Locally Epistatic Genomic Relationship Matrices for Genomic Association, Prediction

Deniz Akdemir & Jean-Luc Jannink
Department of Plant Breeding & Genetics
Cornell University
Ithaca, NY

January 17, 2015
Running Head: Locally Epistatic Models

Key Words: Genomic Selection, Epistasis, Mixed Models

Corresponding Author:

Deniz Akdemir

232 Emerson Hall
Plant Breeding & Genetics
Cornell University
Ithaca, NY 14853
(419) 308 7085 (ph.)
da346@cornell.edu
Abstract

In plant and animal breeding studies a distinction is made between the genetic value (additive + epistatic genetic effects) and the breeding value (additive genetic effects) of an individual since it is expected that some of the epistatic genetic effects will be lost due to recombination. In this paper, we argue that the breeder can take advantage of the epistatic marker effects in regions of low recombination. The models introduced here aim to estimate local epistatic line heritability by using genetic map information and combining local additive and epistatic effects. To this end, we have used semi-parametric mixed models with multiple local genomic relationship matrices with hierarchical designs. Elastic-net post-processing was used to introduce sparsity. Our models produce good predictive performance along with useful explanatory information.
Selection in animal or plant breeding is usually based on estimates of genetic breeding values (GEBV) obtained with semi-parametric mixed models (SPMM) (Hayes et al. (2001), Lee et al. (2008)). In a mixed model, genetic information in the form of a pedigree or markers is used to construct an additive kernel matrix that describes the similarity of line specific additive genetic effects. These models have been successfully used for predicting the breeding values in plants and animals. Studies show that using similarities calculated from sufficient genome-wide marker information almost always lead to better prediction models for the breeding values compared to the pedigree based models (Hayes et al. (2001), Hayes et al. (2009), Habier et al. (2007)). In both simulation studies and in empirical studies of dairy cattle (Hayes et al. (2009), VanRaden et al. (2009)), mice (Lee et al. (2008), Legarra et al. (2008)) and in bi-parental populations of maize, barley and Arabidopsis (Lorenzana and Bernardo (2009), Lorenzana and Bernardo (2009), Heffner et al. (2011)) marker based SPMM GEBVs have been quite accurate.

A SPMM for the $n \times 1$ response vector $y$ is expressed as

$$y = X\beta + Zg + e$$  \hspace{1cm} (1)

where $X$ is the $n \times p$ design matrix for the fixed effects, $\beta$ is a $p \times 1$ vector of fixed effect coefficients, $Z$ is the $n \times q$ design matrix for the random effects; the vector random effects $(g', e')'$ is assumed to follow a multivariate normal distribution with mean 0 and covariance

$$\begin{pmatrix}
\sigma_g^2 K & 0 \\
0 & \sigma_e^2 I_n
\end{pmatrix}$$

where $K$ is a $q \times q$ kernel matrix.

The similarity of the kernel based SPMMs and reproducing kernel Hilbert spaces (RKHS) regression models has been stressed recently (Gianola and Van Kaam (2008)). In fact, this connection was previously recognized by Kimeldorf and Wahba (1970), Harville (1977),
RKHS regression models use an implicit or explicit mapping of the input data into a high dimensional feature space defined by a kernel function. This is often called the "kernel trick" (Schölkopf and Smola (2002)).

A kernel function, \( k(., .) \) maps a pair of input points \( x \) and \( x' \) into real numbers. It is by definition symmetric (\( k(x, x') = k(x', x) \)) and non-negative. Given the inputs for the \( n \) individuals we can compute a kernel matrix \( K \) whose entries are \( K_{ij} = k(x_i, x_j) \). The linear kernel function is given by \( k(x; y) = x'y \). The polynomial kernel function is given by \( k(x; y) = (x'y + c)^d \) for \( c \) and \( d \in \mathbb{R} \). Finally, the Gaussian kernel function is given by \( k(x; y) = \exp(-h(x' - y)(x' - y)) \) where \( h > 0 \).

RKHS regression extends SPMMs by allowing a variety of kernel matrices, not necessarily additive in the input variables, calculated using a variety of kernel functions. Some common choices are the linear, polynomial, and Gaussian kernel functions, though many other options are available (Schölkopf and Smola (2002), Endelman (2011)).

For the marker-based SPMMs, a genetic kernel matrix calculated using a linear kernel matrix incorporates only additive effects of markers. A genetic kernel matrix based on the polynomial kernel of order \( k \) incorporates all of the one to \( k \) order monomials of markers in an additive fashion. The Gaussian kernel function allows us to incorporate additive and complex epistatic effects of the markers implicitly.

Simulation studies, and results from empirical experiments show that the prediction accuracies of models with Gaussian are often higher than the models with linear kernel (De Los Campos (2008), González-Camacho et al. (2012), Heslot et al. (2012)). However, it is not possible to know how much of the increase in accuracy can be transferred to subsequent generations because some of the predicted epistatic effects will be lost by recombination. This is related to the distinction made between the commercial value of a line (defined as the overall genetic effect, additive+non-additive) and the breeding value (the potential for being a good parent, additive only). It can be argued that the linear kernel model estimates the breeding value whereas the Gaussian kernel model estimates the genetic value. In this article, we argue that the breeder
can take advantage of some epistatic marker effects in regions of low recombination. Epistatic interactions that span short map segments (i.e., about 20 cM) are considered "local". The models introduced here aim to estimate local epistatic line heritability by using genetic map information and combining the local additive and epistatic effects. Since only local epistatic effects are used, there is a reduced chance that these effects will disappear with recombination.

The final models we propose are SPMMs with semi-supervised kernel matrices that are obtained as a weighted sum of functions of many local kernels. The principal aim of this article is to measure and incorporate additive and local epistatic genetic contributions since we believe that the local epistatic effects are relevant to the breeder. Locally epistatic models in this article can be adjusted so that the genetic contribution of the whole genome, the chromosomes, or local regions can be obtained.

