Aus dem Institut für Kulturpflanzenwissenschaften
Universität Hohenheim
Fachgebiet: Bioinformatik
Prof. Dr. H.-P. Piepho

Evaluation of alternative statistical methods for genomic selection for quantitative traits in hybrid maize

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt
der Fakultät Agrarwissenschaften

von
Master of Science
Torben Schulz-Streeck
aus Kiel

2012

Die vorliegende Arbeit wurde am 24.07.2012 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften" angenommen.

Tag der mündlichen Prüfung:	27.09.2012

1. Prodekan:	Prof. Dr. A. Fangmeier
Berichterstatter, 1. Prüfer:	Prof. Dr. H.−P. Piepho
Mitberichterstatter, 2. Prüfer:	Prof. Dr. A. E. Melchinger
3. Prüfer:	Prof. Dr. J. Bennewitz

Table of contents

[1] Schulz-Streeck T, Piepho HP (2010) BMC Proceedings 4 (Suppl 1):S8
[2] Schulz-Streeck T, Ogutu JO, Piepho HP (2011) BMC Proceedings 5(Suppl 3):S12
[3] Schulz-Streeck T, Ogutu JO, Piepho HP (2013) Theor Appl Genet 126(1):69-82
[4] Schulz-Streeck T, Ogutu JO, Gordillo A, Karaman Z, Knaak C, Piepho HP (2013) Submitted to Plant Breeding
[5] Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP (2012) Crop Sci. 52:2453-2461

Abbreviations

AIC	Akaike information criterion
AIC$_C$	corrected AIC
BLUP	best linear unbiased predictions
BVN	bivariate-normal
cAIC	conditional AIC
CV	cross-validation
GCV	generalized cross-validation
GEBV	genomic estimated breeding values
GEI	genotype-by-environment interaction
GS	genomic selection
LD	linkage disequilibrium
LS-SVM	least squares support vector machine
MAS	marker-assisted selection
MEI	marker-by-environment interaction
MME	mixed model equations
QTL	quantitative trait locus/loci
RCBD	randomized complete block design
REML	restricted maximum likelihood
RMSE	root mean squared error
RR	ridge regression
RR-BLUP	ridge regression BLUP
SNP	single nucleotide polymorphism
TBV	true breeding values
VTGUL	validation set for tested genotypes in untested locations
VUGTL	validation set for untested genotypes in tested locations
VUGUL	validation set for untested genotypes in untested locations

1. General introduction

Maize (*Zea mays* L.) is one of the most important crop species with an annual production of about 844,358,253 tonnes (FAO, 2011). It is therefore of great economic interest to develop and deploy new techniques that can enhance the efficiency and cost-effectiveness of maize breeding programs. One such technique, genomic selection (GS), also known as genome-wide selection, is a relatively new approach that involves using genome-wide molecular marker information to improve plant and animal breeding. Recent and continuing advances in molecular marker technology have greatly reduced the cost of molecular marker information for plant and animal breeders alike and thereby enabled widespread use of genomic selection in breeding programs. The maize breeding sector was one of the first plant breeding sectors to implement genomic selection in industrial breeding programs. The importance of maize makes the improvement and implementation of GS in maize breeding programs of significant economic interest and importance worldwide.

Genomic selection was first proposed by Meuwissen et al. (2001) to improve the efficiency and cost-effectiveness of plant and animal breeding programs. In contrast to the traditional marker-assisted selection (MAS) methods, where only selected subsets of markers are used, GS uses all the available marker information to predict breeding values. The aim of GS is to predict breeding values instead of detecting single quantitative trait loci (QTL). Thus, no significant tests to identify markers linked to QTL with large effects on a trait are used (Meuwissen et al., 2001). Traditional marker-assisted selection methods are best suited for traits controlled by a few QTL with large effects (Lande and Thompson, 1990; Holland, 2004; Xu and Crouch, 2008; Bernado, 2008), and perform poorly when used to estimate effects of QTL with small effect sizes (Lande and Thompson, 1990; Xu and Crouch, 2008; Bernado, 2008). However, quantitative traits are mainly controlled by many QTL with small effects (Kearsey and Farquhar, 1998; Bernardo, 2002). To accurately predict a quantitative trait it is crucial to use all the QTL affecting the trait in a marker-assisted selection exercise and not only the particular subset of markers which are in high linkage disequilibrium with QTL having large effects (Meuwissen et al., 2001). GS therefore achieves high accuracy by simultaneously estimating the effects of all the available markers without first prescreening the markers using significance tests to identify the most important and relevant markers. To ensure that at least one marker is in high linkage with a QTL the markers must cover the entire genome. Recent advances in single nucleotide polymorphism (SNP) marker genotyping technologies ensure this requirement (Syvänen, 2005; Li et al., 2008). Currently, thousands of SNP markers are available for most livestock species while for maize it is common to use 50,000 SNP markers. The cost of genotyping of these markers is currently reasonably low and will almost certainly continue to decline further as the genotyping technology platforms become increasingly more efficient.

Several studies have demonstrated the importance of GS in maize breeding (Lorenzana and Bernardo, 2009; Piepho, 2009; Crossa et al., 2010; Albrecht et al., 2011; Heslot et al., 2012; Riedelsheimer et al., 2012). GS has also been shown to improve the gain per unit time over traditional marker-assisted selection methods (Heffner et al., 2010). Moreover, GS is advantageous over the classic approach of using only the pedigree information (Crossa et al., 2010; Albrecht et al., 2011). These studies and others have proposed various statistical methods for GS. However, interest in increasing the predictive accuracy and applicability of the methods used for GS argues for the need to develop novel approaches or extensions of existing approaches to enhance their accuracy in GS. In this thesis a few commonly used methods are extended and new approaches are proposed to improve the accuracy of GS in particular in maize breeding.

1.1. Genomic selection for breeding value estimation

Genomic selection is based on using regression models relating phenotypic data to molecular marker information to predict expected genomic breeding values. The statistical models are fitted to a training population consisting of individuals with both phenotypic and genotypic data to estimate the effects of all the markers. The estimated marker effects can then be used to predict genotypic values for target populations with individuals that have been genotyped with the same markers as the training population but which are lacking phenotypic information (Figure 1). The methods commonly assume additive marker effects. Therefore predicted values for the untested genotypes are computed as the sum of all the additive marker effects, called genomic estimated breeding values (GEBV). These GEBVs are then used to guide selection decisions. The GS approach therefore both expedites and increases the cost-effectiveness of the plant breeding program because not all genotypes have to be tested in field trials (Bernado and Yu, 2007; Mayor and Bernado, 2009; Heffner et al., 2010; Jannink et al., 2010). Thus, the gain per unit time is increased (Heffner et al., 2009; 2010).

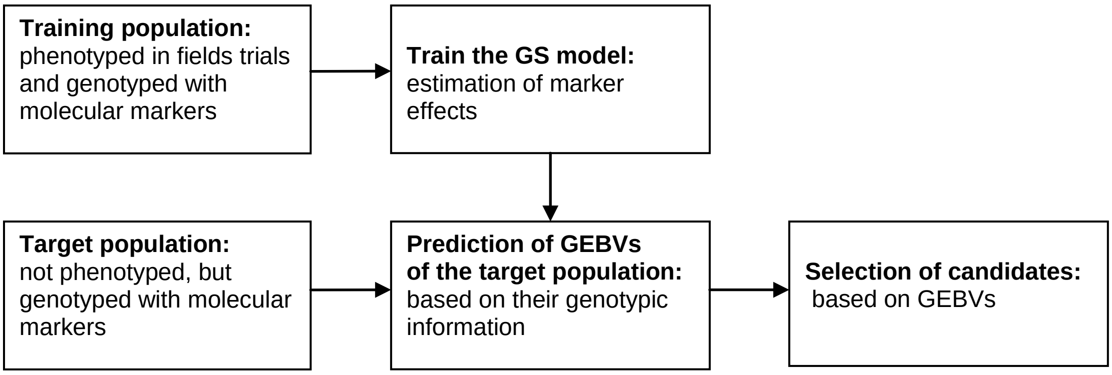
Training population:
phenotyped in fields trials and genotyped with molecular markers
Train the GS model:
estimation of marker effects
Target population:
not phenotyped, but genotyped with molecular markers
Prediction of GEBVs of the target population:
based on their genotypic information
Selection of candidates:
based on GEBVs

Figure 1: A schematic diagram illustrating genomic selection, starting from the training and target populations through prediction of genomic estimated breeding values (GEBV) to select candidates. It is important to note that while only a single instance of model training is depicted here, training can be performed continually as new phenotype and marker data become available (modified from Heffner et al., 2009).

The phenotypic data used for GS in plant breeding experiments are normally adjusted means derived from a statistical model fitted to the raw plot data in a step called phenotypic analysis. The whole process of obtaining GEBVs consists of at least two stages, namely the phenotypic analysis and the genomic analysis stages. The splitting of GS into two or more steps proceeds in the same spirit of stage-wise approaches used in analysis of a series of plant breeding experiments (Smith et al., 2001; Möhring and Piepho, 2009; Welham et al., 2010; Piepho et al., 2011). For a series of plant breeding experiments, a single-stage analysis is regarded as the gold standard because it can fully account for the entire variance-covariance structure of the observed data (Smith et al., 2001). Therefore, if the whole GS analysis is split into different steps, the different steps should collectively be able to approximately recover a single-stage analysis (Piepho, 2009).

1.2. Statistical methods for genomic selection

Using high density markers presents a general problem for GS, because the number of markers (*p*) can far exceed the number of observations (*n*), thus precluding the use of the standard multiple linear regression methods without first performing variable selection to appropriately reduce the number of markers. Many different methods have been proposed to overcome this limitation and used for GS to simultaneously estimate all the marker effects without having to first do significant tests to select a subset of markers. The proposed approaches include, but are not restricted to mixed models (Meuwissen et al., 2001; Piepho, 2009), Bayesian procedures (Meuwissen et al., 2001; Habier et al., 2011), machine learning methods (Long et al., 2007; Ogutu et al., 2011) and regularized regression methods (Heslot et al., 2012; Ogutu et al., 2012).

Meuwissen et al. (2001) were the first to propose a mixed model for GS, where the marker effects or chromosomal segments are taken as random effects, so that the best linear unbiased predictions (BLUP) of the marker effects can be generated. The model assumes that the marker effects are independent random draws from a common normal distribution. All the marker effects are drawn from the same distribution. Therefore each marker has the same variance and all estimated marker effects are shrunken equally towards zero. The marker variance can be estimated as a function of the total genetic variance and used as fixed variance in the GS analysis (Meuwissen et al., 2001; Bernardo and Yu, 2007; Habier et al., 2007). Alternatively, the restrictive assumption of a common fixed marker variance can be lifted by directly estimating the marker variance by restricted maximum likelihood (REML) in the GS analysis (Piepho, 2009). This method is commonly known as ridge regression BLUP (RR-BLUP).

