frontiers in GENETICS

REVIEW ARTICLE
published: 16 October 2014
doi: 10.3389/fgene.2014.00363

Kernel-based whole-genome prediction of complex traits: a review

Gota Morota[1]* and Daniel Gianola[2,3,4]

[1] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, USA
[2] Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, USA
[3] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
[4] Department of Dairy Science, University of Wisconsin-Madison, Madison, WI, USA

Edited by:
Eduardo Manfredi, Institut National de la Recherche Agronomique, France

Reviewed by:
Paola Sebastiani, Boston University, USA
Daniel J. Schaid, Mayo Clinic, USA

***Correspondence:**
Gota Morota, Department of Animal Science, University of Nebraska-Lincoln, Animal Science Building, Lincoln, NE 68533, USA
e-mail: morota@unl.edu

Prediction of genetic values has been a focus of applied quantitative genetics since the beginning of the 20th century, with renewed interest following the advent of the era of whole genome-enabled prediction. Opportunities offered by the emergence of high-dimensional genomic data fueled by post-Sanger sequencing technologies, especially molecular markers, have driven researchers to extend Ronald Fisher and Sewall Wright's models to confront new challenges. In particular, kernel methods are gaining consideration as a regression method of choice for genome-enabled prediction. Complex traits are presumably influenced by many genomic regions working in concert with others (clearly so when considering pathways), thus generating interactions. Motivated by this view, a growing number of statistical approaches based on kernels attempt to capture non-additive effects, either parametrically or non-parametrically. This review centers on whole-genome regression using kernel methods applied to a wide range of quantitative traits of agricultural importance in animals and plants. We discuss various kernel-based approaches tailored to capturing total genetic variation, with the aim of arriving at an enhanced predictive performance in the light of available genome annotation information. Connections between prediction machines born in animal breeding, statistics, and machine learning are revisited, and their empirical prediction performance is discussed. Overall, while some encouraging results have been obtained with non-parametric kernels, recovering non-additive genetic variation in a validation dataset remains a challenge in quantitative genetics.

Keywords: whole-genome prediction, kernel methods, semi-parametric regression, spatial distance, SNP

1. INTRODUCTION

Six years after the rediscovery of Mendel's laws of heredity, Toyama Kametaro's experimental work on silkworm breeding showed the first case of Mendelian inheritance in animals (Onaga, 2010). Yule (1902) made a first attempt at expanding Mendelian theory to factor in quantitative variation, followed by a seminal paper by Fisher (1918) nearly a century ago (Plutynski, 2006). A fundamental concept in quantitative genetics is that of linking genotypes and phenotypes through genetic similarity among individuals, i.e., covariance between relatives (Wright, 1921). The main focus today is to statistically model variation in DNA sequences influencing phenotypic variation in quantitative traits, rather than understanding the biological pathways that are associated with selective genes of interest, which falls in the domain of molecular genetics. The discipline of genome-based prediction is a subfield of quantitative genetics that aims to predict unobserved values by regressing phenotypes on measures of genetic resemblance, obtained from DNA data. Although early attempts took place in the 80's (e.g., Fernando and Grossman, 1989; Lande and Thompson, 1990), implementation of genome-based prediction was largely hindered by scarce molecular information. It was just recently that the subject began to attract widespread attention following the availability of rich DNA variation data spanning the whole genome (e.g., Meuwissen et al., 2001; Gianola et al., 2003). This approach continues to progress rapidly and has been fruitfully applied to a variety of quantitative traits of agronomic importance in animals (e.g., Hayes et al., 2009; VanRaden et al., 2009) and plants (e.g., Crossa et al., 2014). The objective of "genome-enabled selection" is to predict responses by capturing additive genetic effects that may have implication in choosing individuals as parents of the next generation. Statistical methodologies tailored to this application have been reviewed in a number of papers (e.g., Gianola et al., 2009; Calus, 2010; de los Campos et al., 2013a; Gianola, 2013; Meuwissen et al., 2013).

Concurrently, whole-genome prediction of "total" genetic effects has been motivated by the fact that phenotypes and genotypes may not be linearly related and that the additivity assumption, even though useful, is violated (Gianola et al., 2010). Importance of predicting non-additive genetic effects comes into the picture in exploitation of heterosis, mate allocation, cross-breeding, and precision mating in breeding contexts, and more crucially when prediction of phenotypes is a primal point, such as

disease outcome. The objective of this article is to provide a survey of emerging statistical approaches based on kernel methods with emphasis on "prediction" rather than on genome-enabled selection for breeding. Special focus is placed on a semiparametric kernel methodology that condenses genealogical or genomic information into a positive (semi) definite relationship matrix. We highlight insights collected from research conducted in recent years and suggest potential future directions in this area.

In the next section, we go through statistical models involving the use of kernels. Subsequently, we review a variety of kernel matrices that have been applied to date. We then survey applications of kernel methods to real data in a whole-genome prediction framework, and concluding remarks are given in the final section.

2. KERNEL-BASED REGRESSION METHODS

We first review kernel-based prediction models being used for prediction using genomic data. Our aim is to approximate an unknown "true" genetic signal $\mathbf{g}$ with a certain function of a marker genotypes matrix $f(\mathbf{X})$ that maps these genotypes to responses ($\mathbf{y}$). The data generating model is then $\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is the vector of phenotypes and $\boldsymbol{\epsilon}$ is a vector of residuals. In general, all kernel methods differ from each other in the choice of the mapping function $f(.)$ and the type of regularization used to balance goodness of fit and complexity, as discussed later.

2.1. GENOMIC BLUP

Our main interest is to identify the model that gives best prediction among a set of candidate models. To find the best predictive function $f(\mathbf{X})$ there are a few things that we need to set up prior to the search. One is whether we should impose a restriction on the search space or not. The parameter space is a space where all models are characterized by parameters. In a linear model, where response values have linear relationship with respect to the parameters, though, it could be non-linear on covariates such as in the case of polynomial regression. Once the parameters are given, all models are distinct. For example, the two linear models would be

$$\text{Model 1: } y = a_1 + b_1 x_1 + c_1 x_2$$
$$\text{Model 2: } y = a_2 + b_2 x_1 + c_2 x_2.$$

If we give values to parameters a_i, b_i, and c_i then Models 1 and 2 can be differentiable. In other words, we just need to fill out the unknown parameters of the given models. The best linear unbiased predictor (BLUP) is a procedure for filling the unknown values (Henderson, 1975).

Suppose underlying signal is given by $\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}$. Here, $\mathbf{g}$ is the true unknown genetic signal with $\mathbf{g} \sim N(0, \boldsymbol{\Gamma}\siga^2_{\mathbf{g}})$, where $\boldsymbol{\Gamma}$ is a "true" genomic relationship matrix among animals, i.e., at the quantitative trait loci affecting the trait. Since the latter are unknown, we approximate the vector of genetic values $\mathbf{g}$ with a linear function such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an n (number of animals) $\times$ m (number of markers) matrix of additive marker genotypes potentially centered or scaled; $\boldsymbol{\beta}$ is a vector of regression coefficients on marker genotypes; and $\boldsymbol{\epsilon}$ is a residual that includes model misspecification and environmental effects not considered in this analysis. Under the independence assumption between $\mathbf{g}$ and $\boldsymbol{\epsilon}$, the variance-covariance matrix of $\mathbf{y}$ is

$$\begin{aligned}\mathbf{V}_y &= \mathbf{V}_g + \mathbf{V}_\epsilon \\ &= \mathbf{X}\mathbf{X}^T\sigma^2_\beta + \mathbf{I}\sigma^2_\epsilon\end{aligned}$$

often assuming that $\boldsymbol{\beta} \sim N(0, \mathbf{I}\sigma^2_{\boldsymbol{\beta}})$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma^2_{\boldsymbol{\epsilon}})$. Here, $\mathbf{V}_g = \mathbf{X}\mathbf{X}^T\sigma^2_\beta$ is the covariance matrix "due to" markers. The problem is to predict $\mathbf{g}$ such that two conditions are met: (1) E($\hat{\mathbf{g}}$)=E($\mathbf{g}$)=0, and (2) var($\hat{g}_i - g_i$) is minimum for i over all linear functions that satisfy the unbiasedness condition (1). If normality is assumed, the BLUP of $\mathbf{g}(\hat{\mathbf{g}})$ is the conditional mean of $\mathbf{g}$ given the data, and takes the form