In most genome-wide association studies (GWAS) the focus is on estimating the effects of each marker and lower level interactions (CANTOR et al. (2010)). However, the number of SNP markers can easily exceed millions. The methods used in GWAS lack statistical power, and they are computationally exhaustive. The local kernel approach developed in this article remedies these problems by reducing the number of hypothesis tests by focusing on regions.

Another argument for focusing on short segments of the genome as distinct structures comes from the "building blocks" hypothesis in the evolutionary theory. The schema theorem of Holland (HOLLAND (1975)) predicts that a complex system that uses evolutionary mechanisms such as fitness, recombination and mutation tend to generate short and well fit and specialized structures. These basic structures then serve as building blocks. For example, when the alleles associated with an important fitness trait are scattered all around the genome the favorable effects can be lost by independent segregation. Therefore, inversions that group these alleles physically together would be selected.

The sum of the "building blocks" approach we propose in this paper is parsimonious since only a few genomic regions are used in the final model. In addition, importance scores for genomic regions are obtained as a byproduct.
The rest of the paper is organized as follows: In the next section, after briefly reviewing some multiple kernel approaches from the statistics and machine learning literature, we introduce our model that is more suitable to use in the context of traditional SPMMs. We discuss the issues of model set-up, parameter estimation, and hypothesis testing here. We illustrate our model with four benchmark datasets and simulations. We conclude with a section that includes a summary of main findings and discussions.

MATERIALS AND METHODS

**Multiple kernel learning** In recent years, several methods have been proposed to combine and use multiple kernel matrices instead of using a single one. These kernel matrices may correspond to using different notions of similarity or to using information coming from multiple sources.

Some early literature related to use of multiple kernels simultaneously included [HARTLEY and RAO (1967), RAO (1971)] and more recently [BACH et al. (2004) and SONNENBURG et al. (2006)]. Their use in the context of genetic information and mixed models also gained attention ([DE LOS CAMPOS (2008), DE LOS CAMPOS et al. (2010), JARQUÍN et al. (2013) and TUSELL et al. (2014)]).

Multiple kernel learning methods use multiple kernels by combining them into a single one via a combination function. The most commonly used combination function is linear. Given kernels $K_1, K_2, \ldots, K_p$, a linear kernel is of the form

$$K = \eta_1 K_1 + \eta_2 K_2 + \ldots + \eta_p K_p.$$  

The kernel $K$ can also include interaction terms like $K_i \odot K_j$, $K_i \otimes K_j$, or perhaps $-(K_i - K_j) \odot (K_j - K_i)$ where the $\odot$ is the element-wise matrix multiplication operator and the $\otimes$ is the matrix Kronecker product operator. For example, if $K_E$ is the environment kernel matrix and $K_G$ is the genetic kernel matrix, then a component $K_E \odot K_G$ can be used to capture the gene by environment interaction effects.

The kernel weights $\eta_1, \eta_2, \ldots, \eta_p$ are usually assumed to be positive, and this corresponds to
calculating a kernel in the combined feature spaces of the individual kernels. Therefore, once given the kernels, multiple kernel learning boils down to estimating the kernel weights.

**A Locally Epistatic Genomic Model for Genomic Association, Prediction** Our model building approach has three stages:

1. Subsets of the genome: Divide the marker set into $k$ subsets.

2. Local genetic values (GEVs): Use the training data to obtain a model to estimate the local GEVs $\hat{g}^*_{j}(m)$ for each genome region $j = 1, 2, \ldots, k$.

3. Post-processing: Combine the local GEVs using an additive model fitted in the training data set.

In the rest of this section, we will describe each step in more detail.

**Locally epistatic kernels from mapped marker data** In order to obtain $k$ kernels for a marker data, we will need $k$ possibly nested or overlapping subsets of the marker set. These subsets can be obtained using any annotation of the markers. However, since our aim is to capture the additive + locally epistatic genetic effects in the model, we will concentrate only on contiguous though possibly nested and overlapping regions of the genome.

Although it is possible to define genomic regions in an informed fashion (for example see [Xu (2013)] for a division based on recombination hotspots), in our illustrations we will carry out this task hierarchically as illustrated for a hypothetical organism with three chromosomes in Figure 1. In this figure, at the root of the hierarchy we have the whole genome; the second level of the hierarchy divides the genome into chromosomes. The third and the following levels of the hierarchy are obtained in an iterative fashion by splitting each of the partitions of the previous level into a specified number (called "nsplit") of "roughly equal sized" non-overlapping sets of
consecutive markers. For both the simulated and real data, we have used the map positions to order the markers, then, the splits were done such that the partitions at a certain level of splits on a chromosome had approximately the same number of markers. The splitting is stopped at a pre-specified level that is called "depth". The "nsplit" and the "depth" become the hyperparameters of the model. The section on hyperparameter selection and the illustrations herein give some insight into how these parameters can be adjusted.

Multiple kernel SPMMs Some multiple kernel approaches use fixed weights for combining kernels, however, usually the weight parameters need to be learned from the training data. Some principled techniques used to estimate these parameters include minimum norm quadratic unbiased estimation (Rao (1971)), variance least squares approach (Amemiya (1977), Demidenko (2004), p. 223)) or the Bayesian approaches implemented in the R package BGLR (Rodriguez and de los Campos (2012)). However, these methods are more suitable for cases where only a few kernels are being used because they fail to give satisfactory solutions in high-dimensional settings, i.e., when the number of kernels is large. In the remaining of this section, we develop a model that is more suitable for use in high-dimensional settings.

To obtain the local GEVs, one possible approach is to use a SPMM with multiple kernels in the form of

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g}_1 + \mathbf{Z}\mathbf{g}_2 + \ldots + \mathbf{Z}\mathbf{g}_k + \mathbf{e}$$  (2)

where $\mathbf{g}_j \sim N_{qk}(0, \sigma^2_{g_j} \mathbf{K}_j)$ for $j = 1, 2, \ldots, k$, $\mathbf{e} \sim N_n(0, \sigma^2_e \mathbf{I})$ and $\mathbf{g}_1, \ldots, \mathbf{g}_k$, $\mathbf{e}$ are mutually independent.

