suggested by Speed et al. (2012).

The construction of trait-specific $\mathbf{G}'s$ has been considered in previous research although not formally addressed as an estimation problem. Several studies, e.g., Zhang et al. (2010) and Wang et al. (2012), used information on phenotypes in the construction of $\mathbf{G}$ to produce trait-specific similarity (kernel) matrices. Fan et al. (2016) suggested an algorithm that iterates between fixed effects and random effects models that can be used to arrive at trait-informed relationship matrices. Although these methods have intuitive appeal, a cohesive theoretical foundation lacked. Here, we propose a formal Bayesian approach and implement it in an MCMC framework. The point of departure is to regard a similarity matrix $\mathbf{G}$ as an unknown parameter and define it as $\mathbf{G} = \mathbf{XDX}'$ where $\mathbf{D}$ is a diagonal matrix involving unknown quantities. For example, we let $\mathbf{G}(\beta)$ be a function of unknown effect sizes, define a prior density $p(\mathbf{G}(\beta)|\mathbf{X}, Hyp)$ and arrive at a posterior density $p(\mathbf{G}(\beta)|\mathbf{X}, \mathbf{y}, Hyp)$. The unknown similarity matrix can be inferred from MCMC draws from the posterior. The extent of Bayesian learning is assessed from the samples of Frobenius distances between matrices drawn from the prior and posterior distributions of $\mathbf{G}(\beta)$. It is important to note that the proposed method does not use information additional to what is conveyed by the regression model. Rather, it creates a connection between phenotypes and similarity which is ignored when the latter is inferred from markers only. Specifically, $\mathbf{G}(\beta)$ is treated as an unknown (non-linear) function of β, so given a posterior distribution of the vector of regression coefficients, a posterior distribution of the similarity matrix is induced automatically. The basic idea applies to all members of the Bayesian alphabet, such as, e.g., Bayes A, B, C, Cπ, R and L, since obtaining realizations of β is common to all such methods.

Using simulation, it was found that Bayesian learning on similarity does takes place and that the prior and posterior distributions of $\mathbf{G}(\beta)$ become increasingly distinct as sample size grows. A *Pinus* data set employed together with Bayes Cπ provided proof-of-concept. The distribution of Frobenius distances away from $\mathbf{G}_{VR}$ of prior and posterior samples suggested that phenotypes conveyed information on similarities beyond what is provided from markers, these holding for the two traits evaluated. Furthermore, there was evidence that the marker-phenotype informed similarity matrices differed between the two disease traits. A similar picture emerged from the analyses carried out with the wheat data set.

The Bayesian approach provides information about the uncertainty of similarities between any pair of individuals as it is easy to estimate the posterior distribution of any element of $\mathbf{G}(\beta)$. Such variability is expressed conditionally on $\mathbf{X}$ and $\mathbf{y}$ and it measures statistical uncertainty on similarity, but it is not comparable to the metrics presented by Hill and Weir (2011). These authors assumed conceptual repeated sampling of genotypes from a population. The distinction with the Bayesian interpretation is crucial, as some of the genome-enabled prediction literature is unclear with respect to what is fixed and what is random in marker-based models. Genetic variability and covariability is defined in quantitative genetics using the notion that allelic substitution effects are fixed and genotypes are random (Gianola et al., 2009; de los Campos et al., 2015; Gianola et al., 2015), as in Fisher's (2018) model. On the other hand, the Bayesian approach regards substitution effects as random and marker genotypes as fixed. In our paper, we treat $\mathbf{X}$ as a fixed matrix and the randomness in $\mathbf{G} = \mathbf{XD}(\beta)\mathbf{X}'$ derives from Bayesian uncertainty on $\mathbf{D}(\beta)$ which, in turn, depends on the uncertainty on model parameters.