$$\begin{aligned}BLUP(\hat{\mathbf{g}}) &= E(\mathbf{g}|\mathbf{y}) = E[\mathbf{g}] + Cov\left(\mathbf{g}, \mathbf{y}^T\right) Var(\mathbf{y})^{-1}[\mathbf{y} - E(\mathbf{y})] \\ &= Cov\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}^T\right) \cdot \mathbf{V}_y^{-1}\mathbf{y} \\ &= \mathbf{X}\mathbf{X}^T\sigma^2_{\boldsymbol{\beta}}\left[\mathbf{X}\mathbf{X}^T\sigma^2_\beta + \mathbf{I}\sigma^2_\epsilon\right]^{-1}\mathbf{y} \\ &= \left[\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{\sigma^2_\epsilon}{\sigma^2_\beta}\right]^{-1}\mathbf{y}, \end{aligned} \quad (1)$$

assuming that $\mathbf{X}\mathbf{X}^T$ is invertible. Here, $Cov(\mathbf{X}) = \mathbf{X}\mathbf{X}^T$ is a covariance matrix of marker genotypes (provided that X is centered), often considered to be the simplest form of additive genomic relationship kernel, $\mathbf{G}$. We can refine this kernel by relating genetic variance σ^2_g and marker variance σ^2_β under the following assumptions. Again, assume genetic value is parameterized as $g_i = \sum x_{ij}\beta_j$, where both x and β are treated as random and independent. Under Hardy-Weinberg equilibrium, $E(x_{ij}) = 2p_j$ and $Var(x_{ij}) = 2p_j(1 - p_j)$, where p_j is the minor allele frequency of locus j, and assuming linkage equilibrium of markers (all loci are mutually independent),

$$\sigma^2_g = \textstyle\sum_j 2p_j(1 - p_j) \cdot \sigma^2_{\beta_j}.$$

Under the homogeneous marker variance assumption, one obtains the relationship

$$\sigma^2_\beta = \frac{\sigma^2_g}{2\sum_j p_j(1-p_j)}. \quad (2)$$

Replacing σ^2_β in Equation (1) with (2), we get

$$\begin{aligned}BLUP(\hat{\mathbf{g}}) &= \left[\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{\sigma^2_\epsilon}{\frac{\sigma^2_g}{2\sum_j p_j(1-p_j)}}\right]^{-1}\mathbf{y} \\ &= \left[\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma^2_\epsilon}{\sigma^2_g}\right]^{-1}\mathbf{y} \end{aligned} \quad (3)$$

where $\mathbf{G} = \frac{\mathbf{X}\mathbf{X}^T}{2\sum_j p_j(1-p_j)}$ is known as the first **G** matrix introduced in VanRaden (2008).

Likewise, $\sigma_g^2 = m\sigma_\beta^2$ if it is assumed that all markers have variance 1 (following standardizing marker genotypes) so the marked genetic variance is given by the sum of individual marker variances. With this, $\sigma_\beta^2 = \sigma_g^2/m$, and **G** in Equation (3) becomes the second additive genomic relationship kernel of VanRaden (2008), $\frac{\mathbf{X}\mathbf{X}^T}{m}$. One can also approximate $\boldsymbol{\Gamma}$ with a pedigree-based relationship kernel **A** instead of **G**, embedding correlations due to expected additive genetic inheritance. How close **G** approximates $\boldsymbol{\Gamma}$ depends on observed or tagged causal loci in the data. A third type of kernel matrix is a linear combination $\lambda\mathbf{A} + (1-\lambda)\mathbf{G}$, were $0 < \lambda < 1$ is the weight placed on **A** relative to **G** (e.g., Rodríguez-Ramilo et al., 2014).

2.2. REPRODUCING KERNEL HILBERT SPACES REGRESSION

Genomic BLUP (GBLUP) is a linear model characterized by parameters that relate to additive quantitative genetics theory. We now extend the search space by eliminating some restrictions. For instance, semiparametric regressions identify functional forms "before as well as in the midst" of the fitting process (Berk, 2008). A reproducing kernel Hilbert spaces (RKHS) regression represents this type of approach. Here, the true genetic signal $\mathbf{g} = \{g_i\}$ is approximated as an unknown function of genetic effects $g(\mathbf{x}_i)$, contrary to assuming $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$ as in the case of GBLUP. This $g(\mathbf{x}_i)$ function is viewed as the conditional expectation $E(\mathbf{y}_i|\mathbf{x} = \mathbf{x}_i)$, that is, the average phenotypic value of individuals possessing marker genotype $\mathbf{x}_i$ without restricting the form of g_i. RKHS regression proceeds by searching a function and uses the residual sum of squares as a loss function, and assigns the squared norm of **g** under a Hilbert space as a penalty. The objective function to be minimized with respect to **g** is

$$\ell(\mathbf{g}|\lambda) = \|\mathbf{y} - \mathbf{g}\|^2 + \lambda\|g\|^2_{\mathcal{H}}, \tag{4}$$

where λ is a regularization parameter and $\mathcal{H}$ represents a Hilbert space, very rich class of functions. While there are countless candidates for **g** in non-parametric regression, with this setting, the representer theorem developed by Kimeldorf and Wahba (1971) guarantees that the optimizer will be in the span of the functions indexed by the observed covariates. This implies that the objective function reduces to a linear function $\mathbf{K}\boldsymbol{\alpha}$, where **K** is an $n \times n$ kernel constructed from the observed data and $\boldsymbol{\alpha}$ is an $n \times 1$ vector of regression coefficients to be inferred, e.g., by minimizing

$$\ell(\boldsymbol{\alpha}|\lambda) = (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}. \tag{5}$$

Equation (5) is minimized by taking its derivative with respect to $\boldsymbol{\alpha}$ and setting to 0 to obtain:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}, \tag{6}$$

so that the predicted genetic value is given by $\hat{\mathbf{g}} = \mathbf{K}\hat{\boldsymbol{\alpha}}$; this requires λ to be known.

This regression procedure is also known as kernel ridge regression in machine learning, and was first introduced in quantitative genetics by Gianola et al. (2006) and Gianola and van Kaam (2008) in the context of a mixed effects model with a Bayesian treatment. Efficient Gibbs sampling algorithms for RKHS regression have been developed by de los Campos et al. (2010) by exploiting the eigendecomposition of kernels. When the first term in Equation (4) is replaced by the "epsilon-insensitive" loss function, this is equivalent to support vector regression (Moser et al., 2009; Long et al., 2011). Hence, RKHS represents a general and powerful paradigm.

Note that the representer theorem requires assigning the $L2$ norm (Euclidean norm) to regularize the regressions $\boldsymbol{\alpha}$. This regularizer assures that optimal solutions lie in a finite-dimensional rather than in an infinite-dimensional space. In practical situations, sparsity induced by the $L1$ norm (Manhattan norm) may be preferable. This is equivalent to replacing $\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$ in Equation (5) with $\sum_{i=1}^n |\alpha_i|$. However, this violates assumptions of the representer theorem and it no longer guarantees that the optimizer is given by a linear combination of the data points. Nevertheless, it is conceivable that norms other than $L2$ may deliver better predictions.

2.3. RKHS AND BLUP

The important connection between RKHS regression and BLUP was brought up first by Harville (1983); Robinson (1991) and de los Campos et al. (2009) and this part of the review follows their work closely. Suppose we approximate a genetic signal with a vector of additive effects and assume a single record per individual. The model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{7}$$

where **y** is the response variable, **X** is an incidence matrix linking the response to some nuisance effects; $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are regression coefficients and $\boldsymbol{\epsilon}$ is a residual that includes model misspecification and environmental effects not considered in this analysis. The two random components of the model follow the distribution $\boldsymbol{\alpha} \sim N(0, \mathbf{A}\sigma_\alpha^2)$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_\epsilon^2)$, where σ_α^2 is the additive genetic variance and **A** is the numerator relationship matrix between individuals (relationships in the absence of inbreeding). Henderson's mixed model equations (MME) are

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T \\ \mathbf{X} & \mathbf{I} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{y} \end{bmatrix}. \tag{8}$$

Now, transform additive genetic effects into $\boldsymbol{\alpha}^* = \mathbf{A}^{-1}\boldsymbol{\alpha}$ (assuming that $\mathbf{A}^{-1}$ exists) so that Equation (7) is reexpressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\alpha}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{\alpha}^* \sim N(0, \mathbf{A}^{-1}\sigma_\alpha^2)$, and the corresponding MME are

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{A} \\ \mathbf{A}\mathbf{X} & \mathbf{A}^2 + \mathbf{A}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{A}\mathbf{y} \end{bmatrix}, \tag{9}$$

because $\mathbf{A}$ is symmetric. By multiplying the $\boldsymbol{\alpha}^*$ equation by $\mathbf{A}^{-1}$, one obtains