Another SPMM incorporates the marginal variance contribution from each kernel matrix:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g}_j + \mathbf{Z}\mathbf{g}_{-j} + \mathbf{e}_j$$  (3)

where $\mathbf{g}_j \sim N_{qk}(0, \sigma^2_{g_j} \mathbf{K}_j)$ for $j = 1, 2, \ldots, k$, $\mathbf{g}_{-j} \sim N_{qk}(0, \sigma^2_{g_{-j}} \mathbf{K}_{-j})$ is the random effect corresponding to the input components other than the ones in group $j$ and $\mathbf{K}_{-j}$ stands for the
kernel matrix obtained from markers not in group $j$. $e_j \sim N_n(0, \sigma_{e_j}^2 I)$ and $g_j, g_{-j}, e_j$ are mutually independent.

A simpler approach is to use a separate SPMM for each kernel. Let $\hat{\sigma}_{g_j}^2$ and $\hat{\sigma}_{e_j}^2$ be the estimated variance components from the SPMM model in (1) with kernel $K = K_j$. The markers corresponding to the random effect $g_{-j}$ that mainly accounts for the sample structure can now be incorporated as fixed effects via the first principal components. Let the matrix of first few principal components of the markers not in group $j$ is denoted by the matrix $PC_{-j}$. The model is written as

$$y = X\beta + ZPC_{-j}\tau_{-j} + Zg_j + e_j$$

(4)

where $\tau_{-j}$ is considered a fixed effect, $g_j \sim N_{n_k}(0, \sigma_{g_j}^2 K_j)$ for $j = 1, 2, \ldots, k$, $e_j \sim N_n(0, \sigma_{e_j}^2 I)$ and $g_j, e_j$ are independent. In the remaining of this paper, we will combine the fixed effect terms into one as $X^*\beta^*$ for notational ease.

Estimating the parameters of the model in (2) is very difficult with a large number of kernels. The models in (3) or (2) are more suitable for such cases. The model in (4) will be our preferred model for estimating local GEVs in the remaining of this paper since there are very efficient algorithms for estimating the parameters of this model.

Once the fixed effects and the variance parameters of the model in (4) are estimated for the $j$th region, the vector of expected value of the genetic effects (EBLUP) specific to region $j$ can be estimated by

$$\hat{g}_j = \hat{\sigma}_{g_j}^2 K_j Z' (\hat{\sigma}_{g_j}^2 ZK_j Z' + \hat{\sigma}_{e_j}^2 I)^{-1} (y - X^*\hat{\beta}^*)$$

for $j = 1, 2, \ldots, k$ for $j = 1, 2, \ldots, k$.

Post-processing Let $x$ be the $p$ dimensional vector of fixed effects and $m$ be the vector of markers partitioned into $k$ regions. Using the methods discussed in the previous section it is possible to obtain the expected value of the genetic effects (EBLUP) specific to region $j$ for an individual with marker set $m$, which we will denote by $\hat{g}_j(m)$. Also, let $\hat{g}^*_j(m) = \frac{\hat{g}_j(m)}{\hat{\sigma}_{g_j}}$ denote the standardized EBLUPs of random effect components that correspond to the $k$ local kernels for regions $j =$
Consider a final prediction model in the following form:

\[
f(x; \mathbf{m}; \beta, \alpha) = \beta_0 + \sum_{j=1}^{k} \alpha_j \hat{g}^*_j(\mathbf{m}) + \sum_{j=k+1}^{k+p} \beta_j x_j.
\]  

(5)

Estimate the model coefficients using the following

\[
(\hat{\beta}, \hat{\alpha}) = \arg\min_{(\beta, \alpha)} \sum_{i=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{k} \alpha_j \hat{g}^*_j(\mathbf{m}_i) + \sum_{j=k+1}^{k+p} \beta_j x_{ji}))^2 + \lambda \sum_{j=1}^{k} |\alpha_j| + \lambda_1 \sum_{j=1}^{k} |\alpha_j| + \lambda_2 \sum_{j=1}^{k} (\alpha_j)^2
\]  

(6)

where \( \lambda > 0 \) is the shrinkage operator, larger values of \( \lambda \) decreases the number of kernels included in the final prediction model.

When \( k \) is large compared to the sample size \( N \), we can estimate the parameters of the model using the elastic-net penalty, i.e.,

\[
(\hat{\beta}, \hat{\alpha}) = \arg\min_{(\beta, \alpha)} \sum_{i=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{k} \alpha_j \hat{g}^*_j(\mathbf{m}_i) + \sum_{j=k+1}^{k+p} \beta_j x_{ji}))^2 + \lambda_1 \sum_{j=1}^{k} |\alpha_j| + \lambda_2 \sum_{j=1}^{k} (\alpha_j)^2
\]  

(7)

to allow for more than \( N \) non zero coefficients in the final estimation model. \( \lambda_1, \lambda_2 > 0 \) are the shrinkage operators.

In our examples, we have used \( \hat{G}(\mathbf{m}) \hat{\alpha} \) as the estimated genotypic value for an individual with markers \( \mathbf{m} \) where \( \hat{G}(\mathbf{m}) = (\hat{g}^*_1(\mathbf{m}), \hat{g}^*_2(\mathbf{m}), \ldots, \hat{g}^*_k(\mathbf{m})) \). Since \( \hat{G}(\mathbf{m}) \) has standardized columns, \( |\hat{\alpha}| \) can be used as importance scores for the regions in the model.

**Hyperparameters of the model** While fitting the model in (5) we need to decide on the values of a number of hyperparameters. Apart from the model set-up that involve the definition of genomic regions, and inclusion or exclusion of some environmental or structural covariates; these parameters are the kernel parameters, and the parameters related to the elastic-net used in the post-processing step.