A variety of Bayesian approaches have also been proposed to overcome the restrictive assumption of a homogenous marker variance, because it can lead to an underestimation of QTL with large effects (Meuwissen et al., 2001). In a first approach, called BayesA, each marker effect is drawn from its own normal distribution, allowing the variances to differ among the markers. The prior distribution of variances of the markers is a scaled, inverted chi-square distribution. Thus each marker effect is shrunken toward zero differently. In a second Bayesian approach, called BayesB, the possibility that some markers can have zero variances is explicitly accounted for (Meuwissen et al., 2001). BayesB uses a prior distribution for the marker variance that assumes that the marker variance is zero with a known probability (π) but non-zero otherwise. Similar to BayesA each no-zero marker effect is drawn from its own normal distribution. Several refinements have been undertaken to address problems related to the impact of prior distributions assumed for BayesA and BayesB on genomic predictions, including BayesCπ. In BayesCπ, the prior probability (π) that a marker has a zero effect is treated as an unknown to be estimated. For all the other markers with a $(1-\pi)$ probability of having a non-zero effect, a single variance is assumed, similarly to RR-BLUP, instead of a marker-specific variance, as assumed for BayesA and BayesB (Habier et al., 2011).

Despite the different assumptions made by the different Bayesian methods, the predictive accuracies for quantitative traits are often similar to each other (Habier et al. 2011) and to RR-BLUP based on real datasets for animal and plant breeding populations (e.g. Hayes et al., 2009a; VanRaden et al., 2009; Verbyla et al., 2009; Crossa et al., 2010; Heslot et al., 2012). The difference in performance between BayesB and RR-BLUP on real datasets is actually much lower than has been suggested by some simulation studies (Hayes et al., 2009a; VanRaden et al., 2009). Moreover, no method has emerged as clearly the best based on a wide variety of tested traits and species (e.g. Heslot et al., 2012). However, at least in theory, the different methods are best suited to predict different types of traits. Thus, using a method that shrinks the marker effects equally can lead to an underestimation of large effect QTL (Meuwissen et al., 2001; Xu, 2003; Verbyla et al., 2009), whereas BayesA and BayesB can better fit data with a few QTL each with large effects (Hayes et al., 2009b; Verbyla et al., 2009). This can be advantageous when estimating some traits, such as the fat percentage of milk in dairy cattle, which is controlled primarily by a polymorphism in the DGAT1 gene (Grisart et al., 2002). Models allowing for individual marker variances, such as BayesA, can fit such large effects better than models assuming a homogenous marker variance, such as BLUP methods (Hayes et al., 2009b; Verbyla et al., 2009). However, for traits controlled by many genes with small effects, differences between Bayes and BLUP methods are typically negligible (e.g. Hayes et al., 2009b; Verbyla et al., 2009).

The most relevant traits in maize breeding are mostly complex quantitative traits. Thus, it is assumed that an infinitesimal genetic model, which assumes a large number of QTL with small effects, would be reasonable for most relevant traits in maize breeding (Schön et al., 2004; Riedelsheimer et al., 2012). The assumption underling RR-BLUP is thus consistent with the infinitesimal genetic model. Thus, the prediction accuracy of RR-BLUP can be relatively high for quantitative traits from maize breeding populations (Lorenzana and Bernardo, 2009; Albrecht et al., 2011; Heslot et al., 2012; Zhao et al., 2012). Moreover, it is feasible to extend RR-BLUP to allow for heterogeneous marker variances (Meuwissen, 2009: Piepho, 2009) and to exclude markers with estimated zero variances in the spirit of BayesB (Piepho, 2009).

The regularized regression methods can be used even if the number of predictors far exceeds the number of observations, through the use of appropriate penalty functions. The regularized regression methods are intimately related to RR-BLUP (Ruppert et al., 2003). RR-BLUP is a mixed model which can be viewed as a ridge regression model, in which the penalty term is estimated by the quotient of the residual and the marker variance components (Ruppert et al., 2003; Piepho, 2009). The penalty term can be chosen by different data-driven methods, for example by cross-validation (Ruppert et al., 2003). The close connection between regularized regression and mixed models implies that both types of methods may be expected to have similar predictive accuracies for GS. This was recently demonstrated by simulation and empirical studies for several regularized regression methods, including the elastic net, LASSO and ridge regression models (Heslot et al., 2012; Ogutu et al., 2012).

Other machine learning methods with demonstrated high performances in many application domains have also recently been tried for GS, including random forest, boosting, support vector machines and artificial neural networks (González-Recio and Forni, 2011; Ogutu et al., 2011; Heslot et al., 2012). These methods efficiently handle the problem of far more markers than the number of observations and are not only restricted to regression problems involving quantitative traits alone but can also be used for classification problems (Drucker et al., 1997; Bühlmann and Hothorn, 2007; Hastie et al., 2009). Random forest is potentially attractive for GS because it can accommodate complex interactions between markers, nonlinear effects of markers and makes no distributional assumptions about the predictor variables (Breiman, 2001). Nonetheless, evidence available so far suggests similar performance of these machine learning methods and RR-BLUP (Ougtu et al., 2011; Heslot et al., 2012).

Overall, predictive accuracies of statistical methods tested for GS are broadly comparable, even though particular types of methods may be preferable for specific traits. Mixed models possess the appealing property that they can readily be extended by adding more fixed and random effect terms to account for extra sources of variation, such as design and genotype-environment interaction effects, and to accommodate heterogeneous variance components (Piepho, 2009). Therefore, mixed models are of special interest for GS, because they are flexible and competitive with other methods. Consequently, mixed models form the central focus of this thesis.

1.3. Mixed models for genomic selection

The following mixed model for adjusted means of genotypes is commonly used for GS (Piepho, 2009):

$$y = 1_n \mu + Zu + e, \tag{1}$$

where y is an n-vector of adjusted means per genotype, 1_n is an n-vector of ones, μ is a common intercept, Z is an $n \times p$ covariate matrix of p markers for n genotypes, where biallelic markers with alleles A_1 and A_2 are commonly coded as 1 for $A_1 A_1$, as -1 for $A_2 A_2$ and coded as 0 for $A_1 A_2$, $A_2 A_1$ and missing values, u is a vector of random marker effects with

$$u \sim N\left(0, I_p \sigma_u^2\right),$$

where the variance-covariance matrix of u is the product of the p-dimensional identity matrix (I_p), and the variance of marker effects (σ_u^2). The residual error associated with y, e is assumed to follow

$$e \sim N\left(0, I_n \sigma_e^2\right)$$

where I_n is an n-dimensional identity matrix and σ_e^2 is the residual variance.

Estimates of marker effects can be obtained by

$$\tilde{u} = \left(\tilde{Z}^T \tilde{Z} + \hat{\lambda}^2 D\right)^{-1} \tilde{Z}^T y\,,$$

where $\tilde{Z} = \left(1_n \quad Z\right)$, $D = \left(0 \oplus I_p\right)$ with $\oplus$ denoting the direct sum (Searle et al., 1992), $\hat{\lambda}^2 = \hat{\sigma}_e^2 / \hat{\sigma}_u^2$, and $\tilde{u}^T = \left(\mu, u^T\right)$.

This expression is commonly known as the “ridge regression” formulation of BLUP (Ruppert et al., 2003), and is called ridge regression BLUP (RR-BLUP).

Model (1) can be re-written to estimate genotypic instead of marker effects by specifying the variance-covariance of the genotypes in terms of the marker information (Piepho, 2009), which is a useful strategy if the number of markers exceeds the number of genotypes (VanRaden, 2008; Piepho et al., 2012).

$$y = 1_n \mu + g + e\,, \tag{2}$$

where $g = Zu$, $\mathrm{var}(g) = G = \Gamma \sigma_u^2$ and $\Gamma = ZZ^T$, Z^T denotes the transpose of Z and e is defined as in (1). BLUPs of u can be obtained, if Γ is positive-definite, and hence invertible as (Henderson, 1977)

$$\hat{u} = Z^T \Gamma^{-1} \hat{g}\,.$$

In the case Γ is not positive-definite random effects of u can be predicted by

$$\hat{u} = \hat{\sigma}_u^2 Z^T \hat{V}^{-1} \left(y - 1_n \hat{\mu}\right),$$

because the estimated variance-covariance matrix of y ($\hat{V}$) is generally positive-definite even when Γ is not.

In model (2) the covariance between two genotypes is modelled as a function of the distance between their marker profiles, similarly to the idea of modelling the covariances between pairs of observations made at different spatial locations as functions of their separation distances in geostatistics (Schabenberger and Gotway, 2005). The spatial distances are replaced here by genetic distances, calculated as multidimensional Euclidean distances between the markers for all possible pairs of genotypes (Piepho, 2009).

$$\Gamma = \left[f(d_{ii'}) \right]$$

where $d_{ii'}$ is the Euclidean distance between genotypes i and i', defined as $d_{ii'} = \| z_i - z_{i'} \|$, with z_i equal to the i-th row of Z, and $f(d)$ is some monotonically decreasing function of d. It has been shown that the RR-BLUP model is equivalent to a quadratic spatial model, where $f(d)$ has a quadratic function (Piepho, 2009). There are many other different options available for the $f(d)$ function (Schabenberger and Gotway, 2005; Piepho, 2009; Ober et al., 2011). A different but equivalent formulation of this geostatistical model for GS in terms of simple and universal kriging was recently proposed by Ober et al. (2011). These geostatistical models are closely related to the reproducing kernel Hilbert spaces regression model, a regularized regression model that uses kernel functions, such as the Gaussian kernel, to compute the distances between marker profiles for pairs of genotypes for the function $f(d)$ and that can therefore allow for possibly nonlinear marker effects on quantitative traits (Gianola and van Kaam, 2008).

1.4. Extensions and refinements of genomic selection methods

The basic models for GS can be extended or refined in several different ways. For the mixed models, the residual variance can be fixed through an independent estimate made at the phenotypic analysis stage when adjusted means are calculated. Fixing the residual variance can avoid overfitting, because it prevents the markers from capturing the non-genetic error variance (Piepho, 2009). Moreover, polygenetic effects can be included in mixed models to account for genetic variance not captured by the markers. This is commonly done in animal breeding studies using a random genotype effect, where the variance-covariance structure depends on the pedigree information (e.g. Calus and

Veerkamp, 2007). It is also possible to use independent random genotype effects, even if pedigree information is unavailable (Piepho, 2009).

Accurate prediction of GEBVs requires numerous markers; if many markers actually have zero effects and the effects are estimated by BLUP as non-zero, then their cumulative effect adds noise to the estimated marker effects (Goddard and Hayes, 2007). As a consequence predictive accuracy can be enhanced by excluding subsets of markers with no effects that otherwise would compromise the accuracy of prediction of genomic breeding values (Hayes et al. 2009a; Macciotta et al., 2009). Alternatively, some methods automatically select the most significant and pertinent markers, e.g. BayesB, lasso-type models, componentwise and twin boosting algorithms (Meuwissen et al., 2001; Bühlmann and Hothorn, 2007, 2010; Ougtu et al., 2012).

In general, the prediction of GEBVs is influenced by both the phenotypic analysis preceding the marker-based analysis plus non-marker information included in the marker-based methods. Clearly, genomic selection is not necessarily restricted to marker-based prediction but rather is a general approach for accurately predicting GEBVs based on the raw phenotypic data, marker information and non-marker effects using various statistical models applied in several stages.

1.5. Aims and objectives

The overall goal of this study was to develop and comparatively evaluate different approaches for accurately predicting genomic breeding values in GS, with GS viewed as a general approach, incorporating all the different stages from phenotypic analysis of the raw data to the marker-based prediction of the breeding values. In particular, the specific objectives were the following.