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{A} \\ \mathbf{X} & \mathbf{A}+\mathbf{I}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{y} \end{bmatrix}. \tag{10}$$

If phenotypes are pre-corrected for systematic effects (thus $\boldsymbol{\beta} = 0$) a priori, then Equation (10) reduces to

$$\begin{aligned} [\mathbf{A}+\lambda\mathbf{I}]\boldsymbol{\alpha}^* &= \mathbf{y} \\ \hat{\boldsymbol{\alpha}}^* &= (\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{y} \end{aligned} \tag{11}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$ is a regularization parameter. Replacing the pedigree-based relationship kernel $\mathbf{A}$ with a more general term $\mathbf{K}$ yields $\boldsymbol{\alpha}^* = [\mathbf{K}+\lambda\mathbf{I}]^{-1}\mathbf{y}$. Since BLUP is linearly invariant, the BLUP of $\boldsymbol{\alpha}$ is given by $\hat{\boldsymbol{\alpha}} = \mathbf{A}^{-1}\hat{\boldsymbol{\alpha}}^* = \mathbf{A}^{-1}(\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{y}$. This is equivalent to Equation (6) and is the Bayesian kernel ridge regression employed in de los Campos et al. (2010) and Morota et al. (2013). Thus, BLUP of additive effects can be viewed as a regression on pedigree or on additive genomic relationship kernels. It is interesting to note that BLUP, developed in animal breeding, is a special case of RKHS, developed in statistics.

2.4. CONNECTION BETWEEN THE KERNEL AND THE MATRIX OF GENOTYPES

GBLUP is also linked to BLUP of marker regression coefficients. Here, we show how BLUP of regression on markers and BLUP of additive genetic values are related to each other. This relationship was first shown by Henderson (1977) in the context of predicting BLUP of non-phenotyped animals, and was rediscovered recently (e.g., Goddard, 2009). Suppose that the phenotype-genotype mapping function is $\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}$, where the genetic effect is parameterized as $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$. Here, $\mathbf{X}$ is the matrix of marker genotypes and $\boldsymbol{\beta}$ is the marker allele substitution effects. Then, the conditional expectation of $\boldsymbol{\beta}$ given $\mathbf{y}$ assuming known dispersion parameters and $\boldsymbol{\beta} \sim N(0, \mathbf{I}\sigma_\beta^2)$ is

$$\begin{aligned} BLUP(\boldsymbol{\beta}) &= E(\boldsymbol{\beta}|\mathbf{y}) = Cov(\boldsymbol{\beta}, \mathbf{y})Var(\mathbf{y})^{-1}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] \\ &= Cov(\boldsymbol{\beta}, \mathbf{X}\boldsymbol{\beta})\left[\mathbf{X}\mathbf{X}^T\sigma_\beta^2 + \mathbf{I}\sigma_\epsilon^2\right]^{-1}\mathbf{y} \\ &= \sigma_\beta^2\mathbf{X}^T\left[\mathbf{X}\mathbf{X}^T\sigma_\beta^2 + \mathbf{I}\sigma_\epsilon^2\right]^{-1}\mathbf{y} \\ &= \sigma_\beta^2\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\left[\sigma_\beta^2\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\sigma_\epsilon^2\right]^{-1}\mathbf{y} \\ &= \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\left[\mathbf{I} + \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{\sigma_\epsilon^2}{\sigma_\beta^2}\right]^{-1}\mathbf{y}. \end{aligned}$$

Using Equation (3), we get

$$\begin{aligned} BLUP(\boldsymbol{\beta}) &= \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\left[\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_\epsilon^2}{\sigma_\mathbf{g}^2}\right]^{-1}\mathbf{y} \\ &= \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}BLUP(\mathbf{g}). \end{aligned}$$

Thus, once we obtain $\hat{\mathbf{g}}$ from GBLUP, BLUP of marker coefficients is given by $\hat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\hat{\mathbf{g}}$. We arrive at the same prediction regardless of whether we start from the genotype matrix $\mathbf{X}$ or from the $\mathbf{g}$. Note that marker-based regressions can also be "kernelized" when the squared norm of $\boldsymbol{\beta}$ is assigned as penalty function. Ridge regression with markers treated as random effects is known to be equivalent to BLUP, i.e., RR-BLUP (Ruppert et al., 2003). However, the least absolute shrinkage and selection operator (LASSO) does not satisfy this condition. Although RR-BLUP and GBLUP are mathematically equivalent, predictive ability has differed when applied to real data. For example, Zhong et al. (2009) and Massman et al. (2013) observed a superiority of RR-BLUP over GBLUP in the presence of strong LD, and Habier et al. (2013) reported that GBLUP was not able to utilize short-range LD information for prediction.

2.5. RKHS FAMILY

It should be noted that the family of RKHS also extends to independently developed spatial statistics regression (Stein, 1999). Kriging is a variant of BLUP used for predicting values of variables distributed over a space (Stein, 1999). It is a spatial regression technique based on spatial associations at various locations. We first revisit the space which kriging takes place when applied to genomic data.

2.5.1. Spatial genotypic structure

Suppose that all individuals in a sample have been genotyped for biallelic SNPs spanned over the whole genome, with these SNPs used for genome-enabled prediction of quantitative traits. The number of possible genotypic configurations with m SNPs is 3^m, which is an enormous number when m exceeds the thousands. Characterizing the metric space governed by these SNP predictors prior to the analysis is of importance in a spatial prediction problem. Recently, Morota et al. (2013) used SNP codes as coordinates of genotypes in m-dimensional spaces. Therein, an m-dimensional grid graph with vertices representing a vector of individual's genotypes was proposed as a spatial structure of genotypes. This m-dimensional non-Euclidean metric space can be constructed once the total number of SNP genotypes is obtained. Then, each sampled individual in the data set is placed at one of the vertex on this space. With genotypes coded as (0, 1, 2) for "aa," "Aa," and "AA," respectively, two vertices are adjacent if and only if codes at just one SNP locus differ by 1.

In standard regression analysis, independent and identically distributed (i.i.d.) observations is a common assumption while, in kriging, spatially correlated random variables are considered (Isaaks and Srivastava, 1990). The mixed models of Henderson (1975) do not take spatial association of individuals into account. However, in genetics it is known that individuals are often genetically related to each other, especially in animal breeding. For this reason, the numerator relationship matrix $\mathbf{A}$ or the genomic relationship matrix $\mathbf{G}$ among individuals are employed. With this, the set of additive effects in a sample of individuals is assumed to follow a multivariate normal distribution with a zero mean vector and a covariance matrix that is proportional to $\mathbf{A}$ or $\mathbf{G}$, instead of the identity matrix $\mathbf{I}$. On the other hand, kriging explicitly assumes at the onset that observed values are spatially correlated.

Most of observed data are not randomly sampled. This is particularly so in animal breeding, where individuals genotyped tend to be highly related due to intense artificial selection. When we embed these samples into a 3^m grid graph, it is unlikely that points will be uniformly distributed in this space. We would expect to see clusters of individuals because of their genotypic similarity due to selection, or no clusters at all in certain regions. A kriging system attempts to smooth the irregular spatial variation in the p-dimensional space, to arrive at a better prediction of outcomes (Isaaks and Srivastava, 1990). The main goal here is to predict phenotypes at unsampled vertices using data obtained at limited number of neighboring spatial locations. The kriging system, which performs an interpolation on this grid space, is briefly explained in the following subsection.

2.5.2. Kriging

For simplicity, only ordinary kriging is reviewed. Suppose that available phenotypes are viewed as the result of some stochastic process with $y_i = g(\mathbf{x}) + e_i$ having the measurement on individual i ($i = 1, \cdots, n$) at point x_i. A random field g is a random function on a space $\mathbb{D}$, which in our case is the p-dimensional grid graph $\mathbb{Z}_3^p$ of all SNP genotypes, and x_i and e_i are the ith observed vertex and residual. In other words, y_i is a response value associated with a spatial point location $x_.$ which is a vertex. We assume second order stationarity in the observations, that is, the mean is constant but unknown and the covariance depends only on the "distance" between any of two vertices ($||x_i - x_j||$), but not on the locations *per se*. We wish to predict the phenotype of an individual having vertex x_0 using all available data. Kriging is the best predictor (Henderson, 1973) in the mean-squared error sense,

$$\arg\min_{\hat{y}_0}\{\mathrm{E}[\underbrace{\hat{y}_0}_{predicted} - \underbrace{y_0}_{unobserved}]^2\}$$

where y_0 is the phenotype at vertex x_0 to be predicted and $\hat{y}_0$ is its predictor.