Low accuracies due to poor selection of the bandwidth parameter, \( h \), for the Gaussian kernel has been documented previously. Several methods like use of multiple Gaussian kernels simultaneously or model averaging have been recommended to overcome the shortcomings (GONZÁLEZ-CAMACHO et al. (2012), TUSELL et al. (2014)). We have experimented with these
approaches, but they did not produce consistent results for the data sets we analyzed. Therefore, we chose to compare a few predetermined values \( h = \{1/10, 1/5, 1, 5, 10\} / (m * n) \) where \( m \) is the number of markers that contribute to a kernel and \( n \) is the number of genotypes) and selected the value that gave the best ten fold cross-validated accuracy within a given training set.

The other hyperparameters we have varied were the “depth” and the “nsplit” parameters of the hierarchical model set-up scheme in Figure 1. These parameters may be selected by comparing the cross-validated accuracies within the training data set for several reasonable choices.

Inclusion of structural components in the models by which the local GEVs are evaluated is mainly to exclude the co-segregation effects due to family structure in the final model. For the examples in this paper, we have used the first five principal components as covariates during the calculation of local GEVs. These covariates are later dropped from the model during the post-processing step.

In our opinion, the hyperparameter choice for the multiple kernel models should reflect the available resources and the aims of the researcher. For instance, the number of regions that we can define depends on the number of markers, and a more detailed analysis might only be suitable when the number of markers and the number of genotypes in the training data set are large. The hyperparameters of the shrinkage estimators in the post-processing step allows us to control the sparsity of the model. These parameters can be optimized for accuracy using cross-validation, but their value can also be influenced by the amount of sparsity desired in the model. The multiple kernel models provide the user with the flexibility of models with a range of detail and sparsity.

**Illustrations** In this section, for four data sets that represent a variety of situations, we will compare the locally epistatic model with its counterpoints linear and Gaussian kernel SPMMs. The last examples are simulation studies to show that the short-range interactions can be effectively captured by the locally epistatic models and that the cross-validated accuracies within the training set can be used to adjust the hyperparameters.

In the following examples, the measure of accuracy is the correlation between the phenotypic
values and the corresponding estimated genotypic values. Comparisons of accuracy among the different models are based on repeated evaluations of the accuracy in a randomly selected subset (test set) based on models trained with the remaining individuals (training set).

If two traits are similar in terms of their importance scores, we can expect to have a genetic correlation between these traits. Genetic trait correlations are caused either by pleiotropy, close linkage or correlated physiological functions (Chen and Lübberstedt (2010)), and they are important to the breeders for improving correlated traits simultaneously or for reducing linkage drag. For the data sets that have several traits, we can provide a graphical representation of the importance scores from which the closeness of traits can be inferred. This representation is also useful for identifying QTL hotspots that were observed in other studies (Gardner and Latta (2007), Weber et al. (2008), Breitling et al. (2008), Zhao et al. (2011)).

For fitting the mixed models, we have developed and used the EMMREML package (Akdemir and Godfrey (2014)) and for the post-processing step we have used the glmnet package (Friedman and Hastie (2013)) both of which are available in R (R Core Team (2014)). The rest of the software were also programmed in R, and some of these were provided in the supplementary files.

Example 1. (Wheat Data) This dataset was downloaded from triticeaetoolbox.org. 3735 markers on 21 chromosomes (1A-7A, 1B-7B, and 1D-7D) for 337 elite wheat lines (SW-AM Panel) were available for the analysis. The traits (flowering date (FD), heading date (HD), physiological maturity date (MD), plant height (PH), yield (YD), waxiness (WX)) were obtained in two trials during the years 2012 and 2013. We have sampled 90% lines for training the models, and we have used the rest of the lines to evaluate the fit of our models. The whole genome was divided similarly to Figure 1 with a depth of two. The accuracies of the multiple kernel model compared with the linear and Gaussian kernel SPMMs and the mean genome-wide importance scores for regions used in our multiple kernel model over 30 replications of the experiment are summarized for different choices of the number of splits in Figure 2. The mean accuracies of the models for different traits are provided in Table 1. In addition, the importance scores from the multiple kernel
models are used to cluster the traits, and the dendrograms describe the resulting similarities of the traits in Figure 3.

Example 2. (Mouse Data) The Mouse data set we use for this analysis is available as a part of the R package SynbreedData (Wimmer et al. (2013)). Genotypic data consists of 12545 biallelic SNP markers and is available for 1940 individuals. The body weight at age of six weeks [g] and growth slope between six and ten weeks age [g/day] are measured for most of the individuals. The data was described in Valdar et al. (2006) and the heritabilities of these two traits are reported as 0.74 and 0.30. Here, we present the results from replication of the following experiment for the two traits at two different settings 30 times. A random sample of 1500 lines was selected in the training sample. The whole genome was divided in a similar fashion as displayed in Figure 1 with a depth of three. Two different multiple kernel models were obtained by using two versus three splits at each hierarchical level following the split by chromosomes (i.e., each chromosome was divided into four or nine regions). In addition, a Gaussian kernel model and a linear kernel model are fitted. The boxplots comparing accuracies of the models and the mean importance scores for different regions from the multiple kernel models over the 30 replications are displayed in Figures 4 and 5. For both traits, the multiple kernel model is substantially more accurate. In addition, the association derived as an output of this model supports the previously reported association of body weight related traits to the X chromosome (Dragani et al. (1995)).
Example 3. (Barley Data) Alpha-tocotrienols are in the class of fat-soluble chemical compounds related to vitamin E activity. Vitamin E deficiency is connected to many health problems, therefore, increased levels of alpha-tocotrienols is a desirable property for crops. In an experiment carried out by the Barley Coordinated Agriculture Project during the years 2006-2007, alpha-tocotrienol levels for 1723 barley lines were recorded in total of four environments (two years and three locations). 2114 markers on seven chromosomes were available for the analysis.