(1) Develop different approaches for using information from analyses preceding the marker-based prediction of breeding values for GS. Normally, adjusted means obtained from an analysis of the raw phenotypic data are used for GS leading to a stage-wise analysis. From the phenotypic analysis stage an estimate of the error variance can be obtained and used as a fixed residual variance component in the subsequent stage of GS (Chapter 2). Since the adjusted means are often correlated, in particular when trial designs are unbalanced, it may be crucial to preserve and carry forward the information contained in the variance-covariance matrix from the phenotypic analysis stage to the

marker-based prediction stage of GS (Chapter 4). Furthermore, excluding markers with no effects, or with inconsistent effects among crosses, environments or generations can increase predictive accuracy (Chapters 3 and 5). The predictive accuracy of GS may still be further improved by conducting residual diagnostics at the phenotypic analysis stage of field trials to identify and eliminate outlying observations (Chapter 6).

(2) Extend and/or suggest efficient implementations of statistical methods used at the marker-based prediction stage of GS, with a special focus on improving the predictive accuracy of GS in maize breeding. The basic RR-BLUP methods were modified in several different ways in an attempt to enhance their predictive accuracies and/or accommodate non-marker effects. To this end different strategies for modeling polygenic effects were explored and tested (Chapter 2). The models were also modified to accommodate the main marker effects only, as well as the main plus specific marker effects for a variety of environments (Chapters 5). This was achieved by testing if it is necessary to account for marker effects in phenotypic analysis across different environments or if the marker information can be omitted until the very last stage of the stage-wise analysis (Chapter 5). Moreover, GS using multiple populations was compared to analysing each population separately and to an analysis, where both the main and populations-specific marker effects are simultaneously estimated (Chapter 6). Also considered was a modification of the RR-BLUP model allowing different subsets of markers to have different variances (Chapter 3). Beside these modifications of the RR-BLUP method, different spatial models and machine learning methods, including regularized regression methods, namely component-wise boosting, ridge regression, LASSO and the elastic net were also used for GS and their performances compared with those of RR-BLUP (Chapters 2, 3, 4 and 6).

(3) Compare different approaches to reliably evaluate and compare methods for GS. Several cross-validation (CV) schemes were proposed and used to evaluate and compare the different GS approaches (Chapters 3 to 6). One of the proposed schemes attempted to satisfy the fundamental assumption underlying the proper use of CV, namely that the training and validation sets are independent (Chapter 4). Because CV can sometimes be computationally too demanding to implement for certain large problems, several different alternative model selection criteria are proposed and evaluated (Chapter 2 and 5).

References

Albrecht, T., V.Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C.C. Schön. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123:339–350.

Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48:1649–64.

Bühlmann, P., and T. Hothorn. 2007. Boosting algorithms: regularization, prediction and model fitting. Statist. Sci. 22: 477-505.

Bühlmann, P. and T. Hothorn. 2010. Twin Boosting: Improved feature selection and prediction. Statistics and Computing 20(2):119–138.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

Calus, M.P.L., and R.F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124:362–368.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, G. Atlin, J. Burgueño, J.L. Araus, D. Makumbi, J. Yan, V. Arief, M. Banziger, and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Drucker, H., C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9:155–161.

FAOSTAT. 2011. Available at http://faostat.fao.org/?PageID=567#ancor (verified 26 Feb. 2011).

Gianola, D., and J.B.C.H.M. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2305-2313.

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323-330.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43:7.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, N. Cambisano, M. Mni, S. Reid, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222–231.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hastie, T.J., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Springer, New York, USA.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy. Sci. 92: 433-443.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M.E. Goddard. 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51.

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: potential gain per unit time and cost. Crop Sci. 50:1681–1690.

Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49: 1–12.

Henderson, C.R. 1977. Best linear unbiased prediction of breeding values not in the model for records. J. Dairy Sci. 60:783-787.

Heslot, N., H.P. Yang, M.E. Sorrels, and J.L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52:146-160.

Holland, J.B. 2004. Implementation of molecular markers for quantitative traits in breeding programs: Challenges and opportunities. p. 26. In T. Fischer et al. (ed.) New Directions for a Diverse Planet: Proc. for the 4th Int. Crop Science Congress, Brisbane, Australia. 26 Sept.–1 Oct. 2004. Regional Institute, Gosford, Australia.

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomic Proteomic 9: 166–177.

Kearsey, M., and A. Farquhar. 1998. QTL analysis in plants: Where are we now? Heredity 80:137–142.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–56.

Li, C., M. Li, J.R. Long, Q. Cai, and W. Zheng. 2008. Evaluating cost efficiency of SNP chips in genome-wide association studies. Genet. Epidemiol. 32:387–395.

Long, N., D. Gianola, G.J.M. Rosa, K.A. Weigel, and S. Avendan□o. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. 124:377–89.

Lorenzana, R., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161.

Macciotta, N.P.P., G. Gaspa, R. Steri, C. Pieramati, P. Carnier, and C. Dimauro. 2009. Preselection of most significant SNPS for the estimation of genomic breeding values. BMC Proc. 3 (Suppl 1):S14.

Mayor, P.J., and R. Bernardo. 2009. Genome-wide selection and marker-assisted recurrent selection in double haploid versus F2 population. Crop Sci. 49: 1719–1725.

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

Meuwissen, T.H.E. 2009. Accuracy of breeding values of ‘unrelated’ individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.

Möhring, J., and H.P. Piepho. 2009. Comparison of weighting in two-stage analyses of series of experiments. Crop Sci. 49:1977-1988.

Ober, U., M. Erbe, N. Long, E. Porcu, M. Schlather, and H. Simianer. 2011. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188:695-708.

Ogutu, J.O., H.P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection using SNP markers. BMC Proc. 5 (Suppl 3):S11.

Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions (accepted).

Piepho, H.P. 2009. Ridge regression and extensions for genome-wide selection in maize. Crop Sci. 49:1165-1176.

Piepho, H.P., J.O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow. 2012. Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci. doi: 10.2135/cropsci2011.11.0592.

Piepho, H.P., T. Schulz-Streeck, and J.O. Ogutu. 2011. A stage-wise approach for analysis of multi-environment trials. Biuletyn Oceny Odmian 33:7-20.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A.E. Melchinger. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. doi:10.1038/ng.1033.

Ruppert, D., M.P. Wand, and R.J. Carroll. 2003. Semiparametric regression. Cambridge Univ. Press. Cambridge, UK.

Schabenberger, O., and C.A. Gotway. 2005. Statistical methods for spatial data analysis. CRC Press Boca Raton, FL.

Schön, C.C., H.F. Utz, S. Groh, B. Truberg, S. Openshaw, and A.E. Melchinger. 2004. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York.

Smith, A., B. Cullis, and R. Thompson. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138-1147.

Syvänen, A.C. 2005. Toward genome-wide SNP genotyping. Nat. Genet. 37:S5–S10.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Verbyla, K.L., B.J. Hayes, P.J. Bowman, and M.E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genetics Research 91(05):307-311.

Welham, S., B.J. Gogel, A.B. Smith, R. Thompson, and B.R. Cullis. 2010. A comparison of analysis methods for late-stage evaluation trials, Australian and New Zealand Journal of Statistics 52:125-149.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801.

Xu, Y., and J.H. Crouch. 2008. Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. 48: 391–407.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, and J.C. Reif. 2011. Accuracy of genomic selection in European maize elite breeding Populations. Theor. Appl. Genet. 124:769–776.

2. Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models

Torben Schulz-Streeck, Hans-Peter Piepho

Bioinformatics Unit, Institute for Crop Production and Grassland Research, Universität Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

BMC Proceedings 4 (Suppl 1):S8

The original publication is available at http://www.biomedcentral.com/bmcproc/

Abstract

Genome-wide selection (GS) involves estimating breeding values using molecular markers spanning the entire genome. The success of GS approaches will depend crucially on the availability of efficient and easy-to-use computational tools. Therefore, approaches that can be implemented using mixed models hold particular promise and deserve detailed study. A particular class of mixed models suitable for GS is given by geostatistical mixed models, when genetic distance is treated analogously to spatial distance in geostatistics.

We consider various geostatistical mixed models for use in GS. The analyses presented for the QTL-MAS 2009 dataset pay particular attention to the modelling of residual errors as well as of polygenetic effects.

It is shown that geostatistical models are viable alternatives to ridge regression best linear unbiased prediction, one of the common approaches to GS. Correlations between genome-wide estimated breeding values and true breeding values were between 0.879 and 0.889. In the example considered, we did not find a large effect of the residual error variance modelling, largely because error variances were very small. A variance components model reflecting the pedigree of the crosses did not provide an improved fit. Therefore geostatistical models deserve further study as a tool to GS that is easily implemented in a mixed model package.

3. Pre-selection of markers for genomic selection

Torben Schulz-Streeck, Joseph O Ogutu, Hans-Peter Piepho

Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

BMC Proceedings 2011, 5(Suppl 3):S12.

The original publication is available at http://www.biomedcentral.com/bmcproc/

Abstract

Accurate prediction of genomic breeding values (GEBVs) requires numerous markers. However, predictive accuracy can be enhanced by excluding markers with no effects or with inconsistent effects among crosses that can adversely affect the prediction of GEBVs.

Therefore we present three different approaches for pre-selecting markers prior to predicting GEBVs and assess the extent to which pre-selection of markers improves prediction accuracy. Four different best linear unbiased prediction (BLUP) methods, including ridge regression and three geostatistical models were used and the performances of the models were evaluated using 5-fold cross-validation.

Ridge regression BLUP and the geostatistical models gave almost similar fits. Pre-selecting markers was beneficial. Thus excluding markers with inconsistent effects among crosses increased the correlation between GEBVs and true breeding values of the non-phenotyped individuals from 0.607 (using all markers) to 0.625 (using pre-selected markers). Moreover, extension of the ridge regression model to allow for heterogeneous variances between the n (n =5, 10, 50, 100, 250) most significant markers and the remaining markers only marginally increased the accuracy of prediction (from 0.625 to 0.648) for the simulated dataset for the QTL-MAS 2010 workshop.

4. Comparisons of single-stage and two-stage approaches to genomic selection

Torben Schulz-Streeck, Joseph O Ogutu, Hans-Peter Piepho

Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

Theoretical and Applied Genetics 126(1):69-82.

The original publication is available at http://link.springer.com/journal/122

Abstract

Genomic selection (GS) is a method for predicting breeding values for plants or animals using many molecular markers that is commonly implemented in two stages. In plant breeding the first stage usually involves computation of adjusted means for genotypes which are then used to predict genomic breeding values (GEBVs) in the second stage. Classical stage-wise approaches to GS suffer from either approximating, or ignoring correlations among the adjusted means by assuming that the means are independent. We comparatively evaluated the performance of a new stage-wise method for GS, which uses rotation to ensure the means are approximately independent, and that is fully efficient relative to a single-stage approach given known variance components. Specifically, we compared two classical stage-wise approaches, which either ignore or approximate correlations among the means by a diagonal matrix, and the new method to a single-stage analysis for GS. We further evaluated the predictive performance of the new method when implemented by ridge regression best linear unbiased prediction (RR-BLUP) and componentwise linear least squares boosting using 5-fold cross-validation. The new stage-wise approach with rotated means was more similar to the single-stage analysis than the classical two-stage approaches based on non-rotated means for two unbalanced datasets. This suggests that rotation is a worthwhile pre-processing step in GS for the two-stage approaches for unbalanced datasets. Moreover, the predictive accuracy of stage-wise RR-BLUP was higher (5.0 to 6.1%) than that of boosting.