As in BLUP, in kriging, the predictor is restricted to the class of linear functions of the data and given by a weighted linear combination of all available observations, that is,

$$\hat{y}_0 = \sum_{i=1}^{n} w_i y_i, \tag{12}$$

where $w_1, w_2, \cdots, w_n$ are weights that need to be found. The question boils down to how to weight nearby samples of the prediction point in question, and this is pertinent to other whole-genome prediction methods as well. Predicted values are represented as a weighted sum of the observed phenotypes and are functions of a projection matrix $\mathbf{H}$. For example, BLUP and kernel ridge regression are represented as $\hat{\mathbf{y}}_{BLUP} = \mathbf{H}_{BLUP}\mathbf{y}$ and $\hat{\mathbf{y}}_{RKHS} = \mathbf{H}_{RKHS}\mathbf{y}$, respectively, where $\mathbf{H}_{BLUP} = [\mathbf{I} + \mathbf{G}^{-1}\lambda]^{-1}$ and $\mathbf{H}_{RKHS} = \mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}$. Weights may change as we move along when predicting a response at next unobserved vertex in kriging.

In the context of prediction of a random variable, an unbiased predictor requires that the expected value of the difference between predictor and predictand be zero. If the predictor is linear as in Equation (12),

$$\begin{aligned} E[\hat{y}_0 - y_0] &= E\left[\sum_{i=1}^{n} w_i y_i - y_0\right] \\ &= \sum_{i=1}^{n} w_i E[y_i] - E[y_0] \\ &= \mu\sum_{i=1}^{n} w_i - \mu, \end{aligned} \tag{13}$$

where (assuming no nuisance parameters) μ is the mean values of the responses; $E[y_i] = E[y_0] = \mu$ holds because of the stationarity assumption made. Setting Equation (13) to zero gives the unbiasedness condition

$$\mu_1\left(\sum_{i=1}^{n} w_i - 1\right) = 0. \tag{14}$$

Thus, $\hat{y}_0$ is an unbiased predictor if $\sum_{i=1}^{n} w_i = 1$. This is called the "normed weights" condition.

As defined by Henderson (1973), BLUP is a linear unbiased predictor with minimum prediction error variance in such class. Under the unbiasedness condition, this is attained by minimizing the expected squared prediction error.

$$\begin{aligned} Var\left([\hat{y}_0 - y_0]\right) &= E[\hat{y}_0 - y_0]^2 \\ &= Var\left[\sum_{i=1}^{n} w_i y_i - y_0\right] \end{aligned} \tag{15}$$

Adding the unbiasedness conditions to the above equation via a Lagrange multiplier (λ) and setting the $n + 1$ partial first derivatives with respect $\mathbf{w}$ and λ to 0, one obtains the kriging system of linear equations shown below.

$$\underbrace{\begin{bmatrix} \mathbf{V} + \mathbf{I}\sigma_\epsilon^2 & \mathbf{1}' \\ \mathbf{1} & 0 \end{bmatrix}}_{(n+1)\times(n+1)} \cdot \underbrace{\begin{bmatrix} \hat{\mathrm{w}} \\ \lambda \end{bmatrix}}_{(n+1)\times 1} = \underbrace{\begin{bmatrix} \mathbf{C_{0i}} + \mathbf{I}\sigma_\epsilon^2 \\ 1 \end{bmatrix}}_{(n+1)\times 1}, \tag{16}$$

where the left-hand side includes the covariance function $\mathbf{V} = \{\mathbf{Cov}(\mathbf{g}(\mathbf{x_i}), \mathbf{g}(\mathbf{x_j}))\}$, the residual covariance matrix $\mathbf{I}\sigma_e^2$, the vector of ones $\mathbf{1}$. The right-hand side contains covariances between the unsampled location to be predicted and sampled locations $\mathbf{C_{0i}} = \{\mathbf{Cov}(g(x_0), \mathbf{g}(\mathbf{x}_i))\}$. We predict a value at some unsampled vertices by replacing the unknown weights with $\hat{\mathbf{w}}$. Although derived from a different perspective, the predicted value $\hat{y}(x_0)$ is essentially BLUP of Henderson (Robinson, 1991): namely, the projection of yet-to-be-observed data on a linear combination of observed data. Ober et al. (2011) applied two alternatives to this ordinary kriging, named simple kriging and universal kriging, in the context of whole-genome prediction. Kriging has also been used for predicting human disease outcomes by condensing

genomic and RNA expression data into kernel matrices (Wheeler et al., 2014). It is worthwhile to note that kriging is equivalent to Gaussian process regression in the machine learning literature (Rasmussen and Williams, 2005). In summary, BLUP in animal breeding, RKHS in statistics, kriging in geostatistics, and Gaussian process regression in machine learning all share the same spirit.

3. KERNEL MATRICES

A suite of kernel functions has been proposed for whole-genome prediction purposes. Here, we discuss several kernels that can be used conjunction with the aforementioned regression methodologies. All kernels are a special case of a matrix denoted as **K**. A "parametric" kernel aims to capture the signal from some specific gene action, e.g., dominance. This approach permits making claims or interpretation with respect to some theory about hidden genetic architecture. At the other end of the spectrum, use of a "non-parametric" kernel is purely driven by prediction purposes. The kernel may pick up genetic signals regardless of the underlying genetic architecture, and model coefficients typically do not have a theoretical interpretation.

3.1. PARAMETRIC KERNELS

Most of the relationship matrices animal breeders have been using for many years are valid RKHS kernels, e.g., the additive genetic relationship matrix **A**, which is calculated directly from pedigree information. The idea here is to use expectation of genetic relatedness in the absence of genomic information as kernel. The off-diagonals of this matrix are twice the kinship coefficients, and 1 + individual's inbreeding coefficients are placed along the diagonal (Wright, 1922; Malécot, 1948). This is the perhaps the oldest kernel function used in quantitative genetics and is proportional to identical by descent (IBD) probabilities. Similarly, one can also trace expected dominance relationship coefficients using a pedigree (Henderson, 1985). A quantitative trait loci (QTL)-based counterpart of the additive genetic relationship matrix gained attention afterwards (Fernando and Grossman, 1989; Nejati-Javaremi et al., 1997; Villanueva et al., 2005).

The genomic relationship matrix **G** used in GBLUP also represent parametric kernels, using additively coded genotypes. Other types of genomic relationship matrices have been compared and discussed by Toro et al. (2011). As stated earlier, molecular similarity generates covariance even if individuals are not related in the sense of pedigree. Contrary to pre-genomics quantitative genetics, built largely on related individuals, the use of genomic data without genetic relatedness expands quantitative genetics theory (e.g., Yang et al., 2010). Analogous to **G**, the dominance counterpart **D** is constructed by setting up an appropriate dominance contrast between genotypes, for example, AA $= -1$, Aa $= 0$, and aa $= -1$ (Su et al., 2012; Vitezica et al., 2013; Da et al., 2014).

3.2. NON-PARAMETRIC KERNELS

The genomic relationship matrix of VanRaden (2008) represents identical by state (IBS) similarities so there is no requirement to trace back genealogy. It is possible to compute an IBS matrix non-parametrically to enhance predictive performance. It is desirable to pick a kernel matrix that captures characteristics of the data. Such kernels allow interpreting classical relatedness as genomic spatial distances, and are expected to capture some of the complexity of the genome, including non-additive effects. Smoothing of the relatedness encoded by **G** may yield better predictions under complex gene action. We cover a wide variety of kernels that are non-linear in SNPs genotype codes, but appear as linear in a regression model as seen in Equation (6).

In a Gaussian kernel, we embed individuals in the Euclidean space, and the corresponding metric is a squared Euclidean norm. For example, the spatial genetic distance between individuals (i, j) is given by

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\theta d_{ij}^2\right) = \prod_{k=1}^{m} \exp\left(-\theta \left(x_{ik} - x_{jk}\right)^2\right)$$

where θ is a positive bandwidth parameter, $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ik} - x_{jk})^2 + \cdots + (x_{im} - x_{jm})^2}$ is the Euclidean distance, and x_{ik} $(i, j = 1, \cdots, n, k = 1, \cdots, m)$ is the SNP genotype code for individual i at SNP k. The smaller the Euclidean distance is, the stronger the similarity in state between two genotype vectors. Taking exponentiation of the negative Euclidean distance changes the direction of relatedness, that is, a larger distance produces a smaller value of the Gaussian kernel and a smaller spatial genetic similarity. Gianola et al. (2006) introduced the Gaussian kernel in quantitative genetics with the aim of capturing total genetic effects in a whole-genome prediction problem. This kernel is known to be a particular case of what is called a Gaussian radial basis function (RBF). The Gaussian RBF computes spatial genetic distance between n individuals and n' individuals chosen as centroids $(n > n')$. Some applications of this in the context of genome-enabled prediction are in Long et al. (2010) and González-Camacho et al. (2012). While the idea is to choose a minimal set of basis functions, the resulting kernel is no longer semi-positive definite from the RKHS point of view. This approximation approach may help greatly when n is large. When n is chosen to be equal to n', this leads to a standard Gaussian kernel.

