http://www.ncbi.nlm.nih.gov/pubmed/27435734

WALKING THROUGH THE BLACK BOXES OF STATISTICAL BREEDING

Abstract: Intelligent decision making relies on our ability to extract useful information from data to help us achieve our goals more efficiently. Many plant breeders and geneticists perform statistical analyses without understanding the underlying assumptions of the methods or their strengths and pitfalls. In other words, they treat these statistical methods (software and programs) like black boxes. Black boxes represent complex pieces of machinery with contents that are not fully understood by the user. The user sees the inputs and outputs without knowing how the outputs are generated. By providing a general background on statistical methodologies, this review aims (1) to introduce basic concepts of machine learning and its applications to plant breeding; (2) to link classical selection theory to current statistical approaches; (3) to show how to solve mixed models and extend their application to pedigree-based and genomic-based prediction; and (4) to clarify how the algorithms of genome-wide association studies work, including their assumptions and limitations.

Keywords: Gaussian Process; Gibbs sampling; Machine Learning; Mixed Model; Kernel methods.

Introduction

Inferences and models can be either empirical or experimental in design. Empirical methods work best with well-characterized phenomena for which the solution can be found analytically, whereas making inferences from data and using algorithms to identify patterns in the data requires experimental methods. The science that studies these algorithms is known as machine learning. Machine learning also describes the area of artificial intelligence dedicated to building and studying algorithms that are capable of learning from data, endeavoring to find an optimal solution that minimizes a given loss. This makes these machine learning algorithms much more flexible than logical algorithms.

Genetics takes great advantage of two particular branches of machine learning, so-called *supervised* and *unsupervised* learning (Libbrecht and Noble 2015). Supervised learning helps solve problems for which there are both explanatory and response variables. This commonly applies to prediction, selection, and classification in quantitative genetics. Unsupervised learning is used when no response variable exists, for problems like clustering genotypes and finding admixture in populations.

Due to the quantitative nature of most traits of interest, the most-employed type of supervised learning algorithm in plant and animal breeding is Gaussian process (GP) (Rasmussen 2004, Lynch and Walsh 1998). Fisher's infinitesimal model, which forms the basis of the principles of breeding, states that an infinite number of stochastic processes (referring to genes) control the observed phenotype (Orr 2005, Farrall 2004), which converges to a Gaussian distribution according to the central limit theorem. GP represents the basis of selection theory, breeding values, and association studies (Sorensen and Gianola 2002).

In supervised learning procedures, prediction is important to improve quantitative traits and classification is important for decision making and the genetic improvement of categorical traits. Breeding programs develop specific products to meet the needs of a variety of markets (Acquaah 2009, Cleveland and Soleri 2002) and classification models determine the boundaries of the qualities that define these market niches (Lim 1997). In soybeans, adaptation zones define which maturity group (MG) can be cultivated in each region based on the latitude, soil, and climate; in other words, they determine the target environment for breeders (Dardanelli *et al.* 2006). Zhang *et al.* (2007) suggest that soybean adaptation zones have misclassification issues because the growing zone for MG IV to MG VI is much larger than originally thought.

The main goal of this paper is to reveal the inner workings of the black boxes of statistical analysis in plant breeding by explaining the theory and applications of statistical genetics, focusing on widely applied mixed linear models designed for breeding.

Gaussian Process

Quantitative traits all follow some sort of distribution pattern. For example, categorical traits with two classes follow a binomial distribution, as with the color of flowers in soybeans, which are either white or purple. Traits like grain yield and plant height are continuous and often follow a normal distribution. The heritability of traits can assume any value between zero and one, and thus a beta distribution best characterizes this process. Variance components have positive values on a continuous scale and, therefore, they can be described in terms of a chi-squared distribution.

Since most quantitative genetic theory assumes normality, it is particularly important to know how to handle a normal distribution in plant and animal breeding. The normal distribution has a sigmoidal nature, the expectation of any normal random variable X is its mean (μ), and the deviation from the expectation is the variance (σ^2). Once the parameters mean and variance are known, a probability density function (PDF, ϕ) can be used to infer the probability of observing any given event, such as the probability of finding a plant in a given population that yields **x** bu/ac. If this probability is computed for all individuals in the population, the product of these probabilities is the *likelihood* of the data for that particular mean and variance.

Calculating the probability of finding plants with yield equal to or lower than **x** requires a cumulative density function (CDF, Φ), which is the integral of the PDF. Figure 1 shows an example of these calculations for Gaussian traits.

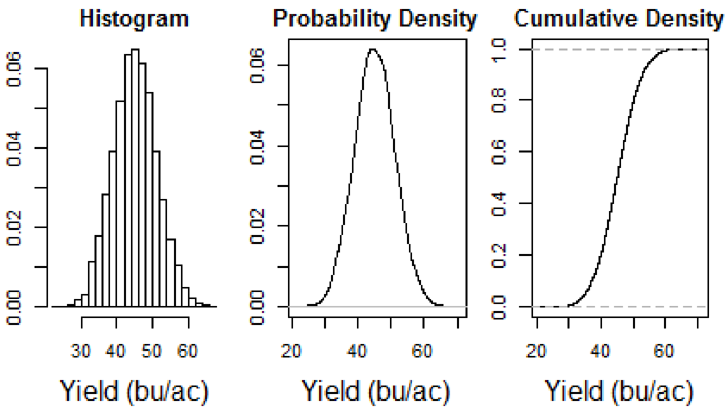


Fig. 1 Probabilistic description of the distribution of yield.

In all datasets, every observation contains some information about unknown parameters. Consequently, more data provides more accurate and precise estimates of mean and variance. There are a variety of methods to estimate the parameters of a distribution. These include likelihood methods and Bayesian procedures. Likelihood methods search for the unknown parameters that maximize the likelihood (L) of the observed data, using the PDF to define the joint probability of the data and parameters $p(\mathbf{X}, \boldsymbol{\theta})$, where $\mathbf{X}$ represents the observed data and the theta ($\boldsymbol{\theta}$) represents the unknown parameters. For a simple normal distribution, $\boldsymbol{\theta} = (\mu, \sigma^2)$. Bayesian procedures estimate parameters using probability distributions assigned to the unknown parameters, referred to as priors, in addition to the likelihood equation.

Infinitesimal Model and Selection Theory

For a normally distributed trait in a population, *directional selection* occurs when a breeder induces the mean to move in the desired direction over generations (Fig. 2). To achieve that, the breeder imposes a selection threshold. Individuals above this threshold are selected as the progenitors of the next generation under the assumption that those individuals provide better genetic properties (Recker *et al.* 2014).

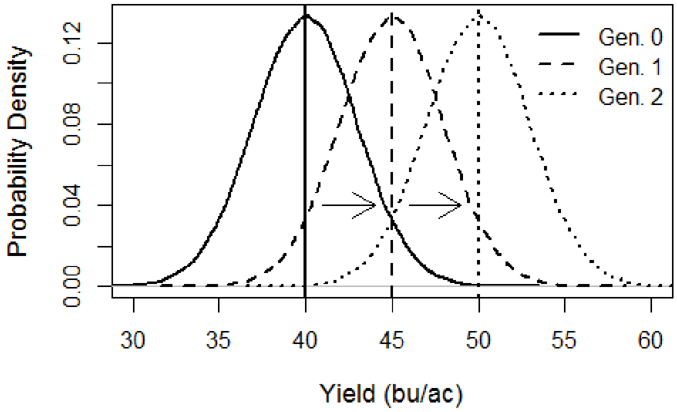


Fig. 2 Illustration of directional selection increasing the population mean over generations.

The genetic properties that affect the phenotype involve alleles with positive and negative effect. Alleles are versions of genes that represent the genetic effect on a given trait. Alleles can interact within the locus, across loci, and by external stimuli; these phenomena are known respectively as *gene action*, *epistasis*, and *expression*. The number of alleles carried by a locus depends on the ploidy level of the individual. Here we focus on diploid organisms, assuming two alleles at each locus.

The selection intensity (i) represents the number of standard deviations from the mean used as the cutoff for the population; in other words, the truncation point above which selected individuals remain in the breeding population as progenitors. This population of selected individuals represents a one-sided, truncated normal distribution. Computing the expectation of the truncated distribution (μ^*) uses the mean (μ) and standard deviation (σ) of the original distribution and, of course, the truncation point ($t = \hat{\mu} + i\hat{\sigma}$) (Wricke and Weber 1986), then estimates the expected mean of the selected population as:

$$E[\mu^*|t] = \mu + \sigma\left[\frac{\phi(i)}{1 - \Phi(i)}\right]$$

where ϕ, Φ, and i respectively represent the normal PDF, CDF, and selection intensity as shown in Figure 3.

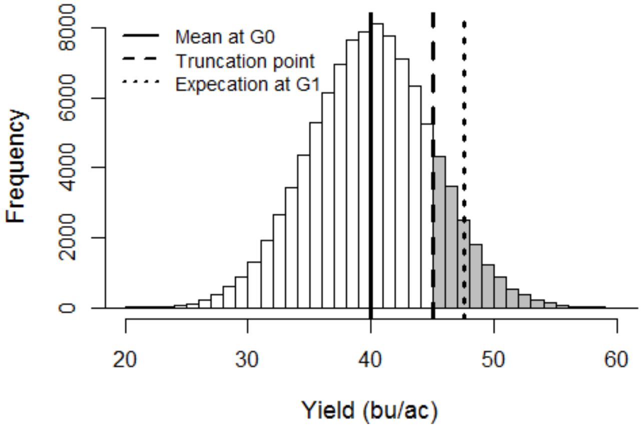


Fig. 3 Scheme of directional selection: Histogram of yield with mean $\hat{\mu} = 40$ and standard deviation $\hat{\sigma} = 5$, expected mean in the next generation $\hat{\mu}^* = 47.6$ and truncation point $t = 45$ for selection intensity $i = 1$. Shaded bars represent the progenitors of the following generation.

The next generation will not have the expected mean μ^*, since the phenotype is not exclusively due to genetic factors (Nyquist and Baker 1991). Despite the fact that alleles interact in a very complex fashion, the observed phenotype is an expression of genetic factors interacting with environmental stimuli (also known as genotype-by-environment interaction). Hence the *realized heritability* (h_r^2) is defined as the ratio between the observed mean of the new generation ($\mu^{(t+1)}$) and its expectation (μ^*) based on the selected progenitors. Realized heritability is not constant across generations. Rather, the selection pressure applied over time in a finite population imposes a major trade-off between the response to selection and genetic gains (Fig. 4).