5. Genomic selection allowing for marker-by-environment interaction

Torben Schulz-Streeck[1,2], Joseph O Ogutu[1], Andrés Gordillo[3,4], Zivan Karaman[5], Carsten Knaak[2], Hans-Peter Piepho[1]

[1] Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

[2] KWS SAAT AG, Grimsehlstraße 31, 37555 Einbeck, Germany

[3] AgReliant Genetics, LLC, 4640 East State Road 32, Lebanon, Indiana 46052, USA

[4] KWS LOCHOW GMBH, Ferdinand-von-Lochow-Strasse 5, 29303 Bergen, Germany

[5] Limagrain Europe, CS 3911, 63720 Chappes, France

Manuscript has been submitted to Plant Breeding

Abstract

Genomic selection (GS) is a method used to predict the effects of molecular markers. The predicted marker effects are then summed to derive genomic breeding values for genotypes. GS has been routinely implemented in plant breeding in two stages. The first stage usually omits the marker information and estimates adjusted means of genotypes across environments. The second stage then uses the adjusted means to predict genomic breeding values. However, if the effects of markers vary substantially between different environments, it may important to account for this variation in genomic prediction for varieties adapted to different environments. Using two maize datasets, we investigated if modelling the marker-by-environment interaction can improve the predictive ability of GS relative to modelling genotype-by-environment interaction alone. Modelling the marker-by-environment interaction only slightly increased the predictive ability of GS relative to modelling only the genotype-by-environment interaction based on two different datasets. Moreover, predictive ability did not reduce substantially even when the number of markers with consistent effects across environments used for genomic prediction was reduced to about 50. Overall, accounting for environment-specific marker effects had a relatively minor influence on predictive ability for the tested datasets. Thus, GS carried out in a stagewise fashion, as is currently commonly done in plant breeding such that the marker information is omitted until the very last stage of the process, may suffice for most practical purposes.

6. Genomic selection using multiple populations

T. Schulz-Streeck[1,2], J. O. Ogutu[1], Z. Karaman[3], C. Knaak[2], and H. P. Piepho[1]

[1] *Bioinformatics Unit, Institute of Crop Science, University of Hohenheim,*
Fruwirthstrasse 23, 70599 Stuttgart, Germany

[2] *KWS SAAT AG, Grimsehlstraße 31, 37555 Einbeck, Germany*

[3] *Limagrain Europe, CS 3911, 63720 Chappes, France*

Crop Science 52:2453–2461

The original publication is available at https://www.crops.org/publications/cs

Abstract

Genomic selection (GS) is a marker-based method for predicting genomic breeding values that involves simultaneous estimation of the effects of many markers or chromosomal segments. Using different populations in GS studies raises the possibility that the marker effects may vary substantially across populations. Models for analysing datasets consisting of multiple populations need to include population-specific marker effects to account for such inter-population variation. However, common models for GS only account for the main marker effects, assuming that they are consistent across populations. Here, we present an approach in which the main plus population-specific marker effects are simultaneously estimated in a single mixed model. The predictive ability of the model were evaluated using 5-fold cross-validation and compared to that of ridge regression best linear unbiased prediction (RR-BLUP) method, involving only either the main marker effects or the population-specific marker effects. We used a maize breeding dataset with 312 testcross genotypes derived from five different biparental populations which were genotyped with a 50k SNP chip. A combined analysis incorporating all the populations was better than separate analyses for each population. Modelling the main plus the population-specific marker effects simultaneously only slightly improved predictive ability compared to modelling only the main marker effects, especially when the number of markers was reduced. The performance of the RR-BLUP method was comparable to that of two popular regularization methods, namely the ridge regression and the elastic net and was more accurate than that of the LASSO. Overall, combining information from related populations improved predictive ability, but further allowing for population-specific marker effects made minor improvement.

7. General discussion

This chapter discusses the key results of this thesis research and anchors them within the broad context of the current literature on genomic selection (GS). The chapter is organized around the following three major thematic areas. First the analyses preceding the marker-based prediction stage of genomic selection are discussed. This is followed by the statistical methods for marker-based prediction in GS and the discussion of the approaches for reliably evaluating and comparing methods for GS. The chapter ends with a summary of the key findings from this research.

Genomic selection is a method that uses molecular markers to predict genomic breeding values. Different methods have been proposed and implemented in an effort to get accurate predictions for GS. This thesis focuses on the use of mixed models, a popular class of models for GS, in particular ridge regression best linear unbiased prediction (RR-BLUP) and its various extensions. Comparative evaluations of the performances of RR-BLUP and its variants for GS using quantitative traits in maize breeding revealed only minor differences in accuracy. Regularized regression and machine learning methods, moreover, did not substantially improve predictive accuracy of GS relative to RR-BLUP or of its different variants. It is demonstrated, furthermore, that the accuracy of genomic prediction using molecular markers is influenced at different stages of the prediction process. This emphasizes the fact that, besides optimizing the marker-based prediction, the decisions made in the analyses preceding the marker-based prediction stage can significantly affect the quality of genomic predictions.

7.1. Analyses preceding the marker-based prediction stage of GS

The quality and reliability of phenotypic data are important to the accuracy of GS and should be enhanced as much as practicable using efficient field trial designs and phenotypic data analyses (Chapter 6). Recommended field trial designs should strive to test genotypes using replicated genotypes distributed over several representative locations and years (Kempton and Fox, 1997). Phenotypic analyses should attempt to identify and exclude potential outlying observations from datasets, as is routinely done in plant breeding studies (Fox et al., 1997). The analysed empirical datasets were based on augmented designs where the genotypes of interest were not replicated within a location and the dataset contained observations from one year only. Due to the lack of replicate observations on genotypes within locations, using the genotypes as fixed effects makes residual diagnostics for outlier detection impossible. However, the use of marker information makes additional diagnostic methods for the phenotypic data available for unreplicated field trials (Chapter 6). Alternative ways of improving the quality of phenotypic data include using designs with replicated genotypes, such as the α-design, or designs where certain proportions of the genotypes are replicated within each location (Piepho, 2006; Smith et al., 2006; Cullis et al., 2006; Williams et al., 2011) or testing in more locations.

The phenotypic analysis can be used to obtain an independent estimate of the residual error variance, which can be used as a fixed variance component at the GS stage to minimize overfitting (Piepho, 2009). If this is done, the markers do not capture the non-genetic portion of the error variance. Adopting such a strategy did not, however, evidently improve predictive accuracy for a simulated dataset prepared for the QTLMAS Workshop 2009 (Chapter 2). This may partly reflect the specific design used in the simulation, which did not include replicated genotypes. Hence only the within-genotype error variance could be estimated and not the between-genotype error variance. For most real plant breeding datasets, replicates are available, so that the non-genetic between-genotype error variance can be estimated and separated from the genetic effects. Another strategy to get an independent estimate for the error variance in GS is to combine the phenotypic and marker-based prediction stages into one analysis (Chapter 4). The error variance can then be estimated simultaneously with the variances of the other random effects in the model as part of the full variance-covariance structure of the observed data. The single-stage strategy produces an independent estimate of the error variance and avoids the need to fix the error variance at the GS stage but is computationally inefficient. A novel and computationally more efficient approach is to split the analysis into several stages, while minimizing the loss of information relative to a single-stage analysis (Chapter 4; Piepho et al., 2011). As expected, this produces results that are slightly more similar to those for the single-stage analysis than results for several other stage-wise methods that use less efficient weighting schemes to transfer the information contained in the variance-covariance matrix of the adjusted means between stages (Chapter 4).

High-quality genotypic data, besides high-quality phenotypic data, will also enhance predictive accuracy of GS. There are several strategies for improving the quality of genotypic data. Firstly, genotypes and markers with missing marker information and markers with too low minor allele frequencies can be identified and excluded from the dataset (e.g. Hayes et al., 2009a). Secondly, markers and genotypes with too many heterozygous loci can also be identified and deleted when using double haploid genotypes which are expected to be completely homozygous in maize breeding experiments. Thirdly, subsets of the most important markers that are most likely to be in high linkage disequilibrium (LD) with QTL and hence to affect the target quantitative trait can be pre-selected and used in GS. Pre-selection of markers can improve predictive accuracy (Hayes et al., 2009a), especially if the selected subset of markers have consistent effects across different populations (Chapter 3) but not necessarily in all circumstances (Weigel et al., 2009; Zhao et al., 2012). Indeed, even when a subset of markers with consistent effects across environments is selected, this may not always guarantee improved predictive ability, in particular when accuracy is assessed separately for individual locations using empirical maize datasets (Chapter 5). Therefore, it is always useful to carefully weight whether pre-selection of markers is warranted because it may not only improve the prediction accuracy, but may even reduce it (Zhao et al., 2012). These studies thus provide evidence that improvements in accuracy due to pre-selection of markers may vary with the dataset and analytical methods used, besides the particulars of the trait of interest. An important and practical limitation of pre-selection of markers is that there is no accepted objective threshold for deciding when to stop marker selection, nor is deciding the desirable threshold straightforward. A further limitation of strategies that set the effects of some markers to zero is the existence of a long-range LD for commercial maize breeding populations (Albrecht et al., 2011; Ching et al., 2002; Riedelsheimer et al., 2012). The existence of a long-range LD makes almost all markers to be in LD with at least one QTL and many markers to be coupled with the same QTL. Thus, it is most likely that each marker has an effect on the trait as discussed by Iwata and Jannink (2011). Due to the long-range LD even randomly selecting a small subset of markers for GS with maize datasets does not necessarily reduce prediction accuracy compared to using many markers (Chapter 6; Zhao et al., 2012). Therefore, for most practical applications in maize breeding, a small number of markers, often less than 1000, may suffice for GS, especially if biparental populations are used (Chapter 4; 5; 6; Lorenzana and Bernardo, 2009; Zhao et al., 2012).

Yet, another important consideration at this stage of GS is to carefully select an appropriate model for the variance-covariance matrix of the markers. Geostatistical models that are now widely used in plant breeding to analyze spatial trends in field trials offer a convenient and flexible means of doing this. Schabenberger and Gotway (2005) present several examples of theoretical semivariograms used in geostatistical models. These models carry over directly to GS and can be used to summarize the information contained in the empirical marker variance-covariance matrix as a function of genomic distances between pairs of genotypes (Chapter 2; 3; Piepho, 2009). Several theoretical models were first fitted to empirical semivariograms and graphical inspection used to select the model with the greatest strength of support in the data (Chapter 3). However, no model was selected as being clearly the best for the tested dataset using the semivariograms and hence all models produced similar marker-based prediction (Chapter 3).