Bayes Cπ was used for the *Pinus* data and a different method would have surely produced a different posterior difference of Frobenius distances away from $\mathbf{G}_{VR}$, perhaps overlapping from

the one produces by Bayes $C\pi$, perhaps not. As noted earlier, similarities based on the Bayesian Lasso applied to wheat grain yield differed from those based on a genomic similarity matrix derived from BLUP point estimates of marker effects. Bayesian learning about $\mathbf{G}$ is entirely dependent on how much is learned about β, which is a function of the trait-environment-method combination examined. For methods 1 and 2, say, and with an orthonormal regression model ($\mathbf{X}'\mathbf{X} = \mathbf{I}$) the squared Frobenius distance between the corresponding similarity matrices is

$$d^2(\mathbf{G}_1, \mathbf{G}_2) = tr\left[\left(\mathbf{XD}(\beta_1)\mathbf{X}' - \mathbf{XD}(\beta_2)\mathbf{X}'\right)^2\right] = tr\left[\mathbf{D}(\beta_1) - \mathbf{D}(\beta_2)\right]^2,$$

where β_1 is the vector of marker effects for method 1, $\mathbf{D}(\beta_1) = diag\left(\frac{\beta_{1j}^2}{\sum_{j=1}^{p}\beta_{1j}^2}\right)$ and similarly for method 2. Clearly, differences in similarity matrices depend on how marker effects are inferred by various methods, a question that is entirely context dependent. If a formal model comparison favors method 1, say, one could state that $\mathbf{G}_1$ provides a better inference of $\mathbf{G}$ than $\mathbf{G}_2$. However, if $\mathbf{G}_1$ is used in a prediction machine that performs better than one based on $\mathbf{G}_2$, this finding does necessarily imply that the inference of similarity based on 2 is better than the one based on 1. The constellation of complex traits (and the number of members of the Bayesian alphabet and of prediction methods) is enormous and it does not seem sensible to offer general prescriptions, especially from a prediction perspective.

Successful application of the concepts developed in this paper depends on the availability of a reliable MCMC implementation. For suitable inference, evidence must indicate that there has not been failure in reaching the target distribution and that sampling has been intensive enough, such that Monte Carlo error does not swamp statistical signals. If the prior distribution is a mixture, for example, so is the posterior (Albert, 2009), but detecting and attaining convergence (or lack thereof) with high-dimensional models is a challenge. With p markers and a four-component mixture, as in Bayes R, the joint posterior has $3p$ parameters so at least 3^p states must be visited, as there are three "free" indicator variables per marker. Rajaratnam and Sparks (2015) studied "convergence complexity" in $n \ll p$ situations and derived formulae that can be used to calculate the minimum amount of sampling required before reaching the equilibrium distribution. Calculations using such formulae (not shown here) suggest that convergence may be computationally difficult to attain when all polymorphic nucleotides available in a DNA sequence are used as covariates in a model, so much more sampling may be required than what is often done in practice. Conceivably, burn-in periods of at least half a million iterations may be needed. Our calculations support observations made by Celeux et al. (2000) on difficulties encountered in convergence of Bayesian mixture (variable selection) models. Inference must be done with caution when high-dimensional data are encountered, especially if sampling is not intensive enough. The approaches presented here do not escape from such pitfalls.

Acknowledgments

DG and CCS acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG; grant no. SCHO 690/4-1). DG was also partially supported by the Jay L. Lush Endowment at Iowa State University. RLF was supported by US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive Grant No. 2018-67015-27957. The Technical University of Munich (TUM) School of Life Sciences, the TUM Institute for Advanced Study, the University of Wisconsin-Madison, Iowa State University and the Institut Pasteur de Montevideo are thanked for providing office space, library and computing resources and logistic support.

Appendix. Similarity matrices for some members of the Bayesian alphabet

A.0.1. Bayes L

The Bayesian Lasso (Bayes L) assigns the same conditional double exponential distribution, with parameter λ, as prior to each of the regression coefficients; marker effects are assumed (given λ) independently distributed *a priori*. Bayes L can be parameterized and implemented in various manners (Park and Casella, 2008; Kärkkäinen and Sillänpää, 2012). Here, we adopt the scale-mixture of normals developed by Gómez-Sánchez-Manzano et al. (2007) and used by Gianola and Fernando (2018) in a multiple-trait generalization of Bayes L. The multi-trait Bayes L implementation in Gianola and Fernando (2018) reduces to a univariate Bayes L model when the number of traits is 1.