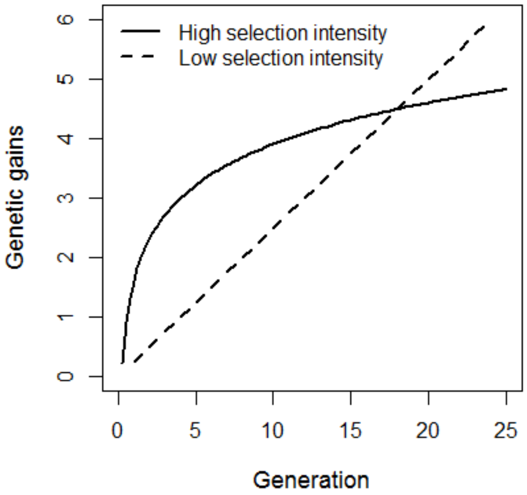


Fig. 4 Illustration of the consequences of high and low selection intensity on genetic gains over generations of selection for a given quantitative trait.

Fisher (1918) proposed that, for a given quantitative trait, there are an infinite number of genes with minor additive contributions that affect the phenotype, the so-called infinitesimal model. In selection theory, the general goal of breeders is to increase the frequency of desirable alleles in a population over time, under the assumption that the allele effect works in an additive fashion. Exceptions to this method include the gains associated with heterosis as exploited by programs that develop hybrids (like maize), or by clonally propagated species (such as potato). According to Fisher's model, the outcome of each gene is additive and is measured by the effect of an allelic substitution. This model is defined as a Gaussian process arising from normally distributed elements of an infinite-dimensional space (also known as Hilbert spaces), where each infinitely small fraction of the genome represents a parameter or dimension.

When applied to finite breeding populations, Fisher's model encounters population genetic issues. For example, finite populations can maintain only a limited number of alleles (Kimura and Crow 1964). Furthermore, multiple evolutionary forces will be acting simultaneously, such as various types of selection and long-term random genetic drift, and this triggers continuous bottlenecks (Wright 1930). This extension of the infinitesimal model for finite populations is called the Wright-Fisher model. The breeding populations of most crops follow the definition of a stochastic Fisher-Wright process (Imhof and Nowak 2006): finite populations with non-overlapping generations, diploid behavior, and ongoing selection.

Crow and Kimura (1970) pointed out that fluctuations that Fisher had defined as noise, Sewall Wright defined as (a slow) evolution. The stability of genetic gain over time relies on selection intensity, mutation rate, and both the total (n) and effective ($\mathrm{N_e}$) population size.

Effective population size is a major limiting factor for efficient selection in plant breeding programs, with serious implications for traditional and genomic-based selection techniques (MacLeod *et al.* 2014). According to Zeng and Hill (1986), the optimal selection intensity occurs when new haplotypes arise at the same frequency with which alleles undergo fixation, such that the population does not exhaust its diversity.

Self-pollinated species are more likely to run out of genetic resources due to their reproductive nature. For example, the effective population size of soybeans in the United States is equivalent to 27 lines (St. Martin 1982) and, not surprisingly, soybean production has nearly reached a yield plateau (Egli 2008a) that is approximately half of the field potential (Specht *et al.* 1999) due to these limited genetic resources (Egli 2008b). Yet, new breeding tools in the *"omics generation"* may improve gains in this currently limited scenario (Rincker *et al.* 2014).

Variance Decomposition and Parsimony

The phenotype of a quantitative trait is in a non-deterministic state. It requires a stochastic model to approximate an infinite population; in other words, a model with random variables defining the variance components of interest. The first model to express variations in phenotype was Fisher's infinitesimal model, in which the phenotypic variance (σ^2_{y}) is a function of genetic (σ^2_{G}) and environmental variances (σ^2_{E}), so that $\sigma^2_{\mathrm{P}} = \sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{E}}$.

Variance component analysis (VCA) is very common in plant breeding and agronomic studies. Two of the most common methods of variance decomposition are the analysis of variance (ANOVA) and restricted maximum likelihood (REML) calculations. Studying the variance due to genotype and environment in soybeans, Carvalho *et al.* (2008) suggested that both methods would provide similar variance components under a balanced experimental design but, under unbalanced conditions, ANOVA methods become biased while REML still provides consistent variance components and the best linear unbiased predictions (BLUPs) (Henderson 1975). This makes REML procedures the most deployed method of VCA in breeding studies, with BLUPs used for variety selection (Piepho *et al.* 2008).

In the infinitesimal model, all variation not explained by genetics is due to environment. In plant breeding, replications allow us to measure the variation due to environment, enabling further decomposition of the variance of the phenotype. Thereby it is possible, for example, to estimate the interaction between genotype and environment ($\sigma^2_{\mathrm{G\times E}}$) and isolate the pure error (σ^2_{ε}). Each term can undergo further decomposition. For example, environmental variance can include year (σ^2_{Y}), location (σ^2_{L}), and management (σ^2_{M}), a component that reflects the controllable environment. Yan and Rajcan (2003) conducted a genotype-by-environment analysis in soybeans, decomposing σ^2_{E} into σ^2_{Y} and σ^2_{L} with all possible interaction terms (ie. $\sigma^2_{\mathrm{G\times Y\times L}}, \sigma^2_{\mathrm{G\times L}}, \sigma^2_{\mathrm{G\times Y}}$), in which they concluded that most variance associated with environment is due to year rather than location.

If genotypic information is provided by genotyping with co-dominant molecular markers, such as single nucleotide polymorphism (SNP), then breeders and geneticists are able to subdivide genetic variance terms as well (Xu 2013). The first decomposition of genetic variation yields the additive genetic variance (σ_{A}^2), the dominance genetic variance (σ_{D}^2), and epistasis (σ_{I}^2). Epistasis represents the interaction among loci, including: additive-by-additive (σ_{AA}^2), additive-by-dominant (σ_{AD}^2), and dominant-by-dominant (σ_{DD}^2).

At this point, it is very important to introduce two concepts: *narrow-* (h^2) and *broad-sense* (H) heritability (Acquaah 2009). In statistics, heritability is known as the intra-class correlation coefficient, which refers to the amount of total variation due to one of its components. Broad-sense heritability is the amount of variation due to genetics ($\mathrm{H} = \sigma_{\mathrm{G}}^2/\sigma_{\mathrm{P}}^2$), also known as *repeatability* (Nyquist and Baker 1991). It illustrates 'nature-versus-nurture,' distinguishing between variation due to genetics and that due to environment. Narrow-sense heritability is the portion of phenotypic variance due to the additive genetic variance only ($\mathrm{h}^2 = \sigma_{\mathrm{A}}^2/\sigma_{\mathrm{P}}^2$), which is associated with the variance transmitted over generations. The latter is important for breeding quantitative traits because it describes how accurately breeding values correspond to the phenotype.

Estimation of the genetic variance component starts by defining the relationship among individuals using a kernel matrix (aka. Kinship matrix). This matrix is a symmetric, square matrix where each cell indicates the relationship between each pair of individuals. The matrix is then used to solve the Henderson's equation (Henderson 1984), a mixed model framework that accommodates terms with independent and non-independent treatments and observations, with the interdependence among observations expressed by the kernel matrix.

Analysis uses multiple types of kernel matrices ($\mathbf{K}$) to represent the relationship among random effects (ie. genotypes). The simplest scenario assumes that random effects are independent, in which case the kernel is then expressed by an identity matrix ($\mathbf{K} = \mathbf{I}$). With regard to non-independent effects, the best known kernels include the pedigree relationship matrix kernel (Wright's 1922), the genomic relationship matrix (VanRaden 2008), and spatial kernels (Piepho 2009).

Kernels used for genomic analysis are built from the genotypic information matrix ($\mathbf{M}$). This matrix has dimensions $q \times m$, where each row (q) represents a genotype and each column (m) represents a molecular marker. Thus, each cell in this matrix represents the locus of a given individual, and each locus is numerically coded to represent {AA, Aa, aa}. Many genomic analysis require specific allele coding for correct interpretation of the results (Strandén and Christensen 2011). For example, the G2A model (Zheng et al. 2005) proposes the centralized allele coding {2q, 1-2p, -2p} to preserve the orthogonality between main and epistatic effects. Table 1 presents some classical set ups for allele coding.

Table 1 Common genomic kernels computed from genomic data, where $\mathbf{M}$ is the genotypic information matrix and $\mathbf{E}$ is the Euclidean distance matrix.

Nature of the kernel	Coding {AA, Aa, aa}	Normalization (α)	Solution
Additive (Linear)	{-1, 0, 1}	$q\ \mathrm{tr}(\mathbf{MM}')^{-1}$	$\alpha\mathbf{MM}'$
Dominance (Linear)	{0, 1, 0}	$q\ \mathrm{tr}(\mathbf{MM}')^{-1}$	$\alpha\mathbf{MM}'$
Add x Add (Linear)	{-1, 0, 1}	$q\ \mathrm{tr}(\mathbf{MM}'\#\mathbf{MM}')^{-1}$	$\alpha(\mathbf{MM}'\#\mathbf{MM}')$
Gaussian (Non-linear)	{-1, 0, 1} or {0, 1, 2}	$\mathrm{Median}(\mathbf{E}^2)^{-1}$	$\exp(-\alpha\mathbf{E}^2)$
Exponential (Non-linear)	{-1, 0, 1} or {0, 1, 2}	$\mathrm{Median}(\mathbf{E})^{-1}$	$\exp(-\alpha\mathbf{E})$

A same model can include multiple genetic terms in order to decompose the total genetic variance using multiple kernels (additivity, dominance, and epistasis). Although it is possible to add as much complexity to the variance decomposition model as desired (Akdemir and Jannink 2015), researchers must take into account two statistical principles: the *hierarchical principle* and the *sparsity principle*. The hierarchical principle states that lower-order terms are generally more important than higher-order ones. In other words, epistasis may contribute little to the total genetic variance and at a high computational cost. The sparsity principle involves statistical parsimony, in which few terms explain most variation. Sparsity plays an important role in genomic analysis because in practical terms not all of the genome contributes to all traits, but rather a reduced number of regions contribute most. These regions are known as quantitative trait loci (QTL).

Lander and Botstein (1989) proposed that the phenotypic variance of quantitative traits was not a single normal distribution, but a Gaussian process consisting of a mixture of distributions associated with the combination of multiple QTL (Fig. 5).