7.2. Statistical methods for marker-based prediction in GS

7.2.1. Ridge Regression Best Linear Unbiased Prediction (RR-BLUP)

The basic RR-BLUP model yields the mixed model equations (MME) for the random marker effects and thus directly provides predictions for the random marker effects. However, if the number of markers is very large, solving the MME can become computationally too demanding or even impossible. Therefore, the basic RR-BLUP model can be rewritten in terms of random genotype effects and the genomic relationship matrix calculated from the markers (e.g. VanRaden, 2008; Hayes et al., 2009b; Piepho, 2009; Piepho et al., 2012). For the basic RR-BLUP method only a single variance component has to be estimated besides the residual error. Fast restricted maximum-likelihood (REML) algorithms for fitting the mixed models which replace complex matrix computations in REML estimation of variance components with arithmetic operations on scalars instead of matrices are now readily available (Kang et al., 2008; Endelman, 2011; Piepho et al., 2012).

In plant breeding experiments an independent estimate of the residual variance or the variance-covariance of the residual error is often available from the analysis that yielded adjusted means. This can be used as a fixed term in the RR-BLUP model (Chapter 2; 4; Piepho, 2009; Piepho et al., 2011). The computationally efficient algorithm for fitting the mixed model of Kang et al. (2008) can be extended for the case of a known residual variance (Piepho et al., 2012). This approach is restricted to the case of independent residual errors. However, the assumption of independent errors can be restrictive, in particular for plant breeding experiments. In plant breeding experiments the error variances are often heterogeneous or experimental designs are unbalanced and hence the variance-covariance matrix of the adjusted genotype means can be complex. The restriction of independent residual errors can be lifted by rotating (orthogonalizing) the adjusted means to ensure independence of the rotated residual errors (Chapter 4). Thus, the mixed model of Kang et al. (2008) can now be generally used with residual errors with arbitrary variance-covariance structures (Chapter 4; Piepho et al., 2012). The implementation of the rotation approach for the mixed model of Kang et al. (2008), extended to apply to the special case with fixed residual error variance, is available as an R package **rrBlupMethod6** in the Comprehensive R Archive Network (Piepho et al., 2012; Schulz-Streeck et al., 2012).

Computationally efficient methods for GS are especially important when cross-validation is used for the evaluation of methods. However, when using a very large number of markers the computation of the genomic relationship matrix dominates the computing time compared to solving the mixed model equations. The matrix multiplications can be optimized for computing time using optimized ‘do’ loops, within specific matrix multiplication subroutines and parallel processing (Aguilar et al., 2011). Additionally, the matrix multiplications can be accelerated by processing the matrix with the marker information in parts (Piepho et al., 2012). Given the efficient algorithms available for implementing it and its relatively high predictive accuracy on many empirical maize breeding datasets (Chapter 4; 5; Crossa et al., 2010; Albrecht et al., 2011; Zhao et al., 2012; Heslot et al., 2012; Riedelsheimer et al., 2012). RR-BLUP would seem an attractive and well-known and tested method for GS in maize breeding. However, the performance of RR-BLUP may vary considerably between populations (Chapter 4; 5; 6; Albrecht et al., 2011; Heslot et al., 2012) and some populations may show relatively low prediction accuracies (Chapter 6). It is yet not definitively known whether this variation in accuracy across populations reflects intrinsic differences between the populations themselves or is an inherent shortcoming of RR-BLUP. Comparisons of the performance of RR-BLUP with those of other methods on the same populations suggest that the former is the likelier of the two possibilities.

7.2.2. Extensions of the RR-BLUP model

The basic mixed model used for GS assumes homogenous variance for the markers (Meuwissen et al., 2001). Different methods have been proposed for GS to relax this strong and restrictive assumption, including BayesA and BayesB. However, different variance components for the markers can also be fitted in a mixed model framework to account for variance heterogeneity (Chapter 3; Meuwissen, 2009, Piepho, 2009). The effect of accounting for variance heterogeneity on predictive accuracy of GS has ranged from slight (Chapter 3; Pszczola et al., 2011) to substantial improvement in accuracy using simulated datasets (Meuwissen et al., 2001; Habier et al., 2007). Using empirical data the differences are negligible in maize breeding (Albrecht et al., 2011; Crossa et al., 2010; Heslot et al., 2012) and animal breeding in the majority of cases (Hayes et al., 2009c; Verbyla et al., 2009; Habier et al., 2011). If large-effect QTL are known to exist for a certain trait, markers linked to these QTL may be modelled with a different variance. In animal breeding using traits with known large effect QTL models allowing individual marker variances, like BayesA, fit these large effects better than models assuming homogenous marker variance, like BLUP methods, where all marker effects are shrunk equally (Verbyla et al., 2009). However, doing so will also not guarantee improvement in predictive accuracy using other traits (Hayes et al., 2009c; Verbyla et al., 2009). Additionally, the markers linked to QTL with known larger effects can be modelled as fixed effects using a mixed model in the spirit of association mapping methods. Moreover, proposals have been made to use the trait-specific relationship matrix in RR-BLUP models based on a small number of simulated QTL to improve accuracy of GS, but this improvement typically decreases as the number of simulated QTL increases (Zhang et al., 2011).

The estimated marker effects can differ between environments in maize breeding (Crossa et al., 2010). However, accounting for environment-specific marker effects had a relatively minor influence on predictive ability for the tested datasets in this thesis. Thus, GS carried out in a stagewise fashion may suffice for most practical purposes, such that the marker information can be omitted until the very last stage of the process. In the case where substantial differences between environments are known the focus of a breeder may shift to predicting the performance of a genotype in a specific subregion. This subregion should be represented by a sample of environments. In this case allowing for subregion-specific marker effects in different environments may enhance accuracy. Moreover, estimated marker effects can differ between populations in maize breeding (Lui et al., 2011). In contrast, analyzing each population separately can be less accurate than a combined analysis that exploits information from related populations (Chapter 6; Albrecht et al., 2011; Jannink et al., 2010; Riedelsheimer et al., 2012; Zhao et al., 2012). It is feasible to account for the main and population-specific marker effects in the same model, but the improvement from using this model compared to modelling only main marker effects was negligible (Chapter 6). Using multiple populations and evaluating the accuracy of GS separately for each population yielded prediction accuracies that differed substantially among the populations, with some populations showing notably low predictive accuracies, similarly to findings of Heslot et al. (2012) and Albrecht et al. (2011). If only some of the populations have high predictive accuracies, this may lead to preferential selection of those populations and loss of diversity represented by the other populations as discussed by Heslot et al. (2012). Weighting the overall selection over the between and within-population selection may remedy this shortcoming (Jannink et al., 2010).

Polygenetic effects can be added to the RR-BLUP model to capture genetic variance not captured by the markers (Chapter 2; Calus and Veerkamp, 2007; Piepho, 2009; Albrecht et al., 2011). Polygenic effects are often modeled by the relationship matrix but can also be estimated using independent effects (Chapter 2; Piepho, 2009). The resulting model can be further extended to include simple random effects of the male and female parent and the crosses themselves (Chapter 2). Modeling polygenic effects is especially useful when using a small number of markers (Calus and Veerkamp, 2007), but merits caution as it may decrease accuracy (Legarra et al., 2008).

The RR-BLUP model is equivalent to a quadratic spatial model (Piepho, 2009). Different other spatial mixed models have been tested on simulated and real datasets but the differences in prediction accuracy among these models and with RR-BLUP have been minor (Chapters 2; 3; Piepho, 2009; Ober et al., 2011). Ober et al. (2011) found that using the Matérn function to model the covariance as a function of the genomic distance between pairs of genotypes can improve accuracy compared to a particular variant of RR-BLUP, called G-BLUP. However, RR-BLUP and its different spatial variants showed higher accuracies than the spatial Matérn model for empirical maize datasets (Appendix A). Thus, the Matérn model does not always perform better than the standard RR-BLUP model. Moreover, the Matérn model is a generalization of several special models, including the Gaussian model, which can converge to the quadratic spatial model under certain circumstances (Chapter 2; Piepho, 2009).

7.2.3. Machine learning and regularized regression methods

The tested machine learning and regularized regression methods did not improve the prediction accuracy of GS relative to RR-BLUP (Chapters 4 and 6) similar to the findings of Iwata and Jannink (2011), Ogutu et al. (2011, 2012) and Heslot et al. (2012). But using a rather small number of simulated QTL, González-Recio and Forni (2011) reported examples in which machine learning methods outperformed Baysian methods.

Overall, the basic RR-BLUP method gave essentially similar results to all its tested extensions and the other alternative methods. These findings imply that RR-BLUP can commonly be used for GS in maize breeding programs. Moreover, the good performance of RR-BLUP was, surprisingly, not restricted only to datasets with many markers and small effects and was similar to that for regularized regression methods or Baysian methods even on datasets with only a few simulated QTL (Pszczola et al., 2011; Ogutu et al., 2012). However, it is still not conclusively established whether the prediction accuracy of RR-BLUP depends primarily on the relationship among genotypes, or on the LD between markers and QTL. If it depends primarily on the relationship among genotypes, then prediction accuracy will decrease fast whereas inbreeding will increase over generations (Habier et al., 2007; Dekkers, 2007). Habier et al. (2007) suggested that the accuracy of RR-BLUP depends more strongly on the relationship among genotypes. Thus, the strength of the relatedness of the genotype functions is similar to the pedigree information (Hayes et al., 2009b; Piepho, 2009). In contrast, the accuracy of some Bayesian methods (e.g. BayesB) depends more sensitively on the LD between markers and QTL, leading to higher long-term gains in selection performance when GS is done using BayesB than using RR-BLUP (Habier et al., 2007).

While datasets covering only one year are used in this thesis, real datasets covering many generations would be needed to empirically confirm if the long-term gains from GS suggested by simulation studies really do materialize. Moreover, weighting low-frequency but favorable alleles so that favorable alleles that are in weak LD with markers are not lost could enhance long-term gain from GS (Jannink, 2010).

Another assumption made in using RR-BLUP is that the marker effects are random draws from a common normal distribution. However, in reality the true distribution is unknown and may differ among traits. Accordingly, different distributions have been proposed for use with Bayes methods (e.g. the double exponential distribution for the Baysian LASSO: Yi and Xu, 2008). Additionally, the distribution of the marker effects may be a mixture of different component distributions (Bennewitz and Meuwissen, 2010; Toro and Varona, 2010). It follows that several methods that do not make stringent distributional assumptions about the marker effects (e.g. boosting, random forest, support vector machine and reproducing kernel Hilbert spaces regression) may therefore be potentially useful for GS (e.g. Gianola et al., 2006; Ogutu et al., 2011).

Since the focus of most GS studies is to evaluate the prediction accuracy of statistical methods for estimating genomic breeding values, only additive genetic effects were considered in this thesis. However, to predict the hybrid performance in maize, dominance and epistatic effects may be important and merit consideration besides the additive genetic effects. This would require the testcross genotypes to be genotyped. A variety of methods are available for capturing complex epistatic interactions, including random forest (Breiman, 2001; Statnikov et al., 2008). Dominance (Xu, 2003; Bennewitz and Meuwissen, 2010) or epistatic (Xu and Jia, 2007) marker effects can also be estimated within the linear model framework.