We have sampled 1500 lines for training the models, and we have used the rest of the lines to evaluate the fit of our models. The whole genome was divided in a similar fashion as displayed in Figure 7 with a depth of three with two splits at each level and only the regions at the most detailed level are used for multiple kernel model building. This is repeated 30 times. Accuracies and associations are summarized in Figure 6.

Example 4. (Maize Data) This data is given in ROMAY et al. (2013) and was also analyzed in WIMMER et al. (2013). 68120 markers on the 2279 USA national inbred maize lines and their phenotypic means for degree days to silking compose the data set. Accuracies for multiple kernel (MK) (5 regions per chromosome), linear kernel (Lin) and Gaussian kernel (Gaus) models for degree days to silking and the importance scores from the MK model are displayed in Figures 7 and 8.

Example 5. (Simulated Phenotypes: Accuracies) The purpose of this simulation study is to compare the accuracies of the SPMM with single Gaussian kernel and the multiple kernel models for traits that are generated by short-range to long-range interactions. For each replication and each setting of the experiment a set of 2000 genotypes was generated using the R Package “hyped” (TECHNOW (2013)) starting by randomly mating two founders with three chromosomes each of...
length 1M and 1000 markers per chromosome. After random mating for 300 generations, selection on a complex quantitative trait over 200 generations with selection intensity 50\% followed by another 100 generations of random mating, the final traits for the analysis were generated using short-range to long-range epistatic marker effects.

The final phenotypes in this example were generated randomly for each replication of the experiment by the following scheme: Given the set markers and the corresponding map, we first randomly select six markers and assign these markers effects from a zero centered normal distribution with variance one. Secondly, in an iterative fashion, we select one of the markers selected in the first step, say $m_1$ and two additional markers from previously not selected markers, say $m_2$ and $m_3$ and make an interaction term using the formula

$$r e^{-(r_1 m_1 + r_2 m_2 + r_3 m_3)w}$$

where $r$ has a zero centered normal distribution with variance three for the low interaction case and to six for the high interaction case. $r_1, r_2$ and $r_3$ independently and identically distributed random variables having a zero centered normal distribution with variance one and $w$ is selected at random from the set \{2, 4, 6\} with equal probability. While generating the local interaction scenarios, the three markers that generate the interaction are restricted to the same chromosome and the maximum distance between these markers is restricted to 1, 4, 10, 20, 30, or 50 cMs depending on the local interaction scenario. For the genome-wide case, no restrictions were imposed while generating the interactions. When total of 100 interaction effects are generated, the genotypic value of an individual is calculated by adding the main and the interaction effects. An independent error term from a zero centered normal distribution is added to each genotypic value to obtain the final phenotypic values while the error variance was chosen such that the heritability of the trait is 1/2. Finally, the markers that were used to generate the phenotypic values were excluded from the marker set before further analysis.

After the genotypes and the phenotypes are generated, a training sample of size 500, 1000, or 1500 was selected at random to estimate the phenotypes of the remaining individuals (test set)
with the following models: A single Gaussian kernel model (SK), multiple kernel models with 2, 5, 10, 20, or 30 splits per chromosome (MK$_2$, MK$_5$, ...MK$_{30}$). The performances of the models measured in terms of mean correlations between the true and the predicted values for the test data over 50 replications are summarized by the plots in Figure 9.

[Figure 9 about here.]

It is clear from these results that as the range or the intensity of the interactions increase the accuracies of all the models decrease. As expected, all models get more accurate as the sample size increases. For all of the cases, there is at least one multiple kernel model that outperforms the single kernel model, and the optimal number of splits seem to be a function of interaction range and the sample size. As the interaction range decreases, a larger number of splits produce more accurate models. On the other hand, for small sample sizes relatively low number of splits should be used. The locally epistatic models adequately capture local interactions leaving out the irrelevant parts of the genome and obtain high accuracies.

**Example 6.** (Simulated Phenotypes: Associations) The purpose of this second simulation is to evaluate the performance of the importance scores obtained from the multiple kernel models. For each replication of the experiment a set of 500, 750, 1000, or 1500 genotypes were generated using the R Package “hypred” (TECHNOW (2013)) starting by randomly mating two founders with three chromosomes each of length 1M and 1000 markers per chromosome. After random mating for 300 generations, the genotypic values were simulated for the individuals by randomly selecting two loci per chromosome and assigning the minor allele at each of these loci an effect of one. The phenotypic value for each individual was obtained by adding the genotypic values an independent error from a zero centered normal distribution with a certain variance so that the heritability of the trait was 1/2. For each of these datasets, multiple kernel models with 10, 20, 30, 40, or 50 splits for chromosome was trained. The whole experiment was replicated 100 times.

The importance scores of the regions obtained from the multiple kernel models cannot be directly used in formal hypothesis testing for association. Nevertheless, they provide important in-
formation about the contribution of genomic regions. For each of the replication of the experiment and the multiple kernel models with differing number of splits per chromosome, we have made a confusion matrix from the true classification of regions based on whether or not they contained an actual QTL and the classification based on whether the regions had zero or non-zero importance scores. In Figure 10, we report the results of our experiments using four measures of classification model performance averaged over the 100 replications. Sensitivity measures the proportion of the regions that contain QTL, which are correctly identified. Specificity measures the proportion of the regions that contain no QTL that are correctly identified. False discovery rate (FDR) measures the proportion of falsely identified positives to all positives that are identified. The F-score is a single measure of performance that is the harmonic mean of precision and sensitivity where precision is defined as one minus the FDR. Although these results are promising, they should be interpreted with caution: The elastic-net penalties we have used are tuned for accuracy and as the sample size increases more terms are allowed to enter the model as an artifact of linkage disequilibrium (LD). Having a non-zero importance score does not immediately imply a formal rejection of no effect in a region since this would also depend on many other things like the size of the importance score, the sample size, the number of regions being fitted, the LD in the population, etc. ...