The hierarchy is: (1) $\beta_j|\sigma^2_{\beta_j}, \lambda \sim N\left(0, \sigma^2_{\beta_j}\lambda\right)$, and (2) $\sigma^2_{\beta_j} \sim Gamma\left(1, \frac{1}{8}\right)$ for all $j = 1, 2, \ldots, p$. Then

$$\mathbf{g} = \mathbf{X}\beta|\mathbf{X}, \sigma^2_\beta, \lambda \sim N\left(\mathbf{0}, \lambda\mathbf{X}\mathbf{V}_{BL}\mathbf{X}'\right), \tag{41}$$

where $\mathbf{V}_{BL} = diag\left\{\sigma^2_{\beta_j}\right\}$ and $\sigma^2_\beta = \left(\sigma^2_{\beta_1}, \sigma^2_{\beta_2}, \ldots, \sigma^2_{\beta_p}\right)$. The structure of the problem is as in Bayes A. The variance of a marker effect (the prior expectation is zero) is

$$Var\left(\beta_j|\lambda\right) = E_{\sigma^2_{\beta_j}} Var\left(\beta_j|\sigma^2_{\beta_j}, \lambda\right) = \frac{\lambda}{8} = \sigma^2_\beta \ \forall j, \tag{42}$$

because $E\left(\sigma^2_{\beta_j}\right) = \frac{1}{8}$ for all j, from properties of a $Gamma(1, \frac{1}{8})$ distribution. Hence, the genomic variance in the Bayesian Lasso (given λ) is

$$\sigma^2_g = \frac{\lambda}{8}\sum_{j=1}^{p} 2q_j(1 - q_j). \tag{43}$$

Given some estimate of σ^2_g and of allelic frequencies at marker loci, λ can be assessed using (43).

A similarity matrix can be defined from the hierarchical representation of Bayes L model. Let $\mathbf{D}_{BL} = diag\left\{\sigma^2_{\beta_j}/\sum_{j=1}^{p}\sigma^2_{\beta_j}\right\}$, and

$$\mathbf{G}_{BL} = \mathbf{X}\mathbf{D}_{BL}\mathbf{X}'. \tag{44}$$

The prior expectation of $\mathbf{G}_{BL}$ is approximately

$$E\left(\mathbf{G}_{BL}|Hyp\right) \approx \frac{1}{p}\mathbf{X}\mathbf{X}' = H\mathbf{G}_{VR}, \tag{45}$$

similar to (20) in Bayes A.

With S samples obtained from the prior and posterior distributions of $\mathbf{G}_{BL}$, the distribution of the Frobenius distance can be estimated by forming the draws

$$dist^{(s)}_{BL} = \sqrt{tr\left(\mathbf{X}\mathbf{D}^{(posterior,s)}_{BL}\mathbf{X}' - \mathbf{X}\mathbf{D}^{(prior,s)}_{BL}\mathbf{X}'\right)^2}; s = 1, 2, \ldots, S. \tag{46}$$

Sampling from the prior is straightforward since $\sigma^2_{\beta_j} \sim Gamma\left(1, \frac{1}{8}\right)$.

A.0.2. Bayes B and Bayes Cπ

Bayes B (Meuwissen et al., 2001) and Bayes C and $C\pi$ (Habier et al., 2011) pose a two-component mixture model as prior distribution. All marker effects follow the same prior distribution and are regarded as independently distributed, *a priori*. In Bayes B, the first component consists of a null effect with known prior

probability π. The second component assigns a $t-$distribution to each marker effect, with parameters as in Bayes A and with probability $1-\pi$. Bayes C also assumes known π but differs from Bayes B in that, for the second component, a multivariate $t-$distribution with null mean is assigned as joint prior to the non-null effects; the scale matrix in Bayes C is $\mathbf{I}S^2_\beta$ and the degrees of freedom are as in Bayes B. A subtle but important difference, is that in Bayes B effects are independent *a priori*, but this is not so for Bayes C. In the latter, marker effects are uncorrelated but not independent, *a priori*. Finally, Bayes $C\pi$ has an additional layer in the hierarchy, with a prior distribution assigned to π. In what follows, we take Bayes B as prototype, as the treatment in Bayes C is similar.