7.3. Approaches for reliably evaluating and comparing methods for GS

Cross-validation (CV) is an omnibus method for performing model selection which has been widely used to evaluate the prediction accuracy of different methods for GS (e.g. Villumsen and Janss, 2009; Crossa et al., 2010; Erbe et al., 2010; Habier et al., 2010; Albrecht et al., 2011). One popular CV procedure is the *k*-fold CV method, in which the dataset is split into *k* subsets, *k*-1 of which are concatenated and used as the training set to select a model and estimate coefficients of the predictor variables, and the *k*th, called the validation set, to validate the selected model. Large simulation studies suggest that the optimal number of splits of the dataset *k* ranges between five and ten (Hastie et al., 2009). The sizes of the training and validation sets affect prediction accuracy in GS. For example, higher accuracies have been obtained for larger training and smaller validation sets (Erbe et al., 2010) and higher variances for smaller validation sets (Lee et al., 2008;

Erbe et al., 2010). Erbe et al. (2010) therefore recommend using fivefold CV in GS as a practical compromise between prediction accuracy and variance.

The degree of relatedness among the genotypes in the training and validation sets also influences the prediction accuracy such that the accuracy of genomic estimated breeding values decreases the more closely related the genotypes are (Habier et al., 2010). Habier et al. (2010) have therefore proposed controlling the additive-genetic relationship among the genotypes in the training and the validation sets. For plant breeding experiments it is also common to account for the degree of relatedness among genotypes in the training and validation sets by using genotypes from the same populations in the training and validation sets, or by using genotypes of from a new population in the validation set (Chapter 3; Albrecht et al., 2011).

An assumption integral to the proper conduct of the *k*-fold CV is that the errors are indepent and identically normally distributed (i.i.d) and hence that the training and validation sets are independent (Arlot and Celisse, 2010). For plant breeding experiments adjusted means are often used for GS and thus for CV. When using unbalanced trial designs, the adjusted means are not independent and thus the basic assumptions of i.i.d. errors are not fulfilled. Rotating the means can satisfy this assumption (Chapter 4). Another way to ensure independence of the validation and the training dataset is to use the dataset for the year following the one in which the training dataset was collected as the validation set. However, for evaluating the long-term gain from GS, it is preferable to have a validation dataset from several subsequent generations.

The computational burden of CV can be very high for some models and datasets. One option to reduce this is to use computationally more efficient procedures for GS (Kang et al., 2008; VanRaden, 2008; Piepho et al., 2012). Another option is to replace the computationally more demanding CV with model selection criteria for GS (Chapter 2; 5). One commonly used model selection criterion, the Akaike information criterion (AIC), often produces nearly-identical rankings of different models to the rankings based on correlations between the observed and the true breeding values, but not always (Chapters 2; 4). Therefore, other model selection criteria that have been proposed, in the literature, especially in connection with smoothing methods, including the conditional AIC (Hastie and Tibshirani, 1990), corrected AIC (Hurvich et al., 1998) and the generalized cross-validation criterion (Craven and Wahba, 1979) and may be explored as potential alternatives to AIC for GS (Chapter 2). Using the simulated dataset prepared for the QTLMAS 2010 workshop (Chapter 3) results of model selection using AIC and the generalized cross-validation were nearly identical with those based on CV, implying that both AIC and generalized cross-validation may be used instead of cross-validation to reduce computing time (Appendix B). However, using an empirical dataset in Chapter 4 showed that the AIC-selected best model did not translate into clearly better prediction ability based on cross-validation. Further tests would be needed to conclusively determine whether the computationally more demanding CV may be replaced with the more efficient model selection criteria without loss of accuracy.

7.4. Conclusions

Genomic selection is a recent, robust and promising approach for integrating information from many molecular markers spanning the whole genome through statistical models for high-dimensional data to predict genotypic values of untested genotypes in plant and animal breeding programs. It is of vast economic interest and great practical appeal because it uses computationally efficient and readily available statistical methods and hence holds great promise of establishing itself as a routine feature of cost-effective breeding programs. For plant breeding it is perhaps most efficient to implement GS in a stagewise fashion, if the loss of information in a stagewise relative to a single-stage approach can be minimized. For most practical purposes, when using stagewise approaches, it may suffice to omit the marker information until at the very last stage, if the marker-by-environment interaction has a only minor influence, as found in the datasets considered in this thesis. Pre-selection of markers is not of much practical relevance in the vast majority of cases in maize breeding, thus greatly simplifying the steps involved in GS. To achieve high prediction accuracy in GS, it is imperative to ensure that both the phenotypic and genotypic data are of reasonably high-quality by using appropriate field trial designs and carrying out adequate quality controls to detect and eliminate observations deemed to be outlying. Further improvement in accuracy may be gained by combining genotypes from different populations into one analysis instead of conducting separate analyses for each population. The widely used and tested ridge regression best linear unbiased prediction method would seem adequate for GS for most practical purposes, thus obviating the need to apply the more complex alternatives, such as the spatial and regularized regression, or machine learning methods.

References

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Albrecht, T., V.Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C.C. Schön. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123:339–350.

Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4:40–79.

Bennewitz, J., and T.H.E. Meuwissen. 2010. The distribution of QTL additive and dominance effects in porcine F2 crosses. J. Anim. Breed. Genet. 127:171-179.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

Calus, M.P.L., and R.F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124:362–368.

Ching, A., K.S. Caldwell, M. Jung, M. Dolan, O. S. Smith, S. Tingey, M. Morgante, and A.J. Rafalski. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3:19.

Craven, P., and G. Wahba. 1979. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377-403.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, G. Atlin, J. Burgueño, J.L. Araus, D. Makumbi, J. Yan, V. Arief, M. Banziger, and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agr. Biol. Envir. St. 11, 381–393.

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331–341.

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

Erbe, M., E.C.G. Pimentel, A.R. Sharifi, and H. Simianer. 2010. Assessment of cross-validation strategies for genomic prediction in cattle. Proceedings of 9th World Congress on Genetics applied to Livestock Production (WCGALP), 1.-6. August 2010, Leipzig, Germany.

Fox, P.N., R. Mead, M. Talbot, and J.D. Corbett. 1997. Data management and validation. In: Kempton, R.A., and P.N. Fox. 1997. Statistical methods for plant variety evaluation. Chapman and Hall, London.

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

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43:7.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. Genet. Sel. Evol. 42:5.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hastie, T., and R. Tibshirani. 1990. Generalized additive models. Chapman and Hall, London.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Data mining, inference, and prediction. Springer Series in Statistics. Springer-Verlag, New York. 2nd edition.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Hayes, B.J., P.M. Visscher, and M.E.Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Res. 91:47–60.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M.E. Goddard. 2009c. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51.

Heslot, N., H.P. Yang, M.E. Sorrels, and J.L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52:146-160.

Hurvich, C.M., J.S. Simonoff, and C. Tsai. 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society Series B 60:271-93.

Iwata, H., and J.L. Jannink. 2011. Accuracy of genomic selection prediction in barley breeding programs: A simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Sci. 51: 1915-1927.

Jannink, J.L. 2010. Dynamics of long-term genomic selection. Genet. Sel. Evol. 42:35.

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomic Proteomic 9: 166–177.

Kang, H.M., N.A. Zaitlin, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1725.

Kempton, R.A., and P.N. Fox. 1997. Statistical methods for plant variety evaluation. Chapman and Hall, London.

Lee, S.H., J.H.J. 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 Genet. 4(10):e1000231.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618.

Liu, W., M. Gowda, J. Steinhof, H.P. Maurer, T. Würschum, C.F.H. Longin, F. Cossic, and J.C. Reif. 2011. Association mapping in an elite maize breeding population. Theor. Appl. Genet. 123:847–858.

Lorenzana, R., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161.

Meuwissen, T.H.E. 2009. Accuracy of breeding values of ‘unrelated’ individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41(35).

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

Ober, U., M. Erbe, N. Long, E. Porcu, M. Schlather, and H. Simianer. 2011. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188:695-708.

Ogutu, J.O., H.P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection using SNP markers. BMC Proc. 5 (Suppl 3):S11.

Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions (accepted).

Pszczola T. Strabel, A. Wolc, S. Mucha, and M. Szydlowski. 2011.Comparison of analyses of the QTLMAS XIV common dataset. I: genomic selection. BMC Proc. 5(Suppl 3):S1.

Piepho, H.P., A. Büchse, and B. Truberg. 2006. On the use of multiple lattice designs and α-designs in plant breeding trials. Plant Breed. 125:523–528.

Piepho, H.P. 2009. Ridge regression and extensions for genome-wide selection in maize. Crop Sci. 49:1165-1176.

Piepho, H.P., T. Schulz-Streeck, and J.O. Ogutu. 2011. A stage-wise approach for analysis of multi-environment trials. Biuletyn Oceny Odmian 33:7-20.

Piepho, H.P., J.O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow. 2012. Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci. doi: 10.2135/cropsci2011.11.0592.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A.E. Melchinger. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Geneti. doi:10.1038/ng.1033.

Schabenberger, O., and C.A. Gotway. 2005. Statistical methods for spatial data analysis. CRC Press Boca Raton, FL.

Schulz-Streeck, T., B. Estaghvirou, and F. Technow. 2012. rrBlupMethod6: Re-parametrization of RR-BLUP to allow for a fixed residual variance. R package, version 1.2. Available at http://cran.r-project.org/web/packages/rrBlupMethod6/index.html

Smith, A.B., P. Lim, and B.R. Cullis. 2006. The design and analysis of multi-phase plant breeding experiments. J. Agr. Sci. 144:393–409.

Statnikov, A., L. Wang, and C.F. Aliferis. 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9:319-324.

Toro, M.A., and L. Varona. 2010. A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42:33.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy. Sci. 91:4414-4423.

Verbyla, K.L., B.J. Hayes, P.J. Bowman, and M.E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genetics Research 91(05):307-311.

Villumsen, T.M., and L. Janss. 2009. Bayesian genomic selection: the effect of haplotype length and priors. BMC Proc. 3(Suppl 1):S11.

Weigel, K.A., G. de los Campos, O. González-Recio, H. Naya, X.L. Wu, N. Long, G.J. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248-5257.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801.

Xu, S., and Z. Jia. 2007. Genomewide analysis of epistatic eff ects for quantitative traits in barley. Genetics 175:1955–1963.

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

Zhang, Z., J.F. Liu, X.D. Ding, P. Bijma, D.J. de Koning, and Q. Zhang. 2010. Best linear unbiased prediction of genomic breeding values using trait-specific marker-derived relationship matrix. PLoS ONE 5(9):e12648.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, and J.C. Reif. 2011. Accuracy of genomic selection in European maize elite breeding Populations. Theor. Appl. Genet. 124:769–776.

8. Summary

Genomic selection (GS) is a new approach for integrating information from many molecular markers spanning the whole genome in the context of plant and animal breeding. Its main aim is to predict genotypic or breeding values for non-phenotyped genotypes using statistical models for high-dimensional data. This approach both expedites and increases the cost-effectiveness of breeding programs because not all genotypes have to be empirically tested in field trials. Thus, GS can considerably increase the gain per unit time. It is crucial that the genomic predictions are accurate. Many approaches have therefore been proposed to enhance predictive accuracy of GS, including mixed models, Bayesian, machine learning and regularized regression methods. All these methods can handle the problem of using more predictor variables than number of observations in the model. This problem is common because by using high density markers for GS, the number of markers often far exceeds the number of observations. Cross-validation is often used to reliably evaluate the relative predictive accuracies of contending approaches. In plant breeding, genotypes are often tested in different environments using several replicates per environment and thus yielding replicate observations per genotype. As a result, GS is normally undertaken in a stagewise fashion. The first stage, called the phenotypic analysis stage, involves computing adjusted means for genotypes across environments. These adjusted means are then used together with the molecular markers in the second stage to estimate the effects of markers, called here the marker-based analysis. The estimated marker effects can then be used to predict genomic breeding values for untested genotypes.