[Figure 10 about here.]

In summary, the importance scores obtained from the multiple kernel models are indicative of the regions that include QTL. As the sample size increases or the number of splits decreases, the ability to identify regions (sensitivity) improves. However, this is accompanied by a deterioration in the ability of correctly identifying regions with no QTL (specificity and FDR). According to the F-score, the overall performance improves as the sample size increases. For larger sample sizes, increasing the number of splits increases the F-score. However, it seems like this trend might not be the same for smaller sample sizes.

Example 7. (Simulated Phenotypes: Number of Splits) The number of splits is an important hyper-parameter affecting the accuracies. In this example, we check the plausibility of a cross-validation
based method for identifying the correct number of splits. For each of the 30 replications of the experiment a set of 1000 genotypes with 1000 markers on each of the three chromosomes were generated as in Example 6. The marker set was divided into 3, 5 or 7 regions per chromosome and the trait values for the genotypes were generated by summing randomly generated locally epistatic effects from these regions and an independent error term with variance adjusted so that the heritability of the trait was 0.5. Models with 1 to 10 number of splits per chromosome were fitted to the genotypic and phenotypic data of a randomly selected training set of genotypes of size 800. The remaining genotypes were assigned to the test set. During the model fitting process, a 10-fold cross-validated accuracy measured in terms of correlation was obtained within the training set for each of these models. In addition, the accuracies of the models were calculated in the test data. The results of this experiment are summarized in Figure 11. The estimates of prediction accuracy obtained within the training set followed the same trend as the accuracies of the models calculated in the test data. The highest accuracies in both cases were observed at the true number of splits.

[Figure 11 about here.]

RESULTS AND DISCUSSION

The locally epistatic models proposed in this paper have good accuracy and explanatory value. Although the final results seem to depend on the complexity of the trait and population structure, similar or better accuracies were obtained for a number of populations compared to single kernel models. The multiple kernel models have the additional advantage that only a small fraction of genomic regions are used in the final model and the importance scores for these regions are readily available as a model output.

The locally epistatic models incorporate only the additive and local epistatic genetic contributions and exclude genome-wide interactions. The breeder can have confidence that these effects can be passed on to several generations. In addition, the information about the importance of genome regions produced from these models is also relevant since they point to regions that are relevant for introgression.
The approaches introduced in this article allow us to use the markers in naturally occurring blocks. The multiple kernel approach overcomes the memory problems that we might incur when the number of markers is very large by loading only subsets of data in the memory at a time. When studying the interactions, an order of magnitude of reduction of complexity can be obtained by only studying the interaction among the blocks instead of interactions among single loci. This block interaction approach will be a subject for future study.

The local kernels use information collected from regions of the genome and, because of linkage, will not be affected by a few missing or erroneous markers. In our examples, we have used the mean imputation to impute missing markers.

The accuracy of the multiple kernel model compared to its genome-wide counterparts partially depends on the trait architecture as illustrated by the contrast between the examples where a trait is affected only by a few regions (the multiple kernel model has high accuracies) or effects distributed homogeneously to the entire genome (genome-wide models have higher accuracies).

Another factor that influences the accuracies is the strength and structure of interactions. If the interactions are, in fact, local the multiple kernel models outperform the single kernel ones. In conclusion, the locally epistatic models are most accurate when a few major genes with local interactions generate the trait under study.

We can obtain local kernel matrices by defining regions in the genome and calculating a separate kernel matrix for each group and region. These regions can be overlapping, hierarchical or discrete. If some markers are associated with each other in terms of linkage or function, it might be useful to combine them together. The whole genome can be divided physically into chromosomes, chromosome arms or linkage groups. Further divisions could be based on recombination hotspots or just merely based on local proximity. We could calculate a separate kernel for introns and exons, non-coding, promoter or repressor sequences. We can also use a grouping of markers based on their effects on low-level traits like lipids, metabolites, gene expressions, or based on their allele frequencies. When some markers are missing for some individuals, we can calculate a kernel for the presence and absence states for these markers. When no such guide is present, one can
use a hierarchical clustering of the variables. It is even possible to incorporate group membership probabilities for markers, so the markers have varying weights in different groups. We are in the process of preparing another manuscript in which different sources of marker annotations are used with the multiple kernel models.

The hierarchical set up defines nested regions from coarse to fine, the regions are divided into sub-regions, and this is repeated to the desired detail level. An advantage to using a hierarchy is the availability of hierarchical testing procedures that have good cost / power properties (BLANCHARD and GEMAN (2005)). Multiple testing procedures where coarse to fine hypotheses are tested sequentially have been proposed to control the family-wise error rate or false discovery rate (REINER et al. (2003), MEINSHAUSEN (2008)). For example, to deal with the inflation of the error probabilities due to testing $k$ hypothesis in the hierarchical set up, Meinshausen’s hierarchical testing procedure (MEINSHAUSEN (2008)) controls the family-wise error by adjusting the significance levels of single tests in the hierarchy. The procedure starts testing the root node at level $\alpha$. When a parent hypothesis is rejected one continues with testing all the child nodes of that parent while the significance level to be used at each node is adjusted by a factor proportional to the number of variables in that node. Recently, there have also been some advances in significance testing for the lasso (LOCKHART et al. (2014)) regression. These procedures can be used with the multiple kernel models to obtain formal significance tests with desirable properties. A manuscript concerning the formal hypothesis testing with the multiple kernel models will also be ready shortly.

SUPPLEMENTARY MATERIALS

Supplementary 1: R-codes used for the illustrations.

S1.zip (Zipped file)

ACKNOWLEDGMENTS

This research was supported by the USDA-NIFA-AFRI Triticeae Coordinated Agricultural Project, award number 2011-68002-30029.
LITERATURE CITED

AKDEMIR, D. and O. U. GODFREY, 2014 EMMREML: Fitting mixed models with known covariance structures. R package version 2.0.

