Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. Albrecht E. Melchinger

Extensions of Genomic Prediction Methods and Approaches for Plant Breeding

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften
vorgelegt
der Fakultät Agrarwissenschaften

von
Frank Technow
aus Bad Langensalza

Stuttgart-Hohenheim
2013

Die vorliegende Arbeit wurde am 13.08.2013 von der Fakultät Agrarwissenschaften als “Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)” angenommen.

Tag der mündlichen Prüfung:	13.12.2013
1. Prodekan:	Prof. Dr. Markus Rodehutscord
Berichterstatter, 1. Prüfer:	Prof. Dr. Albrecht E. Melchinger
Mitberichterstatter, 2. Prüfer:	Prof. Dr. Jörn Bennewitz
3. Prüfer:	Prof. Dr. Hans-Peter Piepho

Contents

[†] Technow, F., C. Riedelsheimer, T.A. Schrag and A.E. Melchinger. 2012. Theor Appl Genet. 125:1181–1194

[§] Technow, F., A. Bürger, and A.E. Melchinger. 2013. G3. 3:197–203 (Suppl. data files av. online)

[‖] Technow, F., and A.E. Melchinger. 2013. Theor Appl Genet. 126:1133–1143 (Suppl. data files and computer programs av. online)

1. General Introduction

Molecular markers provide direct access to the genotypes of individuals. Since they became more widely available in the late 1980s and early 1990s, plant breeding research has focused on utilizing them for shortening breeding cycles and increasing selection intensity (Lande and Thompson, 1990). Two approaches for utilizing molecular marker data for selection purposes might roughly be distinguished. The first, and traditional approach rests on identification of quantitative trait loci (QTL, i.e., loci affecting quantitative traits) as a first step. This is followed by an evaluation of the genetic merit of candidates based on the identified QTL, as a distinct second step (Lande and Thompson, 1990). This approach will henceforth be referred to as 'marker assisted selection' (MAS). The second, and more recent approach is 'genomic selection'. Here, selection is practiced on the basis of genetic values, predicted from the whole molecular marker profile, without a preceding QTL identification step