In the current thesis the efficacy of several contending approaches for GS were tested using different simulation and empirical maize breeding datasets. Here, GS is viewed as a general approach, incorporating all the different stages from the phenotypic analysis of the raw data to the marker-based prediction of the breeding values. The overall goal of this study was to develop and comparatively evaluate different approaches for accurately predicting genomic breeding values in GS. In particular, the specific objectives were to:

(1) Develop different approaches for using information from analyses preceding the marker-based prediction of breeding values for GS.

(2) Extend and/or suggest efficient implementations of statistical methods used at the marker-based prediction stage of GS, with a special focus on improving the predictive accuracy of GS in maize breeding.

(3) Compare different approaches to reliably evaluate and compare methods for GS.

An important step in the analyses preceding the marker-based prediction is the phenotypic analysis stage. One way of combining phenotypic analysis and marker-based prediction into a single stage analysis is presented in Chapter 4. However, a stagewise analysis is typically computationally more efficient than a single stage analysis. Several different weighting schemes for minimizing information loss in stagewise analyses are therefore proposed and explored (Chapter 2 and 4). It is demonstrated that orthogonalizing the adjusted means before submitting them to the next stage is the most efficient way within the set of weighting schemes considered (Chapter 4). Furthermore, when using stagewise approaches, it may suffice to omit the marker information until the very last stage, if the marker-by-environment interaction has only a minor influence, as was found to be the case for the datasets considered in this thesis (Chapter 5). It is also important to ensure that genotypic and phenotypic data for GS are of sufficiently high-quality. This can be achieved by using appropriate field trial designs and carrying out adequate quality controls to detect and eliminate observations deemed to be outlying based on various diagnostic tools (Chapter 6). Moreover, it is shown that pre-selection of markers is less likely to be of high practical relevance to GS in most cases (Chapter 3 and 5). Furthermore, the use of semivariograms to select models with the greatest strength of support in the data for GS is proposed and explored. It is shown that several different theoretical semivariogram models were all well supported by an example dataset and no single model was selected as being clearly the best (Chapter 3).

Several methods and extensions of GS methods have been proposed for marker-based prediction in GS. Their predictive accuracies were similar to that of the widely used ridge regression best linear unbiased prediction method (RR-BLUP). It is thus concluded that RR-BLUP, spatial methods, machine learning methods, such as componentwise boosting, and regularized regression methods, such as elastic net and ridge regression, have comparable performance and can therefore all be routinely used for GS for quantitative traits in maize breeding (Chapter 2, 3, 4 and 6). Accounting for environment-specific or population-specific marker effects had only minor influence on predictive accuracy (Chapter 5 and 6) contrary to findings of several other studies. However, accuracy varied markedly among populations, with some populations showing surprisingly very low levels of accuracy (Chapter 6). Combining different populations prior to marker-based prediction improved prediction accuracy compared to doing separate population-specific analyses (Chapter 6). Moreover, polygenetic effects can be added to the RR-BLUP model to capture genetic variance not captured by the markers. However, doing so yielded minor improvements, especially for high marker densities (Chapter 2 and 6). To relax the assumption of homogenous variance of markers, the RR-BLUP method was extended to accommodate heterogeneous marker variances but this had negligible influence on the predictive accuracy of GS for a simulated dataset (Chapter 3).

The widely used information-theoretic model selection criterion, namely the Akaike information criterion (AIC), ranked models in terms of their predictive accuracies similar to cross-validation in the majority of cases (Chapter 2 and 5). But further tests would be required to definitively determine whether the computationally more demanding cross-validation may be substituted with the more efficient model selection criteria, such as AIC, without much loss of accuracy.

Overall, a stagewise analysis, in which the markers are omitted until at the very last stage, is recommended for GS for the tested datasets. The particular method used for marker-based prediction from the set of those currently in use is of minor importance. Hence, the widely used and thoroughly tested RR-BLUP method would seem adequate for GS for most practical purposes, because it is easy to implement using widely available software packages for mixed models and it is computationally efficient.

9. Zusammenfassung

Die genomweite Selektion stellt einen neuartigen Ansatz dar, um genomweite Markerinformationen, für züchterische Zwecke, sowohl in der Pflanzen- als auch in der Tierzüchtung, zu verwenden. Das Ziel dabei ist die Vorhersage von genetischen Werten oder Zuchtwerten für nicht phänotypisierte Genotypen, wobei statistische Methoden für hochdimensionale Daten verwendet werden. Da durch diesen Ansatz nicht alle Genotypen empirisch in Feldversuchen getestet werden müssen, können Zuchtprogramme somit sowohl beschleunigt als auch kosteneffizienter gestaltet werden, und der Gewinn kann pro Zeiteinheit entscheidend gesteigert werden. Eine genaue Vorhersage hat dabei höchste Priorität. Für diese genaue genomische Vorhersage wurden in der Vergangenheit verschiedene Methoden vorgeschlagen. Dabei spielen unter anderem gemischte Modelle, Bayes-Verfahren, „Machine Learning“ und „regularized regression” Methoden eine große Rolle. All diese genannten Verfahren sind in der Lage mit dem Problem umzugehen, dass mehr Effekte im Modell geschätzt werden als Beobachtungen gegeben sind, welches sich dadurch ergibt, dass bei der Verwendung genomweiter Markerdaten die Anzahl der Marker die der phänotypischen Beobachtungen übersteigen kann. Um die Vorhersagegenauigkeit der verschiedenen Verfahren zur genomweiten Vorhersage zu vergleichen, wird in den meisten Fällen die sogenannte Kreuzvalidierung verwendet. In der Pflanzenzüchtung werden die Genotypen meistens in verschiedenen Umwelten und in jeder Umwelt häufig mit mehreren Wiederholungen getestet, so dass für jeden Genotyp mehrere Beobachtungen vorliegen. Um diese wiederholten Beobachtungen in der genomweiten Vorhersage zu berücksichtigen, wird bei den meisten Methoden schrittweise vorgegangen. Zunächst werden hierbei adjustierte Mittelwerte der Genotypen über die verschiedenen Umwelten berechnet. Diese Analyse wird auch phänotypische Analyse genannt. Die adjustierten Mittelwerte werden im nächsten Schritt verwendet, um Effekte für die Marker zu schätzen. Dieser Schritt wird hier als markerbasierende Analyse bezeichnet. Die geschätzten Markereffekte können dann weiter verwendet werden, um für nicht phänotypisierte Genotypen eine Vorhersage der genomischen Zuchtwerte zu schätzen.

In der vorliegenden Arbeit wurde die Effektivität verschiedener Methoden der genomweiten Selektion untersucht. Hierbei wurden sowohl simulierte Datensätze als auch mehrere reale Datensätze aus der Maiszüchtung verwendet. Die genomweite Selektion wird in dieser Arbeit als ein Verfahren angesehen, welches alle Schritte der genomweiten Selektion, von der Analyse der phänotypischen Rohdaten bis zur markerbasierenden Vorhersage der Zuchtwerte, einschließt. Das Ziel der Arbeit ist es, verschiedene Verfahren auf ihre Vorhersagegenauigkeit von genomischen Zuchtwerten zu bewerten. Die folgenden Zielstellungen wurden im Speziellen behandelt.

(1) Entwicklung verschiedener Verfahren zur Einbindung von Informationen, die vor der marker-basierenden Analyse gewonnen werden, in die genomweite Selektion.

(2) Erweiterung und/oder Empfehlung der effizienten Implementierungen von statistischen Methoden zur marker-basierenden Analyse, wobei im Speziellen die Vorhersagegenauigkeit der genomweiten Selektion in der Maiszüchtung verbessert werden soll.

(3) Vergleich verschiedener Ansätzen zur Beurteilung und zum Vergleich der Güte der Methoden zur genomweiten Selektion.

Ein wichtiger Schritt in den Analysen, die vor der markerbasierenden Analyse stattfinden, ist die Analyse der phänotypischen Daten. Ein Weg um diese Analyse mit der marker-basierenden Analyse in einem einstufigen Verfahren zu kombinieren, wurde in Kapitel 4 gezeigt. Jedoch ist ein schrittweises Vorgehen weniger rechenintensiv, als wenn beide Analysen in einem Schritt kombiniert werden. Deshalb wurden mehrere Gewichtungsansätze für die genomweite Selektion vorgeschlagen, um den Informationsverlust des schrittweisen Verfahrens zu minimieren (Kapitel 2 und 4). Es wurde gezeigt, dass es, im Vergleich mit anderen Gewichtungsansätzen, am effizientesten ist, die adjustierten Mittelwerte nach jedem Analyseschritt zu orthogonalisieren (Kapitel 4). Des Weiteren kann es ausreichend sein, bei diesen schrittweisen Ansätzen die Markerinformation bis zum letzten Schritt zu ignorieren, wenn die Marker × Umweltinteraktion gering ist, wie es in den getesteten Datensätzen der Fall war (Kapitel 5). Weiterhin konnte herausgestellt werden, dass eine hohe Qualität sowohl der genetischen als auch der phänotypischen Daten wichtig ist. Dieses kann erreicht werden, wenn entsprechende Feldversuchsdesigns und geeignete Diagnosemethoden zur Qualitätskontrolle verwendet werden, um Beobachtungen, die außerhalb des erwarten Spektrums liegen, zu entfernen (Kapitel 6). Die Vorselektion von Markern hingegen war in den meisten Fällen nicht von Relevanz für die praktische Anwendung (Kapitel 3 und 5). Außerdem wurde die Verwendung von Semivariogrammen vorgeschlagen und untersucht, um Modelle für die genomweite

Selektion zu ermitteln, die an gegebene Daten am Besten angepasst sind. Es wurde gezeigt, dass die verschiedenen theoretischen Semivariogrammmodelle an die getesteten Daten gut angepasst waren und kein Modell als entscheidend besser zu bewerten war (Kapitel 3).

Mehrere Methoden und Erweiterungen von genomweiten Selektionsmethoden wurden für die markerbasierende Vorhersage vorgeschlagen. Deren Vorhersagegenauigkeiten waren ähnlich zu der häufig verwendeten „ridge regression best linear unbiased prediction“ Methode (RR-BLUP). Somit konnte gezeigt werden, dass RR-BLUP, räumliche Modelle, „machine learning“ Methoden, wie „componentwise boosting“ und „regularized regression“ Methoden, wie „elastic net“ und „ridge regression“ gleichwertige Vorhersagegenauigkeiten zeigen und gleichberechtigt für routinemäßig Anwendung für die genomweite Selektion für quantitative Merkmale in der Maiszüchtung eingesetzt werden können (Kapitel 2, 3, 4 und 6). Im Gegensatz zu Ergebnissen anderer Studien zeigten Erweiterungen mit umweltspezifischen oder populationsspezifischen Markereffekten nur einen geringen Einfluss (Kapitel 5 und 6). Die Genauigkeit der Vorhersage kann sich aber zwischen verschiedenen Populationen stark unterscheiden. Einige Populationen zeigten dabei sehr geringe Vorhersagegenauigkeiten auf (Kapitel 6). Eine Analyse, in der mehrere Populationen simultan verwendet wurden, verbesserte die Vorhersagegenauigkeit gegenüber einer Analyse in der jede Population einzeln ausgewertet wurde (Kapitel 6). Außerdem kann die Methode RR-BLUP um polygenetische Effekte erweitert werden, um die genetische Varianz, die nicht von den Markern erfasst wird, zu berücksichtigen. Dieses zeigte aber nur eine geringe Verbesserung insbesondere bei hohen Markerdichten (Kapitel 2 und 6). Es wurde weiterhin eine Erweiterung der RR-BLUP Methode vorgeschlagen, um auf die Annahme einer homogenen Markervarianz verzichten zu können. Hierbei wurden heterogene Markervarianzen im gemischten Modell vorgeschlagen. Für einen simulierten Datensatz hatte dieses aber nur geringe Auswirkungen (Kapitel 3).