AMEMIYA, T., 1977 A note on a heteroscedastic model. Journal of Econometrics 6(3): 365–370.

BACH, F. R., G. R. LANCKRIET, and M. I. JORDAN, 2004 Multiple kernel learning, conic duality, and the SMO algorithm. In Proceedings of the twenty-first international conference on Machine learning, pp. 6. ACM.

BLANCHARD, G. and D. GEMAN, 2005 Hierarchical testing designs for pattern recognition. Annals of Statistics: 1155–1202.

BREITLING, R., Y. LI, B. M. TESSON, J. FU, C. WU, T. WILTSHIRE, A. GERRITS, L. V. BYSTRYKH, G. DE HAAN, A. I. SU, and OTHERS, 2008 Genetical genomics: spotlight on QTL hotspots. PLoS Genetics 4(10): e1000232.

CANTOR, R. M., K. LANGE, and J. S. SINSHEIMER, 2010 Prioritizing GWAS results: a review of statistical methods and recommendations for their application. The American Journal of Human Genetics 86(1): 6–22.

CHEN, Y. and T. LÜBBERSTEDT, 2010 Molecular basis of trait correlations. Trends in Plant Science 15(8): 454–461.

DE LOS CAMPOS, GUSTAVO AND GIANOLA, D., 2008 Inferring genetic values for quantitative traits non-parametrically. Genetics Research 90(06): 525–540.

DE LOS CAMPOS, G., D. GIANOLA, G. J. ROSA, K. A. WEIGEL, and J. CROSSA, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92(04): 295–308.

DEMIDENKO, E., 2004 Mixed Models: Theory and Applications (Wiley Series in Probability and Statistics). Wiley-Interscience.

DRAGANI, T., Z.-B. ZENG, F. CANZIAN, M. GARIBOLDI, M. GHIARDETTI, G. MANENTI,
and M. P. Pierotti, 1995 Mapping of body weight loci on mouse chromosome X. Mammalian Genome 6(11): 778–781.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4(3): 250–255.

Friedman, J. and T. Hastie, 2013 Tibshirani R. glmnet: lasso and elastic net regularized generalized linear models; 2008. R language.

Gardner, K. M. and R. G. Latta, 2007 Shared quantitative trait loci underlying the genetic correlation between continuous traits. Molecular Ecology 16(20): 4195–4209.

Gianola, D. and J. Van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4): 2289–2303.

González-Camacho, J., G. de Los Campos, P. Pérez, D. Gianola, J. Cairns, G. Mahuku, R. Babu, and J. Crossa, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics 125(4): 759–771.

Habier, D., R. Fernando, and J. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389–2397.

Hartley, H. O. and J. Rao, 1967 Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54(1-2): 93–108.

Harville, D., 1977 Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72(358): 320–338.

Hayes, B., P. Bowman, A. Chamberlain, and M. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2): 433–443.

Hayes, B., M. Goddard, and Others, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829.
HEFFNER, E. L., J.-L. JANNINK, H. IWATA, E. SOUZA, and M. E. SORRELLS, 2011 Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science 51(6): 2597–2606.

HESLOT, N., H.-P. YANG, M. E. SORRELLS, and J.-L. JANNINK, 2012 Genomic selection in plant breeding: a comparison of models. Crop Science 52(1): 146–160.

HOLLAND, J. H., 1975 Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press.

JARQUÍN, D., J. CROSSA, X. LACAZE, P. DU CHEYRON, J. DAUCOURT, J. LORGEOU, F. PIRAUX, L. GUERREIRO, P. PÉREZ, M. CALUS, and OTHERS, 2013 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics: 1–13.

KIMELDORF, G. and G. WAHBA, 1970 A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. The Annals of Mathematical Statistics: 495–502.

LEE, S. H., J. H. VAN DER WERF, B. J. HAYES, M. E. GODDARD, and P. M. VISSCHER, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genetics 4(10): e1000231.

LEGARRA, A., C. ROBERT-GRANIÉ, E. MANFREDI, and J.-M. ELSEN, 2008 Performance of genomic selection in mice. Genetics 180(1): 611–618.

LOCKHART, R., J. TAYLOR, R. J. TIBSHIRANI, R. TIBSHIRANI, and OTHERS, 2014 A significance test for the lasso. The Annals of Statistics 42(2): 413–468.

LORENZANA, R. E. and R. BERNARDO, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and Applied Genetics 120(1): 151–161.

MEINSHAUSEN, N., 2008 Hierarchical testing of variable importance. Biometrika 95(2): 265–
R Core Team, 2014 R: A Language and Environment for Statistical Computing.

Rao, C. R., 1971 Estimation of variance and covariance components MINQUE theory. Journal of Multivariate Analysis 1(3): 257–275.

Reiner, A., D. Yekutieli, and Y. Benjamini, 2003 Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19(3): 368–375.

Robinson, G., 1991 That BLUP is a good thing: The estimation of random effects. Statistical Science 6(1): 15–32.

Rodriguez, P. P. and G. de los Campos, 2012 BGLR: A Statistical Package for Whole-Genome Regression.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts, T. M. Casstevens, R. J. Elshire, C. B. Acharya, S. E. Mitchell, S. A. Flint-Garcia, and others, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology 14(6): R55.

Schölkopf, B. and A. Smola, 2002 Learning with kernels.

Sonnenburg, S., G. Rätsch, C. Schäfer, and B. Schölkopf, 2006 Large scale multiple kernel learning. The Journal of Machine Learning Research 7: 1531–1565.

Speed, T., 1991 [That BLUP is a Good Thing: The Estimation of Random Effects]: Comment. Statistical Science 6(1): 42–44.

Technow, F., 2013 hypred: Simulation of Genomic Data in Applied Genetics. R package version 0.4.