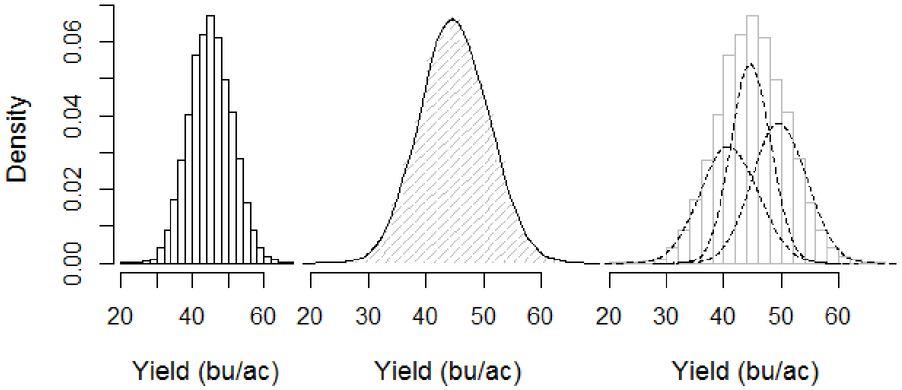


Fig. 5 Histogram of yield (left) illustrating the distribution of a quantitative trait as a single normal distribution (center) compared to a mixture of normal distributions (right) as proposed by Lander and Botstein (1989).

Identifying and locating QTL is extremely important in quantitative genetics. QTL discovery works by comparing the likelihood of two models (Yan *et al.* 2014). The first is the null model (l_0), which contains only the polygenic term defined by an additive kernel (Xu 2013, de los Campos *et al.* 2010). The second is the full model (l_1), that includes the candidate genomic fraction (marker or region) in addition to the polygenic term. The statistical test for this analysis is the likelihood ratio test (LRT), simply calculated as the ratio l_0: l_1. The results can be expressed in terms of the LRT itself, as p-values (LRT~$\chi^2_{v=1}$), or as a logarithm of odds (LOD score) by dividing LRT by 4.61 (Lynch and Walsh 1998).

QTL mapping occurs in both experimental and random populations. There are two major methods to find QTL: *linkage mapping* and *association mapping*. Linkage mapping is a method of tracking QTL as a map function of known genetic distance between markers. It is commonly performed in experimental populations designed for this purpose, with no need for the polygenic term in either the full or reduced models. Association mapping is a test of single markers across the whole genome for experimental or random populations that provides extra scrutiny for the existence of subpopulations.

In both methods, undetected regions will bias the number of QTL downward, and the average effect of QTL upward due to a phenomenon known as the *Beavis effect*. This occurs because the precision and accuracy of finding real QTL relies extensively on the population size (Beavis 1998) and implicit assumptions associated with the population type (Xu 2003a, Nyquist and Baker 1991).

Breeding Values and Variance Components

Only a small fraction of lines developed in breeding programs are released as commercial lines, with selection based on the top-performing genotypes. However there is always more than one trait of interest, so selection can take several forms: one trait at a time (ie. tandem selection), multiple quantitative traits simultaneously (ie. independent culling), or on the combination of traits (ie. index selection). In addition, there are three metrics to evaluate a quantitative trait: phenotypic value, genetic value, and breeding value. While selection based on phenotypic values uses the phenotypes in a straightforward manner, estimation of genetic and breeding values requires the implementation of mixed models.

Mixed model theory is the life's work of the geneticist Charles Henderson, who was motivated to implement and apply Wright's pedigree kernel matrix to breeding and selection. This theory is the foundation of modern genomic prediction methods. A linear model represents a mixed effect when the response variable ($\mathbf{y}$) is a function of a fixed effect term ($\mathbf{Xb}$) and one or more random effects ($\mathbf{Zu}$) other than the residuals ($\mathbf{e}$). The common notation of a mixed model is given by:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$
$$\mathbf{u} \sim \mathrm{N}(0, \mathbf{K}\sigma_{\mathrm{a}}^2)$$
$$\mathbf{e} \sim \mathrm{N}(0, \mathbf{R}\sigma_{\mathrm{e}}^2)$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices of dimensions $n \times p$ and $n \times q$, for fixed and random effects respectively. For these matrices, n represents the number observations, p the number of fixed effect parameters (blocks, covariates, etc.), and q the number of random effect parameters, in this case the number of genotypes. The regression coefficients of fixed and random effects $\mathbf{b}$ and $\mathbf{u}$ are vectors of length p and q. Random term and residual variances are notated as σ_{a}^2 and σ_{e}^2. Matrices $\mathbf{K}$ and $\mathbf{R}$ are the kernels of random effects ($q \times q$) and residuals ($n \times n$) used to define the relationship among random effects (ie. genotypes) and observations, respectively. For this model, the covariance of $\mathbf{y}$ is expressed by the covariance matrix ($\mathbf{V}$), as a function of the random and residual terms ($\mathbf{V} = \mathbf{ZKZ}'\sigma_{\mathrm{a}}^2 + \mathbf{R}\sigma_{\mathrm{e}}^2$). An example of the design matrix appears in the appendix.

A common assumption of linear models is to consider residuals to be independent ($\mathbf{R} = \mathbf{I}$). Yet, the residual relationship matrix can provide a powerful way of dealing with correlated residuals (ie. heteroscedasticity). For example, it is possible to use the $\mathbf{R}$ matrix to inform the model of the pairwise distance among field plots (ie. kriging). This allows us to acknowledge that there might be certain spots in the field where field plots can perform better than others without necessarily knowing where these spots are.

A remarkable property of random effects is the shrinkage that regularizes the regression coefficients based on their contribution to the model. The regularization parameter is notated by lambda (λ), which is analytically defined as the ratio between residual variance and random term variance ($\lambda = \sigma^2_e/\sigma^2_a$), such that the shrinkage of the genetic term is inversely proportional to the heritability of the trait. Thus $h^2 = (1+\lambda)^{-1}$.

The simplest case of a mixed model with a non-independent random term is the so-called animal model. The animal model is Henderson's implementation of Fisher's variance decomposition that attributes everything that is not due to the genetic term to error, since it is possible to include controllable environmental factors in the model as fixed effects. The animal model is the basis of most methods of genomic-based analysis, including genomic prediction and association studies. To facilitate the solution, the simplified mixed model equation (MME) of the animal model assumes that residuals are uncorrelated ($\mathbf{R} = \mathbf{I}$), reducing it to the $\mathbf{Cg} = \mathbf{r}$ problem, as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{Z'X} \\ \mathbf{X'Z} & \mathbf{Z'Z} + \lambda\mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \therefore \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \therefore \mathbf{Cg} = \mathbf{r}$$

where $\mathbf{C}$ is a square matrix comprising the cross-product of the design matrices and kernel matrix, $\mathbf{g}$ is a vector of regression coefficients of fixed and random effects, and $\mathbf{r}$ represents the cross-product of design matrices and response variable.

In this setup, the kernel matrix **K** will define what type of value the model yields for selection purposes. If **K** is an identity matrix then **u** is a vector of genetic values, for which it is particularly important to have replicated observations. If **K** is a pedigree or genomic relationship matrix, then **u** is a vector of breeding values where individuals that share the genetic basis defined in **K** work as partial replications. If **K** is a non-linear kernel (eg. Gaussian) then **u** is a vector of non-linear genomic values, because the Gaussian kernel may account for some level of epistasis. In order to avoid conflicting terminology, from this point the term "breeding value" denotes the random effect coefficients **u**.

If σ^2_e and σ^2_a were known beforehand, finding the coefficients **b** and **u** would not be a problem because the $\mathbf{Cg} = \mathbf{r}$ can be solved via least square regression. However it is necessary to estimate coefficients and variance components from the data simultaneously. The parameters estimated by Henderson's method are described as "Empirical Bayes" because they estimate the prior information necessary to solve the model (ie. σ^2_e and σ^2_a) based on the data (Zhou and Stephens 2014, Gianola *et al.* 1986).

Sorensen and Gianola (2002) showed the Bayesian nature of the mixed model by expressing $\mathbf{X'X}$ as an additional random effect ($\mathbf{X'X} + \lambda\mathbf{K}^{-1}$) that does not undergo regularization (ie. shrinkage) due to the prior knowledge of $\sigma^2_b = \infty$, which results in a null shrinkage ($\lambda = \sigma^2_e/\sigma^2_b = \sigma^2_e/\infty = 0$) with independent terms ($\lambda\mathbf{K}^{-1} = 0 \times \mathbf{K}^{-1} = 0$). Sorensen and Gianola (2002) make a clear distinction between the probabilistic natures of the frequentist and Bayesian mixed models: Under the frequentist framework, the probabilistic model is defined as $\mathbf{y} \sim \mathrm{N}(\mathbf{Xb}, \mathbf{ZKZ}\sigma_a^2 + \mathbf{I}\sigma_e^2)$, whereas under the Bayesian framework it becomes $\mathbf{y} \sim \mathrm{N}(\mathbf{Xb} + \mathbf{Zu}, \mathbf{I}\sigma_e^2)$.

Unless variance components are known a priori, the remaining question is: how can one find a λ parameter that provides a robust estimation of breeding values? The main strategy of supervised machine learning is the use of cross-validation to find the value of λ that provides the best prediction. Cross validation works by dividing the dataset into k subsets and testing the predictability for a wide range of values of λ. The predictability can be computed as the mean square prediction error (lower is better) or the correlation between the predicted and observed (higher is better). A three-fold cross validation to find λ would work as follows:

1. Divide the observed data into three groups (A, B, C);
2. Propose a value for λ;
3. Use AB to predict C; AC to predict B; and BC to predict A;
4. Compute the predictability for this given value of λ;
5. Repeat the previous two steps for a wide range of values of λ;
6. Use the value of λ that provides the highest predictability.

The λ parameter controls the complexity of the model and, consequently, the known tradeoff between bias and variance. Increases in λ add bias that reduces the variance, which often creates a more consistent prediction. As an alternative to cross-validation, it is still possible to estimate the λ value from the variance components ($\lambda = \sigma_e^2/\sigma_a^2$) to provide the best linear unbiased prediction (BLUP).