Using properties of mixtures (e.g., Gianola et al., 2006), the marginal prior distribution of a marker effect in Bayes B has mean zero and variance $\sigma^2_\beta = (1-\pi)\, S^2_\beta \frac{\nu_\beta}{\nu_\beta - 2}$. If hyper-parameter values were set as in Bayes A, the prior variance would be smaller in Bayes B because of the strong assumption that a fraction π of the markers has a null prior variance. The vector of genomic values in Bayes B has mean and covariance matrix

$$\mathbf{g} = \mathbf{X}\beta|\pi, S^2_\beta, \nu_\beta \sim \left[\mathbf{0}, \mathbf{X}\mathbf{X}'(1-\pi)\, S^2_\beta \frac{\nu_\beta}{\nu_\beta - 2}\right]; \tag{47}$$

the distribution of $\mathbf{g}$ is unknown and cannot be written in closed form. The Bayes B genomic variance is structured as

$$\sigma^2_g = (1-\pi)\, S^2_\beta \frac{\nu_\beta}{\nu_\beta - 2} \sum_{j=1}^{p} 2q_j(1-q_j) = pH\,(1-\pi)\, S^2_\beta \frac{\nu_\beta}{\nu_\beta - 2} \tag{48}$$

The original formulation of Bayes B (Meuwissen et al., 2001) uses an implementation based on variance parameters. Given $\sigma^2_{\beta_1}, \sigma^2_{\beta_2}, \ldots, \sigma^2_{\beta_p}$ (see hierarchical model for Bayes A) a marker effect has variance $\sigma^2_{\beta_j}$ with probability $1-\pi$, or variance 0 with probability π. Put

$$\mathbf{D}_{BB} = diag\left(I_j\right);\ \mathbf{G}_{BB} = \mathbf{X}\mathbf{D}_{BB}\mathbf{X}' \tag{49}$$

where I_j takes the value $\frac{\sigma^2_{\beta_j}}{\sum_{j=1}^{p}\sigma^2_{\beta_j}}$ $(j = 1, 2, \ldots, p)$ with prior probability $1-\pi$ or 0 with probability π. The Bayes B MCMC sampler produces an indicator of whether 0 or the realized (normalized) value of $\sigma^2_{\beta_j}$ is placed on position j along the diagonal of $\mathbf{D}_{BB}$. In Bayes $C\pi$, the mixing proportion varies at random because π is unknown. *A priori,* in Bayes B

$$E\left(\mathbf{G}_{BB}|\mathbf{X}, \pi\right) \approx \mathbf{X}\left[\frac{(1-\pi)}{p}\right]\mathbf{X}' = (1-\pi)\, H\mathbf{G}_{VR}. \tag{50}$$

Given S samples from each of the prior and posterior distributions of $\mathbf{D}_{BB}$, Frobenius distances can be compared as discussed previously. The posterior distribution of $\mathbf{G}_{BB}$ can be estimated from posterior samples

$$\mathbf{G}^{(s)}_{BB} = \mathbf{X}\mathbf{D}^{(s)}_{BB}\mathbf{X}';\ s = 1, 2, \ldots, S. \tag{51}$$

Bayes B can also be implemented as in a standard two-components mixture model where assignment to one of the two components of the distribution pertains to an effect and not to a variance. Use is made of data augmentation by including binary (0, 1) indicator variables $\delta = \left(\delta_1, \delta_2, \ldots, \delta_p\right)'$ denoting the component of the mixture responsible for the effect of a marker covariate on phenotypes. The scheme produces metrics for variable selection based on the posterior distribution of the $\delta's$. Given δ, the linear regression model (assuming no intercept) for datum i is

$$y_i = \sum_{j=1}^{p}\delta_j x_{ij}\beta_j + e_i = \sum_{j=1}^{p} x^*_{ij}\beta_j + e_i, \tag{52}$$

or, in matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\Delta}\beta + \mathbf{e} = \mathbf{X}^*\beta + \mathbf{e}, \tag{53}$$