The exponential kernel first explored in Piepho (2009) is closely related to the Gaussian kernel. It takes the form $\exp(-\theta d_{ij})$, so the squared norm is dropped. The Matérn covariance function is a generalization of an Euclidean distance-based kernel that contains the Gaussian and the exponential kernels as particular cases. An advantage of this function is that the actual form of a kernel can be inferred from the data directly, permitting flexible kernel selection. Based on simulated data, the Gaussian kernel appeared to give the best fit within this specific class of kernel (Ober et al., 2011).

Given that SNPs take discrete values only, it seems reasonable to remove redundant areas in the Euclidean space that are never used. The diffusion kernel, which is a discretized Gaussian kernel, can be viewed as functions on discrete spaces, such as a graph. Morota et al. (2013) employed the SNP grid kernel, that is, the diffusion kernel specifically developed to model SNP data distributed on a grid graph, as described earlier. This measures

how similar two vertices are in terms of Manhattan distance on an m-dimensional grid graph. The essence of this kernel is the matrix exponentiation of a graph Laplacian. The SNP grid kernel between two vertices consisting of spatial SNP coordinates on the p-dimensional grid graph is given by

$$K_\theta(\mathbf{x}, \mathbf{x}') \propto \left(\frac{-2e^{-3\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{n_1} \left(\frac{e^{-3\theta}-3e^{-\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{n_2} \left(\frac{4e^{-3\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{m_{11}} \quad (17)$$

where n_s is the number of SNPs at which the copy of the "A" allele between two individuals differ by s, and m_{11} is the number of SNPs at which two individuals share heterozygous states. The bandwidth parameter $\theta > 0$ controls a rate of diffusion (degree of relatedness). A diffusion kernel for binary genotypes (presence or absence) is likewise constructed as

$$K_\theta(\mathbf{x}, \mathbf{x}') \propto \left(\frac{1-\exp(-2\theta)}{1+\exp(-2\theta)}\right)^{d(\mathbf{x},\mathbf{x}')} \quad (18)$$

where $d(\mathbf{x}, \mathbf{x}')$ is the Hamming distance, that is, number of coordinates at which $\mathbf{x}$ and $\mathbf{x}'$ differ. Morota et al. (2013) evaluated the performance of the diffusion and of the Gaussian kernels using dairy cattle and wheat line data. It turned out that differences in predictive ability were negligible, suggesting that the simple Gaussian kernel is very robust.

A similar attempt at refining the Gaussian kernel is in Tusell et al. (2014). These authors proposed a t kernel for m markers taking the form

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left[1 + \frac{(\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)}{m\nu}\right]^{-\frac{(\nu+m)}{2}},$$

where Σ is an $m \times m$ scale matrix and ν is the degrees of freedom (a positive continuous parameter acting as bandwidth). The authors used a diagonal matrix for Σ^{-1}, with its kth element equal to the heterozygosity at the kth SNP locus: $2p_k(1-p_k)$, and evaluated the kernel performance over a fixed grid of values of ν. This kernel aims to expand the underlying metric space with its heavier tails. The t kernel resulted in a similar performance as the Gaussian kernel (Tusell et al., 2014) suggesting that the avenues for enhancing predictive ability through kernel refinement are limited, in agreement with Morota et al. (2013). The picture that emerges here is that use of the Gaussian kernel is probably sufficient for a prediction task.

Covariance functions over the prediction grid need to be specified for interpolation when applying kriging and Gaussian process regression. In this setting, the kernel matrix $\mathbf{K}$ is the covariance matrix of a stochastic process. Here, in order to explicitly interpret a kernel as a covariance function, we assume $\mathrm{E}(g(x_{..})) = 0$. Then the covariance function becomes $\mathrm{Cov}(g(x_{..}), g(x'_{..})) = \mathrm{E}(g(x_{..}), g(x'_{..})) = K(g(x_{..}), g(x'_{..}))$, namely, a kernel.

It is worth keeping in mind that the kernel defines the the inner product, the inner product defines the covariance, and the covariance brings up a new metric called Hilbert space (distance). While some simply structured kernels such as the IBS kernel (Wessel and Schork, 2006), the weighted IBS kernel (Kwee et al., 2008; Wu et al., 2010), and the Wright-Fisher kernel (Zhu et al., 2012) have been applied for GWAS purposes in human genetic epidemiology contexts, their use in animal and plant quantitative genetics has been very limited. In general, constructing non-parametric kernel matrices is computationally more taxing than for their parametric counterparts.

3.3. KERNEL AVERAGING

Kernel methods do not preclude use of several kernels together. An alternative approach is to use "kernel averaging" or "multiple kernel learning," as proposed in de los Campos et al. (2010). Suppose there are three kernels $\mathbf{K}_1$, $\mathbf{K}_2$, and $\mathbf{K}_3$ that are distinct from each other. In this approach, the three kernels are "averaged" to form a new kernel $\mathbf{K} = \mathbf{K}_1\frac{\sigma^2_{K_1}}{\tilde{\sigma}^2_{\mathrm{K}}} + \mathbf{K}_2\frac{\sigma^2_{K_2}}{\tilde{\sigma}^2_{\mathrm{K}}} + \mathbf{K}_3\frac{\sigma^2_{K_3}}{\tilde{\sigma}^2_{\mathrm{K}}}$, where $\sigma^2_{K_1}, \sigma^2_{K_2}, \sigma^2_{K_3}$ are variance components attached to kernels $\mathbf{K}_1, \mathbf{K}_2$, and $\mathbf{K}_3$, respectively, and $\tilde{\sigma}^2_{\mathrm{K}}$ is the sum of the three variances. The ratios of the three variance components are tantamount to the relative contributions of the kernels to the marked genetic variation in the population. For instance, the kernels used can be three Gaussian kernels with different bandwidth parameter values, as employed in Tusell et al. (2014), or one can fit several parametric kernels jointly, e.g., the additive (**G**), dominance (**D**), and additive by dominance (**G#D**) kernels as in Morota et al. (2014b). While there are many possible choices for kernels, the kernel function can be estimated via maximum likelihood by recourse to the Matérn family of covariance function (e.g., Ober et al., 2011) or by fitting several candidate kernels simultaneously through multiple kernel learning.

4. APPLICATIONS OF KERNEL METHODS

A long standing question in quantitative genetics is how important epistasis is for complex traits. To a large extent, animal breeding focuses on additive variability. On the other hand, there is increasing evidence that complex traits are the product of synergistic forces spanned by a large number of genetic polymorphisms along the genome and, theoretically, functional epistasis can play an important role in selection response (Hansen, 2013). Kernel methods are theoretically appealing for accommodating cryptic forms of gene action (Gianola et al., 2006; Gianola and van Kaam, 2008).

4.1. WHOLE-GENOME PREDICTION

Kernel-based whole-genome prediction is being increasingly employed across many species. Within this class of methods, statistical models are linear in the parameters but may be non-linear in the covariates. We gather some of case studies in the published literature here, although not in great depth. The theoretical framework of RKHS regression for genome-enabled prediction first appeared in Gianola et al. (2006), and was subsequently more firmly characterized in Gianola and van Kaam (2008). González-Recio et al. (2008, 2009) reported early applications of the procedures to data on mortality and food conversion rate in broiler chickens. These authors compared RKHS regression

with various linear additive smoothers and concluded that semi-parametric methods had potential for capturing total genetic effects from real data. While same superiority of semi-parametric kernel methods over additive models has been found for body weight of broiler chickens (Long et al., 2010), differences were minimal when applied to litter size in swine (Tusell et al., 2013), dairy sires progeny test (Long et al., 2011; Morota et al., 2013), and phenotypes of dairy cows (Morota et al., 2014b). While various predictive models have not differed substantially, on average, some of the semi-parametric approaches, including RKHS, have performed consistently better.