There are two popular methods for estimating variance components in kernel-based mixed models to obtain a robust value of λ as σ_e^2/σ_a^2 (Robinson 1991): restricted maximum likelihood (REML) (Patterson and Thompson 1971) and Bayesian Gibbs sampling (BGS) (Wang *et al.* 1993). We will also present an alternative involving re-parameterization using reproducing kernel Hilbert spaces (Gianola *et al.* 2006). After presenting kernel-based models, the next section will also present some methods that do not require explicit kernels to provide an equivalent BLUP solution.

REML Algorithm

REML is probably the most employed method for general-purpose estimation of variance components and regression coefficients. It is relatively unbiased when the number of observations is greater than the number of parameters ($n > p$) and much work has gone into making computationally feasible algorithms (Zhou and Stephens 2014, Kang *et al.* 2008, Lee and van der Werf 2016, Misztal *et al.* 2002).

There are a variety of algorithms to compute the REML variance components. This is a numerical optimization problem in which the main goal is to find the variance components and regression coefficients that optimize the restricted maximum likelihood of the data. The restricted (log) likelihood function of the data as a function of the variance components can be expressed as:

$$\mathrm{L}(\sigma_a^2, \sigma_e^2) = -\frac{1}{2}|\mathbf{V}| - \frac{1}{2}\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| - \frac{1}{2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

where $\mathbf{V} = \mathbf{ZKZ}'\sigma_a^2 + \mathbf{R}\sigma_e^2$ (Searle 1979). Popular algorithms include the derivation-free algorithm (Meyer 1989); first-derivative methods, such as expectation-maximization (EM) (Dempster *et al.* 1977); and second-derivative methods, such as Average Information (AI) (Gilmour *et al.* 1995). First- and second-derivative methods have an iterative-analytical solution, but can also be solved numerically via Monte Carlo (Matilainen *et al.* 2013).

The derivation-free approach implemented by Meyer (1989) finds the variance components that maximize the likelihood functions presented above through a heuristic method of optimization called the simplex method (Nelder and Mead 1965), which is similar to a 'guess and check' approach. The classical version is inefficient for complex models with large data, but Kang *et al.* (2008) reintroduced an alternative version that searches directly for the λ that minimizes the negative log likelihood, known as the efficient mixed model association (EMMA) algorithm.

Henderson (1984) presented the expectation maximization (EM-REML) algorithm based on the EM-ML algorithm of Dempster *et al.* (1977), using the first derivative of the restricted log-likelihood as simplified by Searle (1979). The principle of EM is to iteratively update residuals, variances, and coefficients as follows:

1. Propose some starting value for σ_e^2 and σ_a^2;
2. Solve the mixed model to find the coefficients: $\mathbf{g} = \mathbf{C}^{-1}\mathbf{r}$;
3. Compute the residuals: $\mathbf{e} = \mathbf{y} - \mathbf{Wg}$;
4. Update the residual variance: $\sigma_e^2 = \mathrm{n}^{-1}[\mathbf{e}'\mathbf{e} + \mathrm{tr}(\mathbf{WC}^{-1}\mathbf{W}')\sigma_e^2]$;
5. Update the random variance: $\sigma_a^2 = q^{-1}[\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{22})\sigma_e^2]$;
6. Repeat steps 2-5 until the variance components converge;

where $\mathbf{C}^{22}$ represents the $\mathbf{C}_{22}$ term from $\mathbf{C}^{-1}$, and $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$. EM is a very consistent algorithm, but it converges slowly and it requires the inversion of **C** every round to find the regression coefficients. Some numerical strategies can help with solving the MME, such as Cholesky decomposition and Gauss-Seidel algorithm (Legarra and Misztal 2008).

Newton-type methods work by using the gradient obtained by the second-derivatives to update both variance components at the same time. The gradient is generated by a Taylor series converging in the direction in which the parameters minimize the negative log-likelihood (Hofer 1998). Among these methods, the average-information (AI-REML) proposed by Gilmour *et al.* (1995) is the most common because it creates the gradient based on the average of the expected and observed information. The iterative algorithm AI-REML used to find variance components in the animal model is:

$$\begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix}^{t+1} = \begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix}^{t} + 0.5\begin{bmatrix}\mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{PZZ}'\mathbf{Py})\sigma_e^2 & \mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{Py})\sigma_e^4\\ \mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{P})\sigma_e^4 & \mathrm{tr}(\mathbf{y}'\mathbf{Py})\sigma_e^6\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{tr}(\mathbf{PZZ}') - \mathbf{y}'\mathbf{PZZ}'\mathbf{Py}\\ \mathrm{tr}(\mathbf{P}) - \mathbf{y}'\mathbf{PPy}\end{bmatrix}$$

where the parametrization matrix **P** is defined as $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{XV}^{-1}$. The AI-REML is computationally demanding, but it converges within a few iterations to a consistent result. This algorithm has been widely implemented for breeding applications (Gilmour *et al.* 2009, Meyer 2007, Misztal *et al.* 2002). The most time-consuming part of this method is updating the **P** matrix because it requires inversion of the covariance matrix.

It is possible to substantially reduce this computational burden through the Eigendecomposition of $\mathbf{ZKZ}$ to speed up the inversion of $\mathbf{V}$ (Kang *et al.* 2008, Lippert *et al.* 2011). Any square matrix can be Eigendecomposed into eigenvectors ($\mathbf{U}$) and eigenvalues ($\mathbf{D}$), thus $\mathbf{ZKZ} = \mathbf{UDU}'$. Thence, one can obtain $\mathbf{V}^{-1} = \mathbf{U}[\mathbf{D} \times (\sigma_a^2/\sigma_e^2) + 1]^{-1}\mathbf{U}'\sigma_e^{-2}$ and the only inversion required is the elements of a diagonal matrix.

BGS Algorithm

Bayesian Gibbs sampling (BGS) is an algorithm proposed by Gelman and Gelman (1984) that works similarly to the EM algorithm, updating one parameter at a time. Parameters are stored in each cycle, and averaged out at the end. BGS algorithms commonly discard cycles prior to the stationary state (ie. entropy) to provide stability to final estimates (so-called "burn in"). The distribution of the parameters from several cycles are called posterior distribution, often notated as $p(\theta|X)$. In this context, the parameters we are looking for are $\theta = \{\mathbf{b}, \mathbf{u}, \sigma_a^2, \sigma_a^2\}$ (given the data we have), which refers to our matrices, thus $X = \{\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{K}\}$.

The advantage of Bayesian methods is that they initially incorporate some of your expectations about the data, such as the probability distribution of parameters. Wang *et al.* (1993) proposed the first Gibbs sampler algorithm to solve mixed models in the breeding context, where coefficients follow a normal distribution (as they do in REML) and variance components follow a scaled inverse chi-squared distribution ($\chi_{v,S}^{-2}$). This ensures positive estimates of variance components. The algorithm is as follows:

1. Propose some starting value for $\mathbf{g}$, σ_e^2 and σ_a^2;
2. Update each coefficient as $g_i \sim N\left(\text{mean} = g_i^*, \text{variance} = \sigma_e^2 \mathbf{C_{ii}^{-1}}\right)$;
3. Compute the residuals: $\mathbf{e} = \mathbf{y} - \mathbf{Wg}$;
4. Update the residual variance as $\sigma_e^2 = (\mathbf{e}'\mathbf{e} + S^* v^*)/\chi_{n+v^*}^2$;
5. Update the random variance as $\sigma_a^2 = (\mathbf{a}'\mathbf{A^{-1}}\mathbf{a} + S^* v^*)/\chi_{q+v^*}^2$;
6. Update $\mathbf{C}$ with the new value of $\lambda = \sigma_e^2/\sigma_a^2$;
7. Repeat steps 2–6 for a reasonable number of iterations (eg. 1500);
8. Discard the cycles prior to entropy (eg. 500);
9. For each parameter $(\mathbf{g}, \sigma_e^2, \sigma_a^2)$, average the values from iteration (eg. 501-1500);

where $g_i^* = (r_i - \mathbf{C_{i,-i}} g_{-i})\mathbf{C_{ii}^{-1}}$, $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$, and the priors' shape and the degrees of freedom of the variance components are represented by S^* and v^* respectively, and where $S^* = 0.5 \times \text{var}(y)$ and $v^* = 5$ are reasonable priors (Morota *et al.* 2014). Some priors indicate a total unawareness about the expected response in accordance with Laplace's principle of uniform ignorance. These are called flat priors and they should provide results equivalent to REML. To employ a flat prior, set $S^* = 0$ and $v^* = -2$ (Garcia-Cortes and Sorensen 1996).

Parameterization using reproducing kernels Hilbert spaces (RKHS) is an alternative way to solve mixed models that use kernels. This process follows an algorithm proposed by de los Campos *et al.* (2010) that uses the Eigendecomposition of the kernel (ie. $\mathbf{K} = \mathbf{UDU}'$) to obtain the matrix Eigenvectors ($\mathbf{U}$) and the diagonal matrix of Eigenvalues ($\mathbf{D}$). Hence the random term $\mathbf{Zu}$ with $\mathbf{u} \sim \mathrm{N}(0, \mathbf{K}\sigma_a^2)$ can be reparametrized as $\mathbf{Z}^*\boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim \mathrm{N}(0, \mathbf{D}\sigma_a^2)$, where $\mathbf{Z}^* = \mathbf{ZU}$.

When multiple kernels are involved in the same model (additive, dominance, and epistatic kernels), RKHS is often preferable to the traditional methods. RKHS is compatible with the BGS and REML frameworks, and it also allows the solution of mixed models as a ridge regression of Eigenvectors, with a special regularization ($\lambda^* = \mathbf{D}^{-1}\, \sigma_e^2/\sigma_a^2$).

Whole-genome regression (WGR) algorithms

It is also possible to obtain BLUP estimates of breeding values and variance components without kinship matrices. This is especially useful when genotypic information is available (de los Campos *et al.* 2013; VanRaden 2008) allowing a more reliable inference of breeding values (Bernardo and Nyquist 1998). These are called whole-genome regression (WGR) methods. Methods used for WGR are flexible so that they can accommodate hyper-dimensional problems, as when models have more parameters than observations ($p \gg n$), without having to compute large matrices (eg. $\mathbf{M'M}$).

Given the linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{e}$, WGR computes the additive value (α_i) of each marker (m_i) and obtains breeding values by taking the sum of all marker values. Thus, breeding values are estimated as $\mathbf{u} = \mathbf{M}\boldsymbol{\alpha}$. Loci are often coded as {-1, 0, 1} or {0, 1, 2} representing {AA, Aa, aa} but can also be centralized (Zheng et al. 2005, VanRaden 2008), and the vector of regression coefficients $\boldsymbol{\alpha}$ represents the additive value of each allele substitution (Xu 2013).