where $\boldsymbol{\Delta} = diag(\delta_j)$ and $\mathbf{X}^* = \mathbf{X}\boldsymbol{\Delta}$. Conditionally on $\boldsymbol{\Delta}$,

$$\mathbf{g} = \mathbf{X}^*\beta|\boldsymbol{\Delta}, \sigma^2_\beta \sim \left(\mathbf{0}, \mathbf{X}\boldsymbol{\Delta}^2\mathbf{X}'\sigma^2_\beta\right), \tag{54}$$

where $\sigma^2_\beta = S^2_\beta \frac{\nu_\beta}{\nu_\beta - 2}(1-\pi)$; note that $\boldsymbol{\Delta}^2 = \boldsymbol{\Delta}$. Using (48) an unknown similarity matrix is

$$\mathbf{G}_{BB}(\boldsymbol{\Delta}) = \mathbf{X}diag\left\{\frac{\delta_j}{\sum_{j=1}^{p}\delta_j}\right\}\mathbf{X}'. \tag{55}$$

Since $\boldsymbol{\Delta}$ is unknown, $\mathbf{G}_{BB}(\boldsymbol{\Delta})$ is a random matrix. If $\delta_j \sim Bernoulli(\pi)$, $j = 1, 2, \ldots, p$ its prior expectation is

$$E\left[\mathbf{G}_{BB}(\boldsymbol{\Delta})|\mathbf{X}, \pi\right] = \frac{\mathbf{X}\mathbf{X}'}{p} = H\mathbf{G}_{VR}. \tag{56}$$

Now, $\boldsymbol{\Delta}$ in (53) and in (55) has a posterior distribution $\boldsymbol{\Delta}|\mathbf{y}$ that can be estimated from draws $\boldsymbol{\Delta}^{(1)}, \boldsymbol{\Delta}^{(2)}, \ldots, \boldsymbol{\Delta}^{(S)}$ obtained in an MCMC scheme. Thus, samples from the posterior distribution of $\mathbf{G}_{BB}(\boldsymbol{\Delta})$ can be formed as

$$\mathbf{G}^{(s)}_{BB}(\boldsymbol{\Delta}) = \mathbf{X}\boldsymbol{\Delta}^{(s)}\mathbf{X}' = \sum_{j=1}^{p}\delta^{(s)}_j\mathbf{X}_j\mathbf{X}'_j = \sum_{j=1}^{p}\mathbf{G}^{(s)}_{BB,j} \tag{57}$$

which shows clearly how variable selection affects similarity; $\mathbf{G}^{(s)}_{BB,j}$ is the contribution to similarity made by locus j in sample s. The posterior expectation of $\mathbf{G}_{BB,j}(\boldsymbol{\Delta})$ is estimated as

$$\widehat{E}\left(\mathbf{G}_{BB}(\boldsymbol{\Delta})|\mathbf{y}\right) = \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{p}\mathbf{G}^{(s)}_{BB,j} = \sum_{j=1}^{p}\overline{\delta}_j\mathbf{X}_j\mathbf{X}'_j, \tag{58}$$

where $\overline{\delta}_j$ is the estimated posterior expectation of δ_j, also an estimate of the probability of "inclusion" in the model. Since the prior probability of inclusion is $0 < \pi < 1$, the averaging process results in that all markers contribute to similarity, but to different degrees, as

$$\widehat{E}\left(\boldsymbol{\Delta}|\mathbf{y}\right) = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\Delta}^{(s)}. \tag{59}$$

The difference between similarity matrices drawn from their prior and posterior distributions is driven by the difference between the prior and posterior distributions of $\boldsymbol{\Delta}$. In Bayes B, π is taken as a known hyper-parameter, so draws from the prior distribution of $\boldsymbol{\Delta}$ can be obtained by drawing each of its elements from $\delta_j \sim Bernoulli(\pi)$. In Bayes $C\pi$ the draws from the prior distribution can be obtained from composition sampling: (1) draw π from its prior distribution (e.g., a uniform prior) and (2) conditionally on the realization π^*, draw $\delta_j \sim Bernoulli(\pi^*)$, repeating the two steps for each marker. The draws from the posterior distribution of $\boldsymbol{\Delta}$ follow directly from the MCMC algorithm; prior and posterior distributions can be compared by examining the distribution of Frobenius distances.