Applications of semi-parametric regressions in plant breeding have also produced encouraging results. The semi-parametric approach was evaluated in plant breeding by Crossa et al. (2010) who compared RKHS with Bayesian LASSO using CIMMYT data sets. Therein, 599 wheat lines genotypes with 1447 markers and 300 tropical maize lines genotyped with 1148 SNPs were analyzed for this purpose. While RKHS and Bayesian LASSO showed a similar predictive ability in maize, the former outperformed the latter in the wheat lines. This wheat data set has been used in many subsequent studies, and the apparent advantage of RKHS regression was confirmed (de los Campos et al., 2010; Endelman, 2011; Long et al., 2011; Heslot et al., 2012; Morota et al., 2013; Tusell et al., 2014). Heslot et al. (2012) performed an extensive comparison of prediction methods including RKHS, neural networks, support vector machines, ridge-regression BLUP, and random forests. RKHS topped 15 out of 18 trait-comparisons and performed the best on average in terms of predictive correlations. González-Camacho et al. (2012) compared RKHS, radial basis function neural networks (RBFNN) and the Bayesian LASSO using 21 trait-environment combinations of maize data sets. Although the differences were small, RKHS had a consistently higher predictive correlation than RBFNN and the Bayesian LASSO. Subsets of these maize data were reanalyzed in another exhaustive comparison carried out by Ornella et al. (2014). Among six regression methods, RKHS ranked as the best prediction machine when applied to 14 maize data sets and across trait-environment combinations as measured by Pearson's correlation coefficient. Echoing work of González-Camacho et al. (2012), Pérez-Rodríguez et al. (2012) tested RKHS, RBFNN, Bayesian regularized neural networks, and linear additive smoothers, including Bayesian LASSO, Bayesian ridge regression, BayesA, and BayesB. BayesA and BayesB are marker-based Bayesian hierarchical linear regression models developed to capture additive genetic effects of markers (Meuwissen et al., 2001). The data used were 306 elite wheat lines coupled with 1717 diversity array technology (DArT) markers. The authors observed a consistent superiority of non-additive smoothers over additive counterparts, yet, RKHS and RBFNN were equally competitive. Kernel-based whole-genome prediction models have also been applied to genotyping-by-sequencing on maize populations (Crossa et al., 2013). Overall, RKHS performed slightly better than GBLUP when genomic information was the sole source of information used for predictors. A study led by Sun et al. (2012) reported that RKHS and smoothing spline ANOVA model (alternative parametrization of Bayesian kernel ridge regression) coupled with supervised principal component analysis delivered slightly higher predictive correlations in barley and maize data. On the other hand, no clear difference was observed between RR-BLUP and RKHS, which led the authors to conjecture that there appear to be no epistatic effects on six maize traits studied (Riedelsheimer et al., 2012b). In retrospect, RKHS is at least as good as linear additive smoothers: on one hand it delivers better predictive performance when non-additive effects are present and, on the other hand, produces a similar performance when additivity is the main source of genetic variation.

4.2. ASSESSMENT OF PREDICTIVE PERFORMANCE

Avoiding overfitting of prediction models is desirable because one wishes to extract genetic signal but not noise. Since its introduction into animal breeding by Meuwissen et al. (2001), cross-validation (CV) has quickly become the technique of choice for measuring prediction performance. It is being widely used in whole-genome prediction and there are a number of ways of applying it. For instance, CV designs include: (1) two-generation (stratification by generation or date of birth) validation, (2) k-fold validation, and (3) repeated random sub-sampling validation. We highlight here pros and cons of each. The two-generation scheme is a reasonable choice to use when a data set comprises parents and their offspring. The records on parents are used to train the model and prediction is carried out in a testing set that includes offspring of individuals in the training set. While this setting can simulate a standard genetic evaluation scenario and prevent an unrealistic case such as predicting parent responses from offspring records, it generates only a single realized prediction accuracy measurement, and an alternative is to bootstrap the layout. In k-fold validation scheme, the entire data set is first randomly splitted into k disjoint subsets of equal size. Within each fold, k-1 subsets are used to fit the model and the remaining subset is used as a testing set to predict masked phenotypes of individuals. This is repeated until all k subsets are used as testing and results from the k-fold are averaged. Typically, $k = 5$ or 10 is used to assess predictive ability. Unlike two-generation validation, an advantage of this approach is that it is possible to estimate CV uncertainty but at the cost of higher computation load. Recent research (Makowsky et al., 2011; Saatchi et al., 2011; Pérez-Cabal et al., 2012; Kramer et al., 2014) have shown that the k-fold CV gives a slightly better predictive performance than a two-generation validation. In practice, the design of CV must keep the target problem in mind. Lastly, repeated random sub-sampling validation randomly partitions the data set into training (e.g., 90%) and testings (e.g., 10%) sets. Then, an average of say, 50 random repeats of the cross-validation is computed, and this CV layout has been adopted in some past studies (e.g., González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012; Gianola et al., 2014b). While it also permits to obtain a CV distribution, some individuals may never enter into training and testing sets or appear more often than others. While these three CV strategies are widely adopted in practice, it is important to note that the ultimate goal is to successfully predict in cross-study validation using independent data.

Prediction performance on CV in the testing set can be assessed by Pearson's correlation or via the predictive mean-squared error between observed and predicted values. A squared predictive correlation aims to measure the amount of variation

captured by prediction. However, prediction assessment is not limited to these two measures; for example, Ornella et al. (2014) used Cohen's kappa coefficient (Cohen, 1960) and found that it provided similar indicators of the performance of genome-enabled prediction when selecting the best individuals. On the other hand, the behavior of prediction machines differed when the aim was to separate the best and worst individuals, e.g., selecting the best 15% of individuals. This suggests that optimal prediction methods that perform well when predicting individuals in the testing set overall, may result in a poor performance when the target is only the tails of the distribution in the same testing set (Ornella et al., 2014).

Assessment of predictive performance is not exclusively limited to use of CV. Stone (1977) showed that leave-one-out CV and Akaike's information criterion (AIC) are asymptotically equivalent when maximum likelihood estimation is used. An underappreciated point is that Akaike developed AIC to quantify goodness-of-prediction rather than goodness-of-fit (Akaike, 1974). By leveraging this property, it is possible to evaluate predictive ability of models without conducting computationally expensive CV. The prediction model associated with the smallest AIC is considered to fit the data well from a predictive point of view. Note that the proof in Stone (1977) is not restricted to linear models. Piepho (2009) assessed predictive ability of several models using AIC, and the study carried out by Schulz-Streeck and Piepho (2010) found congruence between AIC and the predictive correlation (the smaller the AIC values, the higher the predictive correlations). That said, this pattern did not always hold, suggesting that the CV remains as a viable tool to gauge predictive performance.

Another avenue is to derive deterministic equations that compute expected predictive correlation on the basis of a number of factors. These include size of the training data set, the effective number of independent chromosome segments, the number of markers used, and heritability of the trait (Daetwyler et al., 2008, 2010; Goddard, 2009; Goddard et al., 2011; Erbe et al., 2013). In relation to these, de los Campos et al. (2013b) gave a theoretical upper limit for the achievable predictive ability that does not use number of independent chromosome segments. This is appealing because the number of independently segregating segments is a somewhat idealized concept and not straightforward to estimate. It may be argued that it is worthwhile to employ AIC or deterministic equations prior to performing resource intensive CV, to see what can be expected empirically.

4.3. GWAS-BASED PREDICTION

The first large-scale genome-wide association study (GWAS) involving hundreds of thousands of SNPs was reported by Ozaki et al. (2002). The authors identified some functional variants associated with myocardial infarction using 92,788 genotyped SNPs in more than 2000 individuals. Since then, a number of GWAS results have been reported (Visscher et al., 2012), and the success of GWAS seems largely due to the biochemical technology that generates high dimensional markers spanning the entire genome, rather than the statistical methodologies that have been employed.

Genome-enabled prediction of traits outside of animal and plant domains has been mainly carried out by targeting specific genes or variants identified from inference procedures (GWAS or their meta-analyses) as opposed to using all available markers. This indicates that deciding which genetic variants to include in the prediction equation is a crucial component. In this line, Morota et al. (2014a) classified SNPs based on functional annotation and created kernels for each of several genomic regions. They observed that functionally annotated regions did not always deliver a better predictive performance than intergenic regions in three broiler traits analyzed. A whole-genome regression model incorporating all available quality filtered SNPs attained a similar performance to that from the genomic region (either genic or intergenic) that achieved the best prediction. Likewise, results in Holstein cattle obtained by Erbe et al. (2012) found that the predictive performance of the whole ensemble of SNPs was comparable to that of markers in exonic regions. While the whole-genome prediction approach appears to be adequate for practical purposes, to what extent preselection of SNPs residing in functionally enriched regions aids predictive ability is a subject for further investigation. One plausible explanation for the observed high predictive performance of intergenic regions is in Schierding et al. (2014). Recall that all chromosomes are folded such that certain genic regions (e.g., exons) may physically interact with distantly located gene deserts in the sense of a 3-dimensional space. This creates spatial associations and may allow intergenic regions to regulate gene function from far away on a linear scale.