The simplest WGR model is called ridge regression (RR) or Tikhonov regularization. This is a Gaussian process comprising p stochastic processes, where p is the number of markers in the model ($p = m$), which provides a result equivalent to the previous methods using the additive genomic relationship as kernel matrix (VanRaden 2008, Morota *et al.* 2014). Ridge regression assumes that regression coefficients are normally distributed and provides an interesting framework for working with multicollinearity (Hoerl and Kennard 1970). This is a highly desirable property for handling genomic data and when multiple markers located in a same region carry similar information.

Most WGR methods attempt to minimize a loss function represented by $\mathrm{argmin}(\mathbf{e'e} + \lambda\boldsymbol{\alpha}'\boldsymbol{\alpha})$. Notice that this loss function has two terms: the residual sum of squares ($\mathbf{e'e}$) and the complexity term $\lambda\boldsymbol{\alpha}'\boldsymbol{\alpha}$. The squared penalization of coefficients ($\lambda\boldsymbol{\alpha}'\boldsymbol{\alpha}$) is called L_2 penalization, while L_1 penalization denotes the use of the absolute sum ($\lambda||\boldsymbol{\alpha}||$). L_1 penalization is also known as least absolute shrinkage and selector operator (LASSO) loss (Tibshirani 1996).

Coordinate descent helps to minimize the ridge and LASSO loss functions presented above (Hastie *et al.* 2005), which means that regression coefficients are updated one at a time. Let us begin with the simplest univariate solution: the ordinary least squared (OLS). For a given univariate model $\mathbf{y} = \mathbf{x}\mathrm{b} + \mathbf{e}$, the OLS solution for the regression coefficient is:

$$\mathrm{b} = \frac{\mathbf{x'y}}{\mathbf{x'x}}$$

whereas the ridge regression solution for the same problem is given by:

$$\mathrm{b}_{\mathrm{ridge}} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x} + \lambda}$$

in which the regularization parameter λ imposes shrinkage.

The LASSO univariate solution works slightly differently. For a positive OLS coefficient, the LASSO solution is:

$$\mathrm{b}_{\mathrm{lasso}} = \frac{\mathbf{x}'\mathbf{y} - \lambda}{\mathbf{x}'\mathbf{x}}$$

If the LASSO solution is negative, then the regression coefficient is set as zero. Similarly, when the OLS is negative and the LASSO numerator is given by $\mathbf{x}'\mathbf{y} + \lambda$, the coefficient is set at zero when the LASSO solution is positive. Thus, LASSO performs variable selection in addition to shrinkage, whereas ridge is incapable of providing null regression coefficients. The intermediate model between ridge regression and LASSO is called elastic net (Zou and Hastie 2005), in which regularization minimizes both L_1 and L_2 penalizations, and the univariate solution is:

$$\mathrm{b}_{\mathrm{en}} = \frac{\mathbf{x}'\mathbf{y} - \lambda_1}{\mathbf{x}'\mathbf{x} + \lambda_2}$$

After determining the univariate solution, coordinate descent algorithms follow. For starters, assume we are solving a model where the only fixed effect is the intercept (μ) and the omic data from p parameters is represented by the matrix $\mathbf{X}$, following the model $\mathbf{y} = \mu + \mathbf{Xb} + \mathbf{e}$. The algorithm is simple: reduce the linear model to a univariate version ($\tilde{\mathbf{y}}_{\mathrm{i}} = \mathbf{x}_{\mathbf{i}}\mathrm{b}_{\mathrm{i}} + \mathbf{e}$) and solve one coefficient at a time until convergence. To do so, it is necessary to fit all but the one variable that is being updated. Thus the ridge solution for the i^{th} parameter becomes:

$$\mathrm{b}_{\mathrm{i}} = \frac{\mathbf{x}_{\mathbf{i}}'(\mathbf{y} - \mathbf{X}_{-\mathbf{i}}\mathbf{b}_{-\mathbf{i}})}{\mathbf{x}_{\mathbf{i}}'\mathbf{x}_{\mathbf{i}} + \lambda} = \frac{\mathbf{x}_{\mathbf{i}}'\tilde{\mathbf{y}}_{\mathrm{i}}}{\mathbf{x}_{\mathbf{i}}'\mathbf{x}_{\mathbf{i}} + \lambda}$$

where $\tilde{\mathbf{y}}_{\mathrm{i}}$ represents $\mathbf{y}$ accounting for all parameters except one ($\mathbf{x}_{\mathrm{i}}$). Legarra and Misztal (2008) have provided an efficient framework to prevent the recalculation of $\mathbf{X}_{-\mathrm{i}}\mathbf{b}_{-\mathrm{i}}$ for every regression coefficient, the two-step Gauss-Seidel residual update (GSRU) algorithm, in which the vector of residuals helps in replacement to $\tilde{\mathbf{y}}_{\mathrm{i}}$. The first step involves updating the i^{th} regression coefficient ($\mathrm{b}_{\mathrm{i}}^{\mathrm{t}+1}$) using the current version of residuals:

$$\mathrm{b}_{\mathrm{i}}^{\mathrm{t}+1} = \frac{\mathbf{x}_{\mathbf{i}}'\mathbf{e}^{\mathrm{t}} + \mathbf{x}_{\mathbf{i}}'\mathbf{x}_{\mathbf{i}}\mathrm{b}_{\mathrm{i}}^{\mathrm{t}}}{\mathbf{x}_{\mathbf{i}}'\mathbf{x}_{\mathbf{i}} + \lambda}$$

This is followed by a subsequent update of the residuals:

$$\mathbf{e}^{\mathrm{t}+1} = \mathbf{e}^{\mathrm{t}} - \mathbf{x}_{\mathbf{i}}'(\mathrm{b}_{\mathrm{i}}^{\mathrm{t}+1} - \mathrm{b}_{\mathrm{i}}^{\mathrm{t}})$$

The regularization parameter is commonly estimated by cross-validation in the traditional machine learning framework (Hastie *et al.* 2005), whereas the intercept is the only coefficient updated without regularization ($\lambda = 0$) due to its fixed-effect nature. If the variance components are estimated in each round, one can also update the regularization parameter as $\lambda = \sigma_e^2/\sigma_a^2$.

The Bayesian counterpart of ridge regression (BRR) is solved via a Gibbs sampler, providing a nearly identical solution (de los Campos *et al.* 2013). The main difference is the parameter updating based upon sampling. The BRR algorithm proceeds as follows:

The regression coefficients are sampled from a normal distribution using the GSRU solution as the expected value with a subsequent residual update:

$$\mathrm{b}_\mathrm{i}^{\mathrm{t}+1} \sim \mathrm{N}\left(\mathrm{mean} = \frac{\mathbf{x}_\mathbf{i}'\mathbf{e}^\mathrm{t} + \mathbf{x}_\mathbf{i}'\mathbf{x}_\mathbf{i}\mathrm{b}_\mathrm{i}^\mathrm{t}}{\mathbf{x}_\mathbf{i}'\mathbf{x}_\mathbf{i} + \lambda}, \mathrm{var} = \frac{\sigma_e^2}{\mathbf{x}_\mathbf{i}'\mathbf{x}_\mathbf{i} + \lambda}\right)$$

$$\mathbf{e}^{\mathrm{t}+1} = \mathbf{e}^\mathrm{t} - \mathbf{x}_\mathbf{i}'(\mathrm{b}_\mathrm{i}^{\mathrm{t}+1} - \mathrm{b}_\mathrm{i}^\mathrm{t})$$

Then the variance components are updated from a scaled inverse chi-squared distribution:

$$\sigma_\mathrm{a}^2 = \frac{\mathrm{b}'\mathrm{b} + \mathrm{S}^*\nu^*}{\chi_{\mathrm{p}+\nu^*}^2} \text{ and } \sigma_\mathrm{e}^2 = \frac{\mathrm{e}'\mathrm{e} + \mathrm{S}^*\nu^*}{\chi_{\mathrm{n}+\nu^*}^2}$$

Two models derive from BRR by modifying the regularization setup into non-Gaussian processes: BayesA and Bayesian LASSO. BayesA (Meuwissen *et al.* 2001) is a special case of BRR where each marker has its own variance ($\sigma_{\mathrm{b_i}}^2 = (\mathrm{b}_\mathrm{i}^2 + \mathrm{S}^*\nu^*)/\chi_{1+\nu^*}^2$), implying that each marker will have a unique regularization parameter ($\lambda_\mathrm{i} = \sigma_\mathrm{e}^2/\sigma_{\mathrm{b_i}}^2$). Marker effects follow a t-distribution (thick tails). Breeding values from BayesA are more accurate than BRR, but often biased and sensitive to the prior specification (Lehermeier *et al.* 2013, Gianola 2013). A common set of prior used for BRR and BayesA can intuitively defined as: $S_e^* = 0.5\ \sigma_\mathrm{y}^2$ and $S_b^* = 0.5\ \sigma_\mathrm{y}^2/\sum_j \sigma_{x_j}^2$ with $\nu^* = 5$ for both marker and residual variances.

The Bayesian LASSO (BL) proposed by Park and Casella (2008) has a very particular parametrization that imposes strong shrinkage but, unlike its non-Bayesian counterpart, it is not capable of performing variable selection. In BL, the regularization parameter for each individual parameter (λ_i) is sampled from an inverse-Gaussian distribution with expectation $\sigma_\mathrm{e}\phi/\mathrm{b}_\mathrm{i}$ and shape ϕ^2, such that the distribution of marker effects follows a Laplace distribution.

Non-Gaussian processes (eg. BayesA, BL, LASSO) are able to capture large-effect QTL better than ridge regression and BRR (Fig. 6), whereas kernel methods do not even assign values to each maker. Zhang *et al.* (2010a) proposed a two-step method to incorporate large-effect QTL into kernel methods, thus generating weighted kernels (also known as trait-associated kernels). The first step consists of fitting a WGR to obtain regression coefficients for each marker. The second step involves recoding alleles coded as $\{-|\mathbf{b}|, 0, |\mathbf{b}|\}$ before designing the kernel, so that each allele is weighted according to its association with the trait.