TuSELL, L., P. Pérez-Rodríguez, S. FORNí, and D. Gianola, 2014 Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. Journal of Animal Breeding and Genetics.
Valdar, W., L. C. Solberg, D. Gauguiere, S. Burnett, P. Kleinerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint, 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics 38(8): 879–887.

VanRaden, P., C. Van Tassel, G. Wiggans, T. Sonstegard, R. Schnabel, J. Taylor, and F. Schenkel, 2009 Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1): 16–24.

Weber, A. L., W. H. Briggs, J. Rucker, B. M. Baltazar, J. de Jesus Sanchez-Gonzalez, P. Feng, E. S. Buckler, and J. Doebley, 2008 The genetic architecture of complex traits in teosinte (Zea mays ssp. parviglumis): new evidence from association mapping. Genetics 180(2): 1221–1232.

Wimmer, V., T. Albrecht, H.-J. Auinger, and M. V. Wimmer, 2013 R Package ”synbreed-Data”.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, and C.-C. Schün, 2013 Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195(2): 573–587.

Xu, S., 2013 Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195(3): 1103–1115.

Zhao, K., C.-W. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali, A. H. Price, G. J. Norton, M. R. Islam, A. Reynolds, J. Mezey, and OTHERS, 2011 Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature Communications 2: 467.
| Figure | Title                                      | Page |
|--------|--------------------------------------------|------|
| 1      | Hierarchical set up                        | 28   |
| 2      | Wheat Data: Accuracies                     | 29   |
| 3      | Wheat Data: Associations                   | 30   |
| 4      | Mouse Data: Accuracies and Associations 1  | 31   |
| 5      | Mouse Data: Accuracies and Associations 2  | 32   |
| 6      | Barley Data: Accuracies and Associations   | 33   |
| 7      | Maize Data: Accuracies and Associations 1  | 34   |
| 8      | Maize Data: Accuracies and Associations 2  | 35   |
| 9      | Simulations: Accuracies                   | 36   |
| 10     | Simulations: Associations                 | 37   |
| 11     | Simulations: Number of Splits              | 38   |
Figure 1: A hypothetical hierarchical set up for an organism with three chromosomes. This division has two main parameters, namely the "depth" and the "nsplit". "depth" controls how many levels of splits should be performed. A "depth" of zero corresponds to the root of the tree, a "depth" of one would correspond to chromosomes, a "depth" of two would correspond to splitting the chromosomes and so on. The "nsplit" parameter controls the number of divisions after the chromosome level. In this figure, a setup with "depth" two and "nsplit" three is illustrated.
Figure 2: Wheat Data: Accuracies of the multiple kernel (MK) model compared to the Gaussian (Gaus) kernel model for six traits. Dots below the line in red correspond to the cases where the MK model is more accurate than the Gaus model.
Figure 3: Wheat Data: Associations from the multiple kernel (MK) model for the six traits.
Figure 4: Mouse Data: Accuracies and associations for multiple kernel (MK) and accuracies for linear kernel (Lin) and Gaussian kernel (Gaus) models for "weight at age of 6 weeks [g]".
(a) Four regions per chromosome

(b) Nine regions per chromosome

Figure 5: Mouse Data: Accuracies and associations for multiple kernel (MK) and accuracies for linear kernel (Lin) and Gaussian kernel (Gaus) models for "growth slope between 6 and 10 weeks age [g per day]".
Figure 6: Barley Data: Accuracies and associations for multiple kernel (MK) (4 regions per chromosome) and accuracies linear kernel (Lin) and Gaussian kernel (Gaus) models for tocotrienol levels.
Figure 7: Maize Data: Accuracies and associations for multiple kernel (MK) (25 regions per chromosome) and accuracies for linear kernel (Lin) and Gaussian kernel (Gaus) models for degree days to silking.
Figure 8: Maize Data: Accuracies for multiple kernel (MK) (36 regions per chromosome), linear kernel (Lin) and Gaussian kernel (Gaus) models for degree days to silking.
Figure 9: Simulations: Accuracies of models are compared for traits that are generated by short range to long range interactions. On the horizontal axis the scenarios are displayed in increasing order from left to right with respect to the range of interactions (1, 4, 10, 20, 30, 50 cMs and genome-wide). The vertical axis displays the accuracy.
Figure 10: Simulations: Four measures of classification model performance averaged over the 100 replications for sample sizes 500, 750, 1000 and 1500 and multiple kernel models with 10, 20, 30, 40, or 50 splits per chromosome.
Figure 11: Simulations: The estimated prediction accuracies (measured by correlation, r) by cross-validation in the training set for number of splits equal to 1 to 10 and estimated prediction accuracies in the test set for the same splits. The red vertical line indicates the correct number of splits for each case.
List of Tables

|   | Wheat Data: Accuracies |                         | 40 |
|---|------------------------|-------------------------|----|

39
Table 1: The mean accuracies of models compared for several models for the wheat data

| Model | MK 2 splits | MK 3 splits | MK 4 splits | MK 5 splits | MK 6 splits | Gaus | Lin |
|-------|-------------|-------------|-------------|-------------|-------------|------|-----|
| FD    | 0.48        | 0.47        | 0.46        | 0.48        | 0.47        | 0.48 | 0.47|
| MD    | 0.46        | 0.46        | 0.42        | 0.42        | 0.42        | 0.41 | 0.41|
| PH    | 0.66        | 0.64        | 0.63        | 0.65        | 0.62        | 0.63 | 0.61|
| GP    | 0.53        | 0.47        | 0.5          | 0.51        | 0.48        | 0.46 | 0.44|
| YLD   | 0.81        | 0.81        | 0.82        | 0.82        | 0.83        | 0.83 | 0.83|
| HD    | 0.56        | 0.55        | 0.53        | 0.53        | 0.55        | 0.56 | 0.53|
| WX    | 0.51        | 0.57        | 0.53        | 0.55        | 0.51        | 0.55 | 0.51|