Most GWAS rely on p-values derived from a series of single marker regressions or with inclusion of a genomic relationship matrix to correct for false positive associations due to population structure or relatedness. A new approach, kernel-based GWAS, appears to be emerging. For example, Maity et al. (2012) reported that a multivariate kernel GWAS can potentially accommodate interaction and non-linear effects among markers. However, a pitfall of the two-steps approach (preselection of predictors) based on p-values has been studied by Lazzeroni et al. (2014), leading to the conclusion "While uncertainty is high for a p-value from a single test, p-values obtained from GWAS, or other multiplexed studies requiring multiple testing corrections, provide almost no information with which to make future predictions." due to the high variability of p-values. This suggests that if associated variants are simply picked by non-replicable p-values in GWAS, the resulting prediction step will fail (Wray et al., 2007). Currently, GWAS in animals is predominantly carried out as a by-product of genome-enabled prediction, in which the degree of association is estimated from marker effect sizes. While an ultimate goal is to dissect genetic architecture and to make predictions from validated genetic variants, whether knowledge from "inference" can aid prediction is yet to be answered. Arguably, the future may reside on inference, but there is plenty of room for connecting inference and prediction. Additional criteria need to supplement the use of p-values (Malley et al., 2013). Perhaps, there is a critical need to adopt more CV in GWAS. Significance of detected variants in one data set should be reaffirmed in a separate data set. This is because the fraction of genetic variance explained by QTL or by a subset of markers in a testing set is much less than

that of the training set (e.g., Utz et al., 2000; Makowsky et al., 2011; Würschum and Kraft, 2014).

5. CONCLUSIONS

We reviewed the utility of kernel methods as a tool of choice for a prediction of yet-to-be observed phenotypes, and described their relation to spatial variation as determined by a vector of SNPs. Research aiming to extend Fisher's infinitesimal model to accommodate non-additive effects either parametrically or non-parametrically is a topic of interest. Most studies carried out so far suggest that whole-genome prediction coupled with combinations of kernels may capture non-additive variation (Gianola et al., 2014a; Howard et al., 2014). These approaches are also applicable to whole-genome risk prediction by introducing the concept of liability (Wright, 1934; Falconer, 1965; Gianola, 1982). In many applications, accurate prediction of individual responses rather than of breeding values is a primal interest.

Many challenges remain, however. We conclude by highlighting some potential future directions in kernel-based whole genome prediction methods. Throughput of genomics, transcriptomic and proteomic technologies has advanced tremendously in recent years. Arguably, emerging molecular information makes quantitative genetics more powerful because it permits to expand a theory that was initially largely built around pedigree information during the last century. It could be argued that 21st century quantitative genetics needs to be linked to functional genomics, and not simply relying on hypothetical QTL that may or may not exist.

In addition, more study on methodology specifically tailored for emerging omics data that combines statistical approaches and functional validation is required. For example, the RKHS method accommodates any information set for input variables. This paves the way to link phenotype and genome jointly with intermediate phenotypes such as transcriptomic, proteomic and metabolomic data in a single framework, known as systems genetics (e.g., Civelek and Lusis, 2013). Recent attempts along this line pertinent to genome-enabled prediction include joint evaluations of genome and transcriptome (Bhattacharjee and Sillanpää, 2011) and genome and metabolome (Riedelsheimer et al., 2012a). The question boils down to how we condense and construct kernels from each set of biological information. Also, estimation of non-additive variation in the eQTL context has been explored recently (Powell et al., 2013; Hemani et al., 2014).

Prediction of response variables is largely classified into predicting: (a) additive effects, (b) total genetic effects, and (c) raw phenotypes. The first type is mainstream in genome-enabled selection schemes, and we have given an overview of kernel methods in which their predominant aim is to capture total genetic effects. Although currently receiving less attention, predicting raw phenotypes is especially vital for assessing health or medical outcomes. Arguably, many non-genetic factors affect raw phenotypes, and these cannot be well predicted without environmental information, so imperfect information on environmental variables hinders achieving greater predictive performance (Heslot et al., 2014). One important direction for future study is that of condensing environmental variables into some "environmental kernel," to cast a joint evaluation of genotype and environment via kernel methodology (Jarquín et al., 2014). Research involving analysis of raw phenotypes coupled with environmental variables needs more attention, and even when genetic selection is the sole interest, the use of pre-processed quasi-phenotypes (e.g., estimated breeding value) should be avoided, if possible, as reported by Ekine et al. (2014).

AUTHOR CONTRIBUTIONS

Gota Morota and Daniel Gianola assembled relevant literature and wrote the manuscript.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. New York, NY: Springer.

Bhattacharjee, M., and Sillanpää, M. J. (2011). A bayesian mixed regression based prediction of quantitative traits from molecular marker and gene expression data. *PLoS ONE* 6:e26959. doi: 10.1371/journal.pone.0026959

Calus, M. P. (2010). Genomic breeding value prediction: methods and procedures. *Animal* 4, 157–164. doi: 10.1017/S1751731109991352

Civelek, M., and Lusis, A. J. (2013). Systems genetics approaches to understand complex traits. *Nat. Genet.* 15, 34–48. doi: 10.1038/nrg3575

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Measure.* 20, 37–46. doi: 10.1177/001316446002000104

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Pérez, P., Hickey, J. M., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding program. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS ONE* 9:e87666. doi: 10.1371/journal.pone.0087666

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3:e3395. doi: 10.1371/journal.pone.0003395

de los Campos, G., Gianola, D., and Rosa, G. J. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci.* 87, 1883–1887. doi: 10.2527/jas.2008-1259

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res. (Camb.)* 92, 295–308. doi: 10.1017/S0016672310000285

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013a). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013b). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608

Ekine, C. C., Rowe, S. J., Bishop, S. C., and de Koning, D.-J. (2014). Why breeding values estimated using familial data should not be used for genome-wide association studies. *G3* 4, 341–347. doi: 10.1534/g3.113.008706

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Erbe, M., Gredler, B., Seefried, F. R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the

average accuracy of genomic prediction. *PLoS ONE* 8:e81046. doi: 10.1371/journal.pone.0081046

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., et al. (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. doi: 10.3168/jds.2011-5019

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* 29, 51–76. doi: 10.1111/j.1469-1809.1965.tb00500.x

Fernando, R. L., and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* 21, 467–477. doi: 10.1186/1297-9686-21-4-467

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edn.* 52, 399–433. doi: 10.1017/S0080456800012163

Gianola, D. (1982). Theory and analysis of threshold characters. *J. Anim. Sci.* 54, 1079–1096.

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

Gianola, D., de los Campos, G., González-Recio, O., Long, N., Okut, H., Rosa, G. J. M., et al. (2010). "Statistical learning methods for genome-based analysis of quantitative traits," in *Proceedings of The 9th World Congress on Genetics Applied to Livestock Production* (Leipzig: CD–ROM Communication 0014).

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363. doi: 10.1534/genetics.109.103952

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gianola, D., Morota, G., and Crossa, J. (2014a). "Genome-enabled prediction of complex traits with kernel methods: What have we learned?," in *Proceedings, 10th World Congress of Genetics Applied to Livestock Production* (Vancouver, BC).

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

Gianola, D., and van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Gianola, D., Weigel, K. A., Stella, N. K. A., and Schön, C. C. (2014b). Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS ONE* 9:e91693. doi: 10.1371/journal.pone.0091693

Goddard, M. E. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed Genet.* 128, 409–421. doi: 10.1111/j.1439-0388.2011.00964.x

González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M., and no, S. A. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 178, 2305–2313. doi: 10.1534/genetics.107.084293

González-Recio, O., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol.* 41, 3. doi: 10.1186/1297-9686-41-3

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194, 597–607. doi: 10.1534/genetics.113.152207

Hansen, T. F. (2013). Why epistasis is important for selection and adaptation. *Evolution* 67, 3501–3511. doi: 10.1111/evo.12214

Harville, D. A. (1983). *Discussion on A Section on Interpolation and Estimation. in Statistics an Appraisal*. Ames, IA: The Iowa State University Press.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., et al. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature* 508, 249–253. doi: 10.1038/nature13005

Henderson, C. R. (1973). "Sire evaluation and genetic trends," in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr Jay. L. Lush*, (Champaign, IL: American Society of Animal Science and American Dairy Science Association), 10–41.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430

Henderson, C. R. (1977). Best linear unbiased prediction of breeding values not in the model for records. *J. Dairy Sci.* 60, 783–787. doi: 10.3168/jds.S0022-0302(77)83935-0

Henderson, C. R. (1985). Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *J. Anim. Sci.* 60, 111–117.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480. doi: 10.1007/s00122-013-2231-5

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop. Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)* 4, 1027–1046. doi: 10.1534/g3.114.010298

Isaaks, E. H., and Srivastava, R. M. (1990). *An Introduction to Applied Geostatistics*. New York, NY: Oxford University Press.