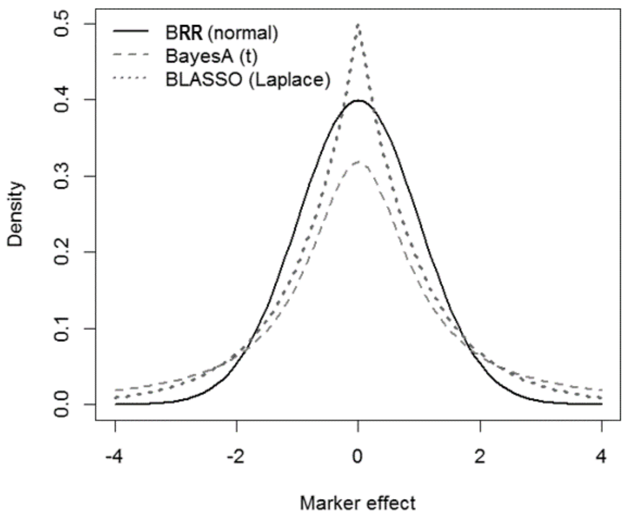


Fig. 6 Density function of whole-genome regressions BRR, BayesA, and Bayesian LASSO, where marker effects follow normal, t, and Laplace distributions, respectively.

The prediction accuracy provided by the various methods (kernel and regression) changes according to the genetic architecture of the trait (de los Campos *et al.* 2013) and the model with more realistic assumptions often provides the most accurate prediction. Although all models are likely to provide robust predictions, looking for the optimal method may require breeders to evaluate multiple models through cross-validation.

One may believe that not all markers contribute to the trait of interest but shrinkage does not eliminate markers from the model. There are two ways to tackle this problem: either using L_1 loss or adding a variable selection term into the L_2 model. Indeed, each Bayesian model presented earlier has a variable selection counterpart: BayesA has BayesB (Meuwissen *et al.* 2001), BRR has BayesCπ (Habier *et al.* 2011), and BL has an expanded version proposed by Legarra *et al.* (2011b).

Meuwissen *et al.* (2001) proposed variable selection using the Metropolis-Hasting algorithm, which suggests that markers be included into the model at random. The proposed changes are accepted when the new model provides a better likelihood. Meuwissen's approach is robust, but at a high computational cost. Alternatively, efficient variable selection can be incorporated in the Gibbs sampler (O'Hara and Sillanpää 2009) via the following three methods:

1. Stochastic search variable selection (George and McCulloch 1993);
2. Unconditional prior (Kuo and Mallick 1998);
3. Gibbs variable selection (Dellaportas *et al.* 2002).

Table 2 summarizes this section, showing the computation of breeding values to aid selection through kernel and WGR methods. The procedures of screening the whole-genome for large effect QTL by testing one marker at a time conditional to a kernel-based polygenic term are called genome-wide association studies (GWAS).

Table 2 Comparison of methods used to generate breeding values.

Method	Class	Process	Variable Selection	Large-effect QTL	Loss function
Pedigree BLUP	Kernel	Gaussian			REML/L_2
Linear GBLUP	Kernel	Gaussian			REML/L_2
Spatial GBLUP	Kernel	Gaussian			REML/L_2
Weighted GBLUP	Kernel	Gaussian	X	X	REML/L_2
Ridge	Regression	Gaussian			REML/L_2
LASSO	Regression	Laplace	X	X	L_1
Elastic Net	Regression	Mixture	X	X	L_1/L_2
BRR	Regression	Gaussian			L_2
BayesA	Regression	t		X	L_2
BL	Regression	Laplace		X	L_2
BayesB	Regression	Mixture	X	X	L_2
BayesCπ	Regression	Gaussian	X		L_2

Because non-Gaussian WGR methods are capable of capturing major effect alleles, these methods can be used directly to perform GWAS. LASSO and BayesCπ have been widely used to detect QTLs (Colombiani *et al.* 2012, Fang *et al.* 2012, Li and Sillanpää 2012, Yi and Xu 2008). Furthermore, a comparison study performed by Legarra *et al.* (2015) pointed out the superiority of these methods over the traditional framework, which is based on comparing the likelihood of null and full models.

Data Quality Control and Association Analysis

Understanding the underlying genetics of quantitative traits informs strategies for crop improvement (Sonah *et al.* 2014). The most basic procedure to associate genetics and phenotypes with molecular tools is to find the markers associated with phenotypes and consequently determine which genes are involved. Regardless of the genetic resources (ie. type of population), association studies have four fundamental steps: phenotyping, genotyping, mapping, and validation. Validation consists of performing the first three procedures of phenotyping, genotyping, and mapping on an experimental population specially designed for this purpose (eg. near isogenic lines). Therefore, we will emphasize only the three initial steps.

Phenotyping

When traits are governed by many loci, sensitivity to environmental variation increases. The external stimuli affect the genetic expression of the various loci at different levels. The genetic expression of complex traits, like yield and drought tolerance, is highly variable across the genome (Guimarães-Dias *et al.* 2012, Le *et al.* 2011). In the context of minimizing environmental noise in phenotypes, research on field phenomics aims to generate or improve high-throughput and high-precision phenotyping techniques, but the integration of various sources of omic data has been primarily used to improve abiotic stress (Deshmukh *et al.* 2014).

The use of replications is always highly desirable, because having multiple observations always improves the accuracy of estimates of the true genetic value. It is possible to further reduce noise due to field variation by using spatial statistics, such as kriging (Basso *et al.* 2000), which allows adjustment for spatial correlation among field trials (Banerjee *et al.* 2010, Zas 2006). For example, Lado *et al.* (2013) were able to improve accuracy of genomic prediction in wheat by controlling field variation through spatial adjustments using a simple mixed model with a moving-mean covariate structure.

Kriging methods to control for field variation can complement experimental design and unreplicated trials (Banerjee *et al.* 2010, Lado *et al.* 2013). Phenotypic data contains genetic information, micro- and macro-environmental variation, and interactions among environmental and genetic factors. For this application of kriging, we can employ a mixed effect model with an additional term to define field correlation among field plots. Thus:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Iv} + \mathbf{e}$$

where the observed phenotype ($\mathbf{y}$) is a function of some fixed effect ($\mathbf{Xb}$) such as block or environment. The genetic effect ($\mathbf{Zu}$) allows specification of the association among individuals given $\mathbf{u} \sim \mathrm{N}(0, \mathbf{K}\sigma_a^2)$, where $\mathbf{K}$ represents the additive genetic relationship matrix. The field variation ($\mathbf{Iv}$) term represents a spatial relationship (ie. Euclidean distance between plots in the field) defined by a spatial kernel (eg. Gaussian) such that $\mathbf{v} \sim \mathrm{N}(0, \mathbf{S}\sigma_s^2)$. While the residual term ($\mathbf{e}$) incorporates random errors and higher-order interactions. There is also an alternative approach that assumes the residuals are correlated, so that $\mathbf{e} \sim \mathbf{N}(0, \mathbf{S}\sigma_e^2)$, thus avoiding the additional term ($\mathbf{Iv}$) in the model.

Accounting for spatial variation is particularly important in unreplicated trials (eg. progeny rows) when pedigree and genotypic information is scarce. Thus the distinction between genetics and environment is a complex problem and the use of a replicated check may be the only true indicator of field variation. With reduced environmental noise, genotypic values tend to have a more stable performance across environments, which can be measured using a Pearson or Spearman correlation. Another measure of improvement provided by accounting for field variation is the increase in broad- and narrow-sense heritabilities, for which increased variance is expected to be due to genetic factors.

Genotyping

High-throughput genotyping techniques have become very popular in plant breeding (Jarquín *et al.* 2014, Sohan *et al.* 2014), often with poor genotyping quality and a large amount of missing data (Halprin and Stephan 2009) that makes mapping and selection challenging (Jarquín *et al.* 2014, Poland and Rife 2012). In such scenarios, the accurate imputation of missing loci and good correction of SNP miscalls becomes essential for robust downstream analyses (Marchini and Howie 2010, Xavier *et al.* 2016). Two popular methods of genotypic imputation in plant breeding are hidden Markov models (HMM) and random forest (Swarts *et al.* 2014, Rutkoski *et al.* 2013).

HMM are commonly employed in genetics and genomics for stochastic modeling of Markov processes, such as the computation of haplotypes. In genetic terms, the three possible states of a diploid organism with two alleles for a given locus m are: $\mathrm{M_1M_1}$, $\mathrm{M_1M_2}$, and $\mathrm{M_2M_2}$, disregarding linkage phase. Assuming ordered markers, the HMM estimates the most likely path of states (ie. genotype) based on the transition probability of marker $\mathrm{m^t}$ to change state given the previous marker $\mathrm{m^{t-1}}$. HMM is the most common method for imputation of missing genotypes. In addition, Marchini and Howie (2010) showed that HMM can boost the power and resolution of genome-wide association studies.

Random forest is a non-parametric method for predicting, classifying, and imputing mixed data types. It establishes a combination of decision-tree predictors, in which decision trees are bootstrapped to generate random independent vectors that constitute training forests. This is particularly useful for imputing unordered markers. Rutkoski *et al.* (2013) reported random forest as a promising imputation method for genotyping-by-sequencing (GBS) data in wheat, and Xavier *et al.* (2016) showed that random forest is as efficient as HMM in soybeans.

Other quality parameters that have a major impact on analysis are the minor allele frequency (MAF) of molecular markers (Tabangin *et al.* 2009) and the marker's ability to carry a gene. This latter is estimated from the marker heritability (Forneris *et al.* 2015) when markers are seen as molecular phenotypes. It is used to identify markers that do not follow Mendelian segregation due to biased inheritance of alleles (Glémin 2010).

Minor alleles are very important for population stratification. Wen *et al.* (2008) found as many as nine subpopulations when evaluating the structure of 393 landraces and 196 native populations of soybeans in China. However, low MAF has two major drawbacks in association analysis: (1) it may increase the rate of false discoveries if one disregards the existence of a subpopulation; and (2) if an allele has a major effect but is only present in a low frequency, this particular gene will become undetectable due to the lack of power associated with its low signal-to-noise ratio (Tabangin *et al.* 2009). Jarquín *et al.* (2014) found that an MAF threshold as high as 0.30 improved prediction accuracy of genomic selection models in soybeans.

Gene Mapping

The improvement seen in gene mapping is one of the rare occasions in which machine learning is concerned with enhancing inference accuracy. The principles of mapping were discussed previously, where we showed that associations between marker and trait can be estimated by the improvement in (restricted) likelihood provided by the marker, conditional to a polygenic term (ie. additive kernel) that accounts for the existence of subpopulations.