It is important to emphasize that (51) and (55) define two distinct similarity matrices. The prior and posterior distributions of (51) depend on the $\sigma^2_{\beta_j}$ parameters, on which limited Bayesian learning takes place (Gianola et al., 2009). The second matrix involves the δ_j indicators, which enable learning about π in Bayes $C\pi$.

A.0.3. Bayes R

Erbe et al. (2012) proposed a four-component mixture distribution as prior for each of the markers entering into the linear regression model and the procedure was termed Bayes R. The prior for the effect of marker j is

$$\beta_j|\pi_1,\pi_2,\pi_3,\pi_4,Hyp \sim \pi_1 N_1+\pi_2 N_2+\pi_3 N_3+\pi_4 N_4;\quad j=1,2,\ldots,p, \tag{60}$$

where N_i $(i=1,2,3,4)$ denotes a normal distribution with null mean and variance σ_i^2, $\pi = \{\pi_i\}$ is the vector of probabilities of membership. In Bayes R the hyper-parameters in *Hyp* include the variances of the four normal distributions, which are set arbitrarily. Actually, component 1 has null mean and variance in Bayes R so it represents a point-mass with probability π_1, producing a "spike and slab" model. In turn, the probabilities of membership π are assigned a four-dimensional Dirichlet prior distribution with hyper-parameters $\alpha=(1,1,1,1)'$, resulting in a uniform prior. Hence,

$$Var\left(\beta_j|\pi,Hyp\right)=\sum_{i=1}^{4}\pi_i\sigma_i^2, \tag{61}$$

recalling that $\sigma_1^2=0$. Because of the nullity of the means of the component distributions, averaging over the Dirichlet distribution yields

$$Var\left(\beta_j|Hyp\right)=E_\pi\left[Var\left(\beta_j|Hyp\right)\right]=\frac{1}{4}\sum_{i=1}^{4}\sigma_i^2=\sigma_\beta^2\vee j, \tag{62}$$

so the variance of the marginal prior distribution of a marker effect is an unweighted average of the variances of the mixture components. For Bayes R,

$$\mathbf{g}=\mathbf{X}\beta|Hyp\sim\left(\mathbf{0},\mathbf{X}\mathbf{X}'\sigma_\beta^2\right), \tag{63}$$

and the genomic variance takes the form

$$\sigma_g^2=\sum_{j=1}^{p}2q_j(1-q_j)\sigma_\beta^2. \tag{64}$$

As in any mixture model, the joint posterior can be augmented with indicator variables. In Bayes R, a 4×1 vector $\varphi_j=(\varphi_1,\varphi_2,\varphi_3,\varphi_4)'$, points to the distribution "responsible" for marker effect j; e.g., if marker j has a null effect, then $\varphi_j=(1,0,0,0)'$. From the point of view of learning a similarity matrix informed by phenotypes, what matters is whether a given marker has a null effect or not. Hence, one can define a binary indicator taking the value 1 if the marker has an effect $(\varphi_{j1}=0)$ and 0 otherwise, i.e., if either φ_{j2}, φ_{j3} or φ_{j4} take the value 1; let the binary indicator be γ_j $(j=1,2,\ldots,p)$. A similarity matrix for Bayes R would be

$$\mathbf{G}_{BR}(\Gamma)=\mathbf{X}'\Gamma\mathbf{X}, \tag{65}$$

where $\Gamma=diag\left(\gamma_j/\sum_{j=1}^{p}\gamma_j\right)$. *A priori,*

$$E\left[\mathbf{G}_{BR}(\Gamma)|Hyp\right]\approx\mathbf{X}'diag\left(\frac{\pi_1}{p\pi_1}\right)\mathbf{X}=H\mathbf{G}_{VR}, \tag{66}$$

where π_1 is the prior probability of a marker having a null effect on the trait. The MCMC sampler produces draws from the posterior distribution of each φ_j (and therefore, of each γ_j). The mean of the posterior distribution of $\mathbf{G}_{BR}(\Gamma)$ is estimated as

$$\widehat{\mathbf{G}}_{BR}(\Gamma)=\frac{1}{S}\sum_{s=1}^{S}\mathbf{X}'\Gamma^{(s)}\mathbf{X};\ s=1,2,\ldots,S. \tag{67}$$