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P. D., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi: 10.1007/s00122-013-2243-1

Kimeldorf, G., and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* 33, 82–95. doi: 10.1016/0022-247X(71)90184-3

Kramer, M., Erbe, M., Seefried, F. R., Gredler, B., Bapst, B., Bieber, A., et al. (2014). Accuracy of direct genomic values for functional traits in Brown Swiss cattle. *J. Dairy Sci.* 97, 1774–1781. doi: 10.3168/jds.2013-7054

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.* 82, 386–397. doi: 10.1016/j.ajhg.2007.10.010

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756.

Lazzeroni, L., Lu, Y., and Belitskaya-Lévy, I. (2014). *P*-values in genomics: apparent precision masks high uncertainty. *Mol. Psychiatry*. doi: 10.1038/mp.2013.184. [Epub ahead of print].

Long, N., Gianola, D., Rosa, G. J., and Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123, 1065–1074. doi: 10.1007/s00122-011-1648-y

Long, N., Gianola, D., Rosa, G. J., Weigel, K. A., Kranis, A., and González-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genet. Res.* 92, 209–225. doi: 10.1017/S0016672310000157

Maity, A., Sullivan, P. F., and ing Tzeng, J. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.* 36, 686–695. doi: 10.1002/gepi.21663

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., et al. (2011). Beyond missing heritability: prediction of complex traits. *PLoS Genet.* 7:e1002051. doi: 10.1371/journal.pgen.1002051

Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Paris: Masson et Cie.

Malley, J. D., Dasgupta, A., and Moore, J. H. (2013). The limits of *p*-values for biological data mining. *BioData Min.* 6:10. doi: 10.1186/1756-0381-6-10

Massman, J. M., Gordillo, A., Lorenzana, R. E., and Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theor. Appl. Genet.* 126, 13–22. doi: 10.1007/s00122-012-1955-y

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

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2013). Accelerating improvement of livestock with genomic selection. *Annu. Rev. Genet.* 1, 221–237. doi: 10.1146/annurev-animal-031412-103705

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014a). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics* 15:109. doi: 10.1186/1471-2164-15-109

Morota, G., Boddhireddy, P., Vukasinovic, N., Gianola, D., and DeNise, S. (2014b). 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. doi: 10.3389/fgene.2014.00056

Morota, G., Koyama, M., Rosa, G. J. M., Weigel, K. A., and Gianola, D. (2013). Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet. Sel. Evol.* 45, 17. doi: 10.1186/1297-9686-45-17

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., and Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41, 56. doi: 10.1186/1297-9686-41-56

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* 75, 1738–1745.

Ober, U., Erbe, M., Long, N., Porcu, E., H, M. S., and Simianer (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics* 188, 695–708. doi: 10.1534/genetics.111.128694

Onaga, L. (2010). Toyama Kametaro and Vernon Kellogg: silkworm inheritance experiments in Japan, Siam, and the United States, 1900-1912. *J. Hist. Biol.* 43, 215–264. doi: 10.1007/s10739-010-9222-z

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J. M., Burgueño, J., Zhang, X., et al. (2014). Genomic-enabled prediction with classification algorithms. *Heredity* 112, 616–626. doi: 10.1038/hdy.2013.144

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., et al. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* 32, 650–654. doi: 10.1038/ng1047

Pérez-Cabal, M. A., Vazquez, A. I., Gianola, D., Rosa, G. J., and Weigel, K. A. (2012). Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front. Genet.* 3:27. doi: 10.3389/fgene.2012.00027

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., and Dreisiqacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3* 2, 1595–1605. doi: 10.1534/g3.112.003665

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* 49, 1165–1176. doi: 10.2135/cropsci2008.10.0595

Plutynski, A. (2006). What was Fisher's fundamental theorem of natural selection and what was it for? *Stud. Hist. Philos. Biol. Biomed. Sci.* 37, 59–82. doi: 10.1016/j.shpsc.2005.12.004

Powell, J. E., Henders, A. K., McRae, A. F., Kim, J., Hemani, G., Martin, N. G., et al. (2013). Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet.* 9:e1003502. doi: 10.1371/journal.pgen.1003502

Rasmussen, C. E., and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012a). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44, 217–220. doi: 10.1038/ng.1033

Riedelsheimer, C., Technow, F., and Melchinger, A. E. (2012b). Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* 13:452. doi: 10.1186/1471-2164-13-452

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* 6, 15–51. doi: 10.1214/ss/1177011926

Rodríguez-Ramilo, S. T., García-Cortés, L. A., and González-Recio, O. (2014). Combining genomic and genealogical information in a reproducing kernel hilbert spaces regression model for genome-enabled predictions in dairy cattle. *PLoS ONE* 9:e93424. doi: 10.1371/journal.pone.0093424

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511755453

Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., et al. (2011). Accuracies of genomic breeding values in american angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol.* 43, 40. doi: 10.1186/1297-9686-43-40

Schierding, W., Cutfield, W. S., and O'Sullivan, J. M. (2014). The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell. *Front. Genet.* 5:39. doi: 10.3389/fgene.2014.00039

Schulz-Streeck, T., and Piepho, H. P. (2010). Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proc.* 4:S8. doi: 10.1186/1753-6561-4-S1-S8

Stein, M. L. (1999). *Interpolation of Spatial Data*. New York, NY: Springer. doi: 10.1007/978-1-4612-1494-6

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. Ser. B* 39, 44–47.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS ONE* 7:e45293. doi: 10.1371/journal.pone.0045293

Sun, X., Ma, P., and Mumm, R. H. (2012). Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. *PLoS ONE* 7:e50604. doi: 10.1371/journal.pone.0050604

Toro, M. A., García-Cortés, L. A., and Legarra, A. (2011). A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet. Sel. Evol.* 43, 27. doi: 10.1186/1297-9686-43-27

Tusell, L., Pérez-Rodríguez, P., Forni, S., and Gianola, D. (2014). Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. *J. Anim. Breed Genet.* 131, 105–115. doi: 10.1111/jbg.12070

Tusell, L., Pérez-Rodríguez, P., Wu, S. F. X.-L., and Gianola, D. (2013). Genome-enabled methods for predicting litter size in pigs: a comparison. *Animal* 7, 1739–1749. doi: 10.1017/S1751731113001389

Utz, H. F., Melchinger, A. E., and Schon, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154, 1839–1849.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

VanRaden, P. M., Tassell, C. P. V., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92, 16–24. doi: 10.3168/jds.2008-1514

Villanueva, B., Pong-Wong, R., Fernández, J., and Toro, M. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* 83, 1747–1752.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *Am. J. Hum. Genet.* 9, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi: 10.1534/genetics.113.155176

Wessel, J., and Schork, N. J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* 79, 792–806. doi: 10.1086/508346

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., et al. (2014). Poly-omic prediction of complex traits: omickriging. *Genet. Epidemiol.* 38, 402–415. doi: 10.1002/gepi.21808

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* 17, 1520–1528. doi: 10.1101/gr.6665407

Wright, S. (1921). Systems of mating. I. The biometric relations between offspring and parent. *Genetics* 6, 111–123.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* 56, 330–338. doi: 10.1086/279872

Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics* 19, 506–536.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanockand, S. J., Hunter, D. J., et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* 86, 929–942. doi: 10.1016/j.ajhg.2010.05.002

Würschum, T., and Kraft, T. (2014). Cross-validation in association mapping and its relevance for the estimation of QTL parameters of complex traits. *Heredity* 112, 463–468. doi: 10.1038/hdy.2013.126

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yule, G. U. (1902). Mendel's laws and their probable relation to intra-racial heredity. *New Phytol.* 1, 192–207, 222–238. doi: 10.1111/j.1469-8137.1902.tb07336.x

Zhong, S., Dekkers, J. C. M., Fernando, R. L., and Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182, 355–364. doi: 10.1534/genetics.108.098277

Zhu, H., Li, L., and Zhou, H. (2012). Nonlinear dimension reduction with WrightFisher kernel for genotype aggregation and association mapping. *Bioinformatics* 28, i375–i381. doi: 10.1093/bioinformatics/bts406

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 August 2014; paper pending published: 01 September 2014; accepted: 29 September 2014; published online: 16 October 2014.

*Citation: Morota G and Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. **5**:363. doi: 10.3389/fgene.2014.00363*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.