Early mapping studies from random populations ignored population structure, which may have led to a great number of false discoveries (Xu and Shete 2005). Yu *et al.* (2006) proposed a mixed model framework that accounts for background genetics called the unified mixed model (UMM), also known as the $\mathrm{K+Q}$ method. In this method, a fixed-effect population structure term ($\mathbf{Q}$) is complementary to the polygenic term derived from a kernel ($\mathbf{K}$) of pedigree, genomic data or both. The population structure is often defined by clusters from the software STRUCURE (Pritchard *et al.* 2000) or Eigenvectors computed using the software EIGENSTRAT (Price *et al.* 2006). UMM has some undesirable properties, including redundancy (once the information of **Q** is extracted from **K**) and the computational burden from the estimation of variance components for every marker.

In order to avoid computing the mixed model every round, Aulchenko *et al.* (2007) proposed an approximated method known as the genome-wide rapid association using mixed model and regression (GRAMMAR) algorithm. The authors proposed fitting the animal model and analyzing the residual term as un-structured phenotypes without needing to include a polygenic term, so that the mixed model only needs to be solved once. Although conveniently fast, the original GRAMMAR approach provides biased estimates of SNP effects. Some have proposed variations from the original algorithm to overcome this limitation, including the GRAMMAR-gamma (Svishcheva *et al.* 2012) and BOLT-LMM (Loh *et al.* 2015). Due to its computational feasibility, GRAMMAR is often the model of choice for analyzing a large number of markers.

In the previous section we mentioned the use of Eigendecomposition for the efficient computation of mixed models. Kang *et al.* (2008) proposed the EMMA algorithm to provide a computational solution for the K + Q model by using numerical optimization methods to search for a λ that maximizes REML. And in this algorithm, Eigendecomposition plays a major role in simplifying the computation of the likelihood function. The EMMA algorithm became a popular solution for single-kernel mixed models, implemented in popular R packages such as rrBLUP, EMMREML, and NAM (Endelman 2011, Akdemir and Jannink 2015, Xavier *et al.* 2015). However, EMMA is impractical for association analysis in large datasets.

To overcome the computational limitations seen in EMMA, some have proposed approximation methods known as compressed mixed models. These include the EMMA eXpedited (EMMAX) algorithm (Kang *et al.* 2010) and the population parameters previously determined (P3D) algorithm (Zhang *et al.* 2010b). EMMAX and P3D generate the polygenic term for clusters of individuals in order to compress the information of **K**. These methods also assume the variance components in the full modes to be equivalent to the null model and thus estimate variance components only once. The compression of the polygenic term entails substantial information loss, but the **Q** term helps to preserve part of this information.

Others have presented more efficient solutions for the mixed model without compression, also known as exact methods. Lippert *et al.* (2011) proposed the factored spectrally transformed (FaST) algorithm, while Zhou and Stephens (2012) introduced the genome-wide efficient mixed model association (GEMMA) algorithm. GEMMA utilizes the full-rank kernel, genomic relationship matrix (ie. uses all Eigenvectors) and this provides stability to the algorithm and very robust control over admixture. FaST, on the other hand, was designed to accommodate larger numbers of individuals with reduced-rank kernel, which also prevents the double-fitting of markers discussed below.

In general, mixed models can increase power and prevent false positives at a reasonable cost, but this approach also presents some pitfalls (Yang *et al.* 2014), such as the loss of power in case-control studies and (often) double-fitting markers into the model, where the marker under evaluation in the full model is also used to build the kernel (genomic relationship matrix). The use of WGR as a GWAS method could easily satisfy the limitation of double-fitting once each marker effect is inferred, conditional to all other parameters. Table 3 summarizes the properties of the main association algorithms.

Table 3 Comparison of association methods based on mixed models.

Algorithm	Class	Double-fitting	Computation	Full-rank GRM
EMMA	Exact	X	Slow	X
GRAMMAR-gamma	Approximate	X	Fast	X
BOLT-LMM	Approximate		Fast	
EMMAX / P3D	Approximate	X	Intermediate	
GEMMA	Exact	X	Intermediate	X
FaST-LMM	Exact		Intermediate	
WGR	Exact		Fast	X
Emp. Bayes	Exact		Intermediate	X

Recently, some have proposed more flexible models to relax assumptions made by the GWAS algorithm and to deal with complex structured populations, including next-generation panels, such as the multi-parent advanced generation inter-cross (MAGIC) and nested association mapping (NAM) populations. The empirical Bayes algorithm (Xavier *et al.* 2015, Wei and Xy 2015) endeavors to further increase power and resolution of GWAS by treating markers as a random effect to shrink the background noise to zero, also implementing a sliding window to overcome double-fitting markers by removing the local markers from the polygenic term. In addition, if any stratification factor is known *a priori*, the algorithm reparameterizes the markers to haplotypes, thus accounting for some level of epistasis and thereby relaxing assumptions about the linkage phase between marker and QTL in different subpopulations. Figure 7 compares the GWAS algorithms.

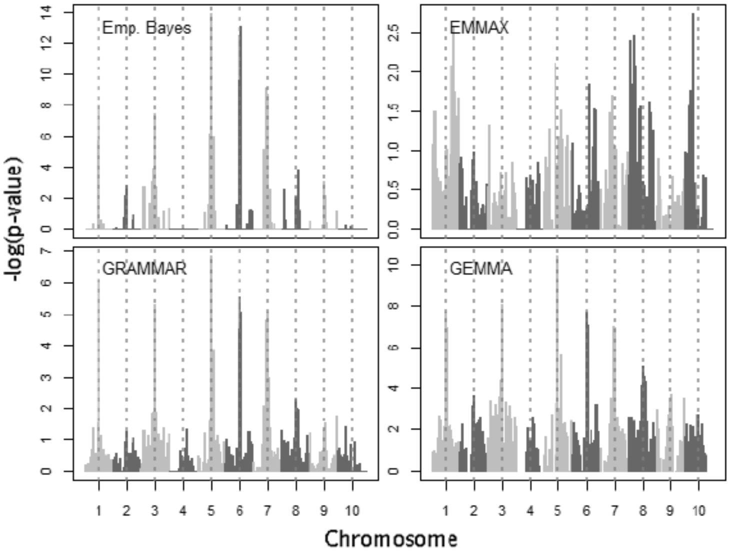


Fig. 7 Manhattan plots using four different algorithms of association analysis for a simulated nested association panel dataset with one QTL in the center of each chromosome.

Conclusions

The various models and algorithms all make important assumptions. Knowing how the computations work may help improve statistical analysis and decision making. Most statistical procedures in breeding theory are based on Gaussian process and can be computed through mixed models using kernels and regression models. We have illustrated some of the flexibility possible when using principles of machine learning and mixed models for selection, prediction, and mapping, as well as when inferring variance components.

References

Acquaah G (2009) Principles of plant genetics and breeding. John Wiley and Sons. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK.

Akdemir D, Jannink JL (2015) Locally epistatic genomic relationship matrices for genomic association and prediction. Genetics 199(3):857-71.

Aulchenko YS, De Koning DJ, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177(1): 577-585.

Banerjee S, Finley AO, Waldmann P, Ericsson T (2010) Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association 105(490): 506-521.

Basso B, Ritchie JT, Pierce FJ, Braga RP, Jones JW (2001) Spatial validation of crop models for precision agriculture. Agricultural Systems 68(2): 97-112.

Beavis WD (1998) QTL analyses: power, precision, and accuracy. Molecular dissection of complex traits, 145-162.

Bernardo R, Nyquist WE (1998) Additive and testcross genetic variances in crosses among recombinant inbreds. Theoretical and applied genetics 97(1-2): 116-121.

Carvalho AD, Fritsche Neto R, Geraldi IO (2008) Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares. Crop Breeding and Applied Biotechnology 8(3): 219-224.

Cleveland DA, Soleri D (Eds.) (2002) Farmers, scientists, and plant breeding: integrating knowledge and practice. CABI.

Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V, Robert-Granié C (2013) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of dairy science 96(1): 575-591.

Crow JF, Kimura M (1970) An introduction to population genetics theory. An introduction to population genetics theory.

Dardanelli JL, Balzarini M, Martínez MJ, Cuniberti M, Resnik S, Ramunda SF, *et al.* (2006) Soybean maturity groups, environments, and their interaction define mega-environments for seed composition in Argentina. Crop science 46(5): 1939-1947.

Dellaportas P, Forster JJ, Ntzoufras I. (2002) On Bayesian model and variable selection using MCMC. Statistics and Computing 12(1): 27-36.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327-345.

de Los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92(04): 295-308.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance components models. Journal of the American Statistical Association 76(374): 341-353.

Deshmukh RK, Sonah H, Patil G, Chen W, Prince S, Mutava R, *et al.* (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Plant Genetics and Genomics, 5, 244.

Diffey S, Welsh A, Smith A, Cullis BR (2013) A faster and computationally more efficient REML (PX) EM algorithm for linear mixed models. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 2-13, 8.

Egli DB (2008a) Soybean yield trends from 1972 to 2003 in mid-western USA. Field crops research 106(1): 53-59.

Egli DB (2008b) Comparison of corn and soybean yields in the United States: Historical trends and future prospects. Agronomy journal, 100(Supplement_3), S-79.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4(3):250-5.

Fang M, Jiang D, Li D, Yang R, Fu W, Pu L, *et al.* (2012) Improved LASSO priors for shrinkage quantitative trait loci mapping. Theoretical and Applied Genetics 124(7): 1315-1324.

Farrall M (2004) Quantitative genetic variation: a post-modern view. Human molecular genetics, 13(suppl 1), R1-R7.

Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh, 52: 399-433.

Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, Cantet RJ (2015) Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. Genetics 199(3): 675-681.

García-Cortés LA, Sorensen D (1996) On a multivariate implementation of the Gibbs sampler. Genetics Selection Evolution 28(1): 121-126.

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6): 721-741.

Gengler N, Mayeres P, Szydlowski M (2007) A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1(1): 21-28.

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88(423): 881-889.

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194(3): 573-596.

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761-1776.

Gianola D, Foulley JL, Fernando RL (1986) Prediction of breeding values when variances are not known. Genetics Selection Evolution 18(4): 485-498.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics, 1440-1450.

Glémin S (2010) Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics 185(3): 939-959.