At any round of the MCMC $\gamma_j^{(s)}=0$ if $\varphi_{j1}^{(s)}=1$, or $\gamma_j^{(s)}=1$ otherwise; $j=1,2,\ldots,p$. The difference between the prior and posterior distributions of Γ informs how much phenotypes affect similarity between individuals over and above markers. Equivalently, one can examine the distribution of Frobenius distances between pairs of matrices drawn from the prior and posterior distributions of $\mathbf{G}_{BR}(\Gamma)$.

References

Albert, J., 2009. Bayesian Computation with R, second ed. Springer, New York.

Bernardo, R., 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34, 20–25.

Bravington, M.V., Skaug, H.J., Anderson, E.C., 2016. Close-kin mark recapture. Statist. Sci. 2, 259–274.

Celeux, G., Hurn, M., Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distributions. J. Amer. Statist. Assoc. 95, 957–970.

Cheng, H., Garrick, D., Fernando, R., 2015. XSim: Simulation of descendants from ancestors with sequence data. In: G3: Genes, Genomes, Genetics, Vol. 5. pp. 1415–1417.

Cheng, H., Garrick, D.J., Fernando, R.L., 2016. JWAS: Julia implementation of whole-genome analyses software using univariate and multivariate Bayesian mixed effects models. http://QTL.rocks.

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., Fernando, R.L., 2018. Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics http://dx.doi.org/10.1534/genetics.118.300650.

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., et al., 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186, 713–724.

Day-Williams, A.G., Blangero, J., Dyer, T.D., Lange, K., Sobel, E.M., 2011. Linkage analysis without defined pedigrees. Genet. Epidemiol. 35, 360–370.

de los Campos, G., Gianola, D., Allison, D.A.B., 2011. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Rev. Genet. 11, 880–886.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193, 327–345.

de los Campos, G., Sorensen, D., Gianola, D., 2015. Genomic heritability: what is it? PLoS Genet. 11 (5), e1005048. http://dx.doi.org/10.1371/journal.pgen.1005048.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM Algorithm. J. R. Stat. Soc. B 39, 1–38.

Erbe, M., Hayes, B.J., Matukumali, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., Goddard, M.E., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95, 4114–4129.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics, fourth ed. Longmans Green, Harlow, UK.

Fernando, R.L., Cheng, H., Sun, X., Garrick, D.J., 2017. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. J. Anim. Breed. Genet. 134, 213–223.

Fisher, R.A., 1935. The fiducial argument in statistical inference. Ann. Eugen. 6, 391–398.

Fisher, R.A., 2018. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 53, 399–433.

García-Ruiz, A., Cole, J.B., Van Raden, P.M., Wiggans, G.R., Ruiz-López, F.J., et al., 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA 113, E3995–4004.

Gianola, D., 2013. Priors in whole genome regression: the Bayesian alphabet returns. Genetics 194, 573–596.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R.L., 2009. Additive genetic variability and the Bayesian alphabet. Genetics 187, 347–363.

Gianola, D., de los Campos, G., Toro, M.A., Naya, H., Schön, C.-C., Sorensen, D., 2015. Do molecular markers inform about pleiotropy? Genetics 201, 23–29.

Gianola, D., Fariello, M.I., Naya, H., Schön, C.-C., 2016. Genome-wide association studies with a genomic relationship matrix: a case study with wheat and Arabidopsis. In: G3: Genes, Genomes. Genetics. http://dx.doi.org/10.1534/g3.116.034256.

Gianola, D., Fernando, 2018. A multiple-trait Bayesian Lasso for genome-enabled association and prediction of complex traits. submitted for publication, https://www.biorxiv.org/content/10.1101/852749v1.

Gianola, D., Heringstad, B., Odegaard, J., 2006. On the quantitative genetics of mixture characters. Genetics 173, 2247–2255.

Gianola, D., Okut, H., Weigel, K.A., Rosa, G.J.M., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. http://dx.doi.org/10.1186/1471-2156-12-87.