Das häufig verwendete Modellselektionskriterium „Akaike information criterion“ (AIC) zeigte in den meisten Fällen ähnliche Ergebnisse in der Beurteilung der genomweiten Selektionsmethoden wie die standardmäßig verwendet Kreuzvalidierung (Kapitel 2 und 5). Es sind aber weitere Tests notwendig, um grundlegend zu klären, ob die rechenintensive Kreuzvalidierung mit den effizienteren Modellselektionskriterien, wie zum Beispiel dem AIC, ersetzt werden kann, ohne dass ein Genauigkeitsverlust erfolgt.

Auf Grundlage der analysierten Daten kann in den meisten Fällen eine schrittweise Analyse empfohlen werden, wobei die Marker erst im letzten Schritt berücksichtigt werden müssen. Die zu verwendende Methode für die markerbasierende Vorhersage ist von geringerer Bedeutung, weshalb in den meisten Fällen die RR-BLUP Methode für die genomweite Vorhersage empfohlen werden kann, da diese einfach mit gängigen Software zur Analyse gemischter Modelle zu implementieren ist und rechenzeiteffizient ist.

10. Appendix

Appendix A

The dataset used here is the same as that used in Chapter 4 and 5. A fivefold cross-validation with five replicates was used to evaluate the predictive ability of the RR-BLUP and the four different spatial models. The predictive ability was calculated as the Pearson correlation between predicted values and adjusted means of the validation set using cross-validation.

Table A1: Predictive ability of different genetic covariance models. The predictive ability was calculated as the Pearson correlation between predicted values and adjusted means of the validation set using a fivefold cross-validation.

Model	Predictive ability	
	Dataset A	Dataset B
RR-BLUP	0.7164	0.4858
Exponential spatial model	0.7161	0.4956
Gaussian spatial model	0.7157	0.4756
Spherical spatial model	0.7161	0.4949
Matérn spatial model	0.7157	0.4765

Appendix B

The dataset used here is the same as that used in Chapter 3. A detailed description of the dataset can also be found in Szydlowski and Paczyńska (2011). Different genetic covariance models were compared using a fivefold cross-validation, an independent dataset with known true breeding values but lacking of phenotypic values and different model selection criteria. The Prediction ability was calculated as the between predicted values and observed values of the validation sets using the CV. The prediction accuracy was calculated as the Pearson correlation between predicted values and true genetic values for non-phenotyped individuals.

Table B1: Selection of different genetic covariance models using information criteria (Akaike Information Criterion (AIC), conditional AIC (cAIC), corrected AIC (AICc) and generalized cross-validation criterion (GCV), Pearson correlation between GEBVs and observed values in the validation sets (CV), and Pearson correlation between genomic estimated breeding values and true breeding values for non-phenotyped individuals (TBV). The best three methods for each criterion are printed in boldface.

					Correlation	
Model	AIC	cAIC	AICc	GCV	CV	TBV
Ridge Regression BLUP (RR)						
RR	16432	16205	11.91	145871	0.530	0.607
Pre-selection of SNPs (method 2[$])						
RR (500 markers)	16312	16184	11.88	143422	0.570	0.599
RR (1000 markers)	**16275**	16124	**11.86**	139965	**0.583**	**0.623**
RR (2000 markers)	16295	16124	**11.86**	140199	0.579	**0.625**
RR (3000 markers)	16312	16130	11.87	140726	0.576	0.617
Gaussian spatial model (GAU)						
GAU (9570 markers)	16430	16182	11.93	145679	0.530	0.600
Pre-selection of SNPs (method 2)						
GAU (500 markers)	16303	16154	11.88	142179	0.569	0.596
GAU (1000 markers)	**16271**	16102	**11.86**	**139250**	**0.583**	0.614
GAU (2000 markers)	16292	16100	11.87	139628	0.580	0.614
GAU (3000 markers)	16310	16111	11.88	140380	0.577	0.608
Exponential spatial model (EXP)						
EXP (9570 markers)	16440	**15715**	12.66	146827	0.530	0.607
Pre-selection of SNPs (method 2)						
EXP (500 markers)	16302	15986	12.06	141816	0.572	0.599
EXP (1000 markers)	**16275**	15888	12.13	**139427**	**0.583**	0.620
EXP (2000 markers)	16297	15829	12.23	140005	0.582	**0.621**
EXP (3000 markers)	16317	**15807**	12.29	140943	0.580	0.615
Linear spatial model (LIN)						
LIN (9570 markers)			Did not converge			
Pre-selection of SNPs (method 2)						
LIN (500 markers)	16300	15987	12.06	141813	0.572	0.596
LIN (1000 markers)	**16273**	15888	12.13	**139425**	**0.584**	0.614
LIN (2000 markers)	16295	15833	12.22	139995	0.582	0.614
LIN (3000 markers)	16315	**15809**	12.29	140936	0.580	0.608

[$] Pre-selection method 2 is explianed in Chapter 3.

References

Szydlowski, M., and P. Paczyńska. 2011. QTLMAS 2010: simulated dataset. BMC Proc. 5(Suppl 3):S3.

Acknowledgments

First of all, I am very grateful to Prof. Dr. H.-P. Piepho for his advice, suggestions, the many very helpful hints in all statistical issues and the continuous and excellent support during this thesis.

Thanks to Prof. Dr. A. E. Melchinger and Prof. Dr. J. Bennewitz for serving on my graduate committee.

Sincere thanks I would like to spend to Dr. J. O. Ogutu for many discussions, his persistent support and his patience in proofreading.

Many thanks to Dr. A. Gordillo, Dr. Z. Karaman, and Dr. C. Knaak for the excellent collaboration, the fruitful discussions and for being co-authors of the publications.

Special thanks go to the colleagues at the Institute of Crop Science, namely Kathrin Kleinknecht, Karin Hartung, Jens Möhring, André Schützenmeister, Andrea Richter, Ulrike Semmler-Busch, and Boubacar Estaghvirou.

I would also like to thank Anita Rapp, Zeynep Akyildiz, and Gerdi Frankenberger for their great help in organisational matters.

Many thanks go to AgReliant Genetics for providing the data, for the fruitful collaboration, and for the financial support.

Most sincere thanks to my family and Liane Lühmann, who encouraged and supported me throughout the years. Thank you very much.

Curriculum vitae

Torben Schulz-Streeck

Personal Details

Date of birth	March 24, 1981
Place of birth	Kiel, Germany
Nationality	German

Professional experience

02/2009-03/2012	Research Assistant at the Institute of Crop Science, Working Group Bioinformatics at the University of Hohenheim
12/2008	Research Assistant at the Department of Crop Sciences, Division of Plant Breeding at the Georg-August University of Göttingen
08/2007	Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

Education

10/2006-12/2008	Master's program for Agricultural Sciences with emphasis in Plant Sciences at the Georg-August University of Göttingen, Degree: Master of Science
10/2003-09/2006	Bachelor's program for Agricultural Sciences with emphasis in Plant Sciences at the Georg-August University of Göttingen, Degree: Bachelor of Science
04/2003-09/2003	Bachelor's program for Agricultural Sciences at the Christian-Albrechts-Universität zu Kiel
1992-2001	High school (Gymnasium Kronshagen)
1988-1992	Elementary school (Grundschule Gettorf)

Torben Schulz-Streeck
Stuttgart, den 08.03.2012

List of publications

Schulz-Streeck, T., and H.P. Piepho. 2010. Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. BMC Proceedings 4(Suppl 1):S8.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2011. Pre-selection of markers for genomic selection. BMC Proceedings 5(Suppl 3):S12.

Ogutu, J.O., H.P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection with SNP markers. BMC Proceedings 5(Suppl 3):S11.

Piepho H.P., T. Schulz-Streeck, and J.O. Ogutu. 2011. A stage-wise approach for analysis of multi-environment trials. Biuletyn Oceny Odmian 33:7-20.

Piepho, H.P., J.O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow. 2012. Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Science 52:1093-1104.

Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings 6 (Suppl 2):S10.

Piepho, H.P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. Analysis of multi-environment trials by stage-wise approaches. Biometrical Journal 54(6):844-60.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. Theor Appl Genet 126(1):69-82.

Schulz-Streeck, T., J.O. Ogutu, A. Gordillo, Z. Karaman, C. Knaak, and H.P. Piepho. 2013. Genomic selection allowing for marker by environment interaction. Submitted to Plant Breeding.

Schulz-Streeck, T., J.O. Ogutu, Z. Karaman, C. Knaak, and H.P. Piepho. 2012. Genomic selection using multiple populations. Crop Sci. 52:2453–2461.

Software

Schulz-Streeck, T., B. Estaghvirou, and F. Technow. 2012. rrBlupMethod6: Re-parametrization of RR-BLUP to allow for a fixed residual variance. R package, version 1.2. Available at http://cran.r-project.org/web/packages/rrBlupMethod6/index.html

List of conference contributions

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2009. Modelling multiple crosses for genomwide selection in plant breeding using spatial models. Joint workshop of the BMBF-funded projects Synbreed and FUGATO-plus GenoTrack "Statistical Analysis of genomic data" held in Göttingen from the 10th to the 11th of December 2009.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2010. Pre-selection of markers for genomic selection. QTLMAS 2010 workshop held at Poznan in Poland from the 16th to the 18th of May 2010.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2010. Pre-selection of markers for genomic selection. Synbreed project meeting held at University of Hohenheim on 27th of May 2010.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2010. Genomic selection by ridge regression and geostatisitcal models with pre-selection of markers. Genomic based breeding: Conference of the genome analysis section of the German Society of Plant breeding held in Gießen from the 26th to the 28th of October 2010.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2011. Comparisons of single-stage and two-stage approaches to genomic selection. Synbreed project meeting, held in Cuxhaven from the 13th to the 14th of December 2011.

Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit mit dem Titel „Evaluation of alternative statistical methods for genomic selection for quantitative traits in hybrid maize“ von mir selbst verfasst und lediglich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt wurde. Wörtlich oder inhaltlich übernommene Stellen wurden als solche gekennzeichnet. Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Institution oder Prüfungsbehörde vorgelegt. Insbesondere erkläre ich, dass ich nicht früher oder gleichzeitig einen Antrag auf Eröffnung eines Promotionsverfahrens unter Vorlage der hier eingereichten Dissertation gestellt habe und ich nicht die Hilfe einer kommerziellen Promotionsvermittelung oder −beratung in Anspruch genommen habe.

Stuttgart, den 08.03.2012

____

Torben Schulz-Streeck