Guimarães-Dias F, Neves-Borges AC, Viana AAB, Mesquita RO, Romano E, Grossi-de-Sa MDF, *et al.* (2012) Expression analysis in response to drought stress in soybean: Shedding light on the regulation of metabolic pathway genes. Genetics and molecular biology 35(1): 222-232.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics 12(1): 186.

Halperin E, Stephan DA (2009) SNP imputation in association studies. Nature biotechnology 27(4): 349-351.

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2):83-85.

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics, 423-447.

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 12(1):55-67.

Hofer A (1998) Variance component estimation in animal breeding: a review. Journal of Animal Breeding and Genetics 115(1|6): 247-265.

Imhof LA, Nowak MA (2006) Evolutionary game dynamics in a Wright-Fisher process. Journal of mathematical biology 52(5): 667-681.

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC genomics 15(1): 740.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42(4): 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178(3): 1709-1723.

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49(4): 725.

Kuo L, Mallick B (1998) Variable selection for regression models. Sankhya: The Indian Journal of Statistics, Series B, 65-81.

Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, Belzile F, *et al.* (2013) Increased Genomic Prediction Accuracy in Wheat Breeding Through Spatial Adjustment of Field Trial Data. G3: Genes| Genomes| Genetics 3(12): 2105-2114.

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121(1): 185-199.

Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K, Tran LSP (2011) Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. DNA research, dsr015.

Lee SH, van der Werf JH (2016) MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 10:btw012.

Legarra A, Croiseau P, Sanchez MP, Teyssèdre S, Sallé G, Allais S, *et al.* (2015) A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. Genetics Selection Evolution 47(1) 6.

Legarra A, Ricardi A, Filangi O (2011) GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel. snp.toulouse.inra.fr/~alegarra/.

Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S (2011) Improved Lasso for genomic selection. Genetics research 93(01): 77-87.

Legarra A, Misztal I (2008) Technical note: Computing strategies in genome-wide selection. Journal of dairy science 91(1): 360-366.

Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Statistical applications in genetics and molecular biology 12(3): 375-391.

Li Z, Sillanpää MJ (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theoretical and Applied Genetics 125(3): 419-435.

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nature Reviews Genetics. 1:16(6):321-32.

Lim C (1997) An econometric classification and review of international tourism demand models. Tourism economics 3(1): 69-81.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. Nature Methods 8(10): 833-835.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature genetics 47(3):284-90.

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits (Vol. 1). Sunderland: Sinauer.

MacLeod IM, Hayes BJ, Goddard ME (2014) The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198(4):1671-1684.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11(7): 499-511.

Matilainen K, Mäntysaari EA, Lidauer MH, Strandén I, Thompson R (2013) Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood Estimation of Genetic Parameters. PloS One 8(12) e80821.

Meuwissen TMH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Meyer K (2007) WOMBAT: A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B 8(11): 815-821.

Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm, Genetics Selection Evolution (21): 317-340.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2002) BLUPF90 and related programs (BGF90). In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. Session 28. (pp. 1-2). Institut National de la Recherche Agronomique (INRA).

Morota G, Boddhireddy P, Vukasinovic N, Gianola D, DeNise S (2014) Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. Front. Genet. 5(56):10-3389.

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell Online 21(8): 2194-2202.

Nelder JA, Mead R (1965) A simplex method for function minimization. The computer journal 7(4): 308-313.

Nyquist WE, Baker RJ (1991) Estimation of heritability and prediction of selection response in plant populations. Critical reviews in plant sciences 10(3): 235-322.

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods:

what, how and which. Bayesian analysis 4(1): 85-117.

Orr HA (2005) The genetic theory of adaptation: a brief history. Nature Reviews Genetics 6(2): 119-127.

Park T, Casella G (2008) The bayesian lasso. Journal of the American Statistical Association 103(482): 681-686.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3): 545-554.

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics, genetics-114.

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1-2): 209-228.

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Science 49(4):1165-76.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5(3): 92-102.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38(8):904-9.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. Genetics 155(2):945-59.

Rasmussen CE (2004) Gaussian processes in machine learning. In Advanced Lectures on Machine Learning, pp. 63-71. Springer Berlin Heidelberg.

Recker JR, Burton JW, Cardinal A, Miranda L (2014) Genetic and Phenotypic Correlations of Quantitative Traits in Two Long-Term, Randomly Mated Soybean Populations. Crop Science 54(3): 939-943.

Rincker K, Nelson R, Specht J, Sleper D, Cary T, Cianzio SR, *et al.* (2014) Genetic improvement of US soybean in Maturity Groups II, III, and IV. Crop Science.

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statistical science, 15-32.

Rutkoski JE, Poland J, Jannink JL, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes| Genomes| Genetics 3(3): 427-439.

Searle SR (1979) Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology. Paper BU-673M, Biometrics Unit, Cornell University.

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2014) Identification of loci governing eight agronomic traits using a GBS|GWAS approach and validation by QTL mapping in soya bean. Plant biotechnology journal.

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer.

Specht J E, Hume DJ, Kumudini SV (1999) Soybean yield potential-a genetic and physiological perspective. Crop Science 39(6): 1560-1570.

Strandén I, Christensen OF (2011) Allele coding in genomic evaluation. Genet Sel Evol 43(1).

St. Martin SK (1982) Effective population size for the soybean improvement program in maturity groups 00 to IV. Crop Science 22(1): 151-152.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. Nature genetics 44(10): 1166-1170.

Swarts K, Li H, Romero Navarro JA, An D, Romay MC, Hearne S, *et al.* (2014) Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant Genome 7(3).

Tabangin ME, Woo JG, Martin LJ (2009, December) The effect of minor allele frequency on the likelihood of obtaining false positives. In BMC proceedings, Vol. 3, No. Suppl 7, p. S41. BioMed Central Ltd.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91(11): 4414-4423.

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genetics Selection Evolution 25:41-62.

Wei J, Xu S (2016). A Random Model Approach to QTL Mapping in Multi-parent Advanced Generation Inter-cross (MAGIC) Populations. Genetics. 202(2):471-486.

Wen ZX, Zhao TJ, Zheng YZ, Liu SH, Wang CE, Wang F, Gai JY (2008) Association analysis of agronomic and quality traits with SSR markers in Glycine max and Glycine soja in China: I. Population structure and associated markers. Acta Agronomica Sinica 34(7): 1169-1178.

Wricke G, Weber E (1986) Quantitative genetics and selection in plant breeding. Walter de Gruyter.

Wright S (1930) Evolution in Mendelian populations. Genetics 16(2): 97.

Wright S (1922) Coefficients of inbreeding and relationship. American Naturalist, 330-338.

Xavier A, Xu S, Muir WM, and Rainey KM (2015) NAM: Association Studies in Multiple Populations. Bioinformatics, btv448.

Xavier A, Muir WM, Rainey KM (2016). Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC bioinformatics 17(1):1.

Xu S (2013) Mapping quantitative trait loci by controlling polygenic background effect. Genetics 195(4): 1209-1222.

Xu H, Shete S (2005) Effects of population structure on genetic association studies. BMC genetics 6(Suppl 1):S109.

Xu S (2003a) Theoretical basis of the Beavis effect. Genetics 165(4): 2259-2268.

Xu S (2003b) Estimating polygenic effects using markers of the entire genome. Genetics 163(2): 789-801.

Yan W, Rajcan I (2003) Prediction of cultivar performance based on single-versus multiple-year tests in soybean. Crop Science 43(2): 549-555.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nature genetics 46(2): 100-106.

Yi N, and Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179(2): 1045-1055.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, *et al.* (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38(2): 203-208.

Zas R (2006) Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree genetics and genomes 2(4): 177-185.

Zeng ZB, Wang T, Zou W (2005) Modeling quantitative trait loci and interpretation of models. Genetics 169(3):1711-25.

Zeng ZB Hill WG (1986) The selection limit due to the conflict between truncation and stabilizing selection with mutation. Genetics 114(4): 1313-1328.

Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q (2010a) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PloS one. 2010 Sep 9;5(9):e12648.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, *et al.* (2010b) Mixed linear model approach adapted for genome-wide association studies. Nature genetics 42(4): 355-360.

Zhang LX, Kyei-Boahen S, Zhang J, Zhang MH, Freeland TB, Watson CE, Liu X (2007) Modifications of optimum adaptation zones for soybean maturity groups in the USA. Crop Management, 6(1).

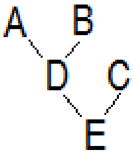
Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nature genetics 44(7): 821-824.

Zhou X, Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods 11(4): 407-409.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2):301-20.

APPENDIX: Numerical example of design matrices

Suppose that a breeding program is conducting a test with a three-way hybrid $(A \times B \times C)$ to find out the narrow-sense heritability of the trait of interest. The only genetic information available is a short pedigree that describes the three-way cross, as follows:



This evaluation was conducted in a single environment, growing two replicates of each parent (A, B, C) and the final hybrid (E). Considering that a plot with genotype C was lost during the growing season, the design matrices are given by:

$$\mathbf{y} = \begin{array}{c} \text{Yield} \\ \left[\begin{array}{c} 25 \\ 27 \\ 27 \\ 21 \\ 20 \\ 21 \\ 27 \end{array}\right] \end{array} \quad \mathbf{X} = \begin{array}{c} \text{Intercept} \\ \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}\right] \end{array} \quad \mathbf{Z} = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \quad \mathbf{K} = \begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 1 & 0 & 0 & 0.5 & 0.25 \\ B & 0 & 1 & 0 & 0.5 & 0.25 \\ C & 0 & 0 & 1 & 0 & 0.5 \\ D & 0.5 & 0.5 & 0 & 1 & 0.5 \\ E & 0.25 & 0.25 & 0.5 & 0.5 & 1 \end{array}$$

The example above was run using the Gibbs sampling algorithm shown in the manuscript, with the prior suggested here ($v^* = 5$ and $S^* = 0.5 \times var(\mathbf{y}) = 5.17$). The outcome was:

$$\mathbf{b} = \left[\begin{array}{c} 23.812 \end{array}\right] \quad \mathbf{u} = \left[\begin{array}{c} 1.191 \\ 0.172 \\ -1.291 \\ 0.799 \\ -0.060 \end{array}\right] \quad \sigma^2_a = 4.004 \quad \sigma^2_e = 6.987$$

which yields a narrow-sense heritability of 0.364, and breeding values $(\mathbf{u})$ computed for all genotypes, including the parental line D not grown in the field.