Gómez-Sánchez-Manzano, Gómez-Villegas, M.A., Marín, J.M., 2007. Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. Comm. Statist. Theory Methods 37, 972–985, 2008.
Habier, D., Fernando, R., Dekkers, J.C.M., 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389–2397.
Habier, D., Fernando, R., Kizilkaya, K., Garrick, D.J., 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12, 185.
Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92, 433–443.
Henderson, C.R., 1976. A Simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32, 69–83.
Hill, W.G., Weir, B.S., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93, 47–64.
Karaman, E., Lund, M.S., Anche, M.T., Janss, L., Su, G., 2018. Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. http://dx.doi.org/10.1534/g3.118.200673, G3.
Kärkkäinen, H.P., Sillänpää, M.J., 2012. Back to basics for Bayesian model building in genomic selection. Genetics 191, 969–987.
Kempthorne, O., 1954. The correlation between relatives in a random mating population. Proc. R. Stat. Soc. B 143, 103–113.
Lehermeier, C., de los Campos, G., Wimmer, V., Schön, C.-C., 2017. Genomic variance estimates: with or without disequilibrium covariances? J. Anim. Breed. Genet. 134, 232–241.
Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative usage of fixed and random-effect models for powerful and efficient genome-wide association studies. PLoS Genet. http://dx.doi.org/10.1371/journal.pgen.1005767.
Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., 2011. Marker-assisted prediction of non-additive genetic values. Genetica 139, 843–854.
Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc, Sunderland, MA.
Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.
Nejati-Javaremi, A., Smith, C., Gibson, J.P., 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75, 1738–1745.
Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 681–686.
Pérez, P., de los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198, 483–495.
Rajaratnam, B., Sparks, D., 2015. MCMC-based inference in the era of big data: a fundamental analysis of the convergence complexity of high-dimensional chains. Technical Report, University of California, Davis.
Sethuraman, A., 2018. Estimating genetic relatedness in admixed populations. In: G3: Genes, Genomes, Genetics. http://dx.doi.org/10.1534/g3.118.200485.
Sorensen, D., Gianola, D., 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer, New York.
Speed, D., Hemani, G., Johnson, M.R., Balding, D.J., 2012. Improved heritability estimation from genome-wide SNPs. Amer. J. Hum. Genet. 91, 1011–1021.
Strandén, I., Garrick, D.L., 2009. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92, 2971–2975.
Sukhatme, P.V., 1938. On Fisher and Behrens' test of significance for the difference in means of two normal samples. Sankhyã 4, 39–48.
Sun, X., Qu, L., Garrick, D.J., Dekkers, J.C.M., Fernando, R.L., 2012. A fast EM algorithm for Bayes A-like prediction of genomic breeding values. PLoS ONE 7, e49157–9.
Thompson, E.A., 1975. The estimation of pair-wise relationships. Ann. Hum. Genet. 39, 173–188.
Van Raden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.
Van Raden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., et al., 2009. Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92, 16–24.
Visscher, P.M., Medland, S.E., Ferreira, M.A.R., Morley, K.I., Zhu, G., et al., 2006. Assumption free-estimation of heritability from genome-wide identity by descent sharing between full-siblings. PLoS Genet. 2, e41.
Walker, G.A., Saw, J.G., 1978. The distribution of linear combinations of t-variables. J. Amer. Statist. Assoc. 73, 876–878.
Walsh, B., Lynch, M., 2018. Evolution and Selection of Quantitative Traits. Oxford University Press, Oxford.
Wang, H., Misztal, I., Aguilar, Legarra, A., Muir, W.M., 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94, 73–83.
Wang, B., Sverdlov, S., Thompson, E., 2017. Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. Genetics 205, 1063–1078.
Wiggans, G.R., Cole, J.B., Hubbard, S.M., Sonstegard, 2017. Genomic selection in dairy cattle: the USDA experience. Ann. Rev. Anim. Biosci. 5, 309–327.
Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.J., et al., 2010. Best linear unbiased prediction of genomic breeding value using a trait-specific marker-derived relationship matrix. PLoS ONE 5, e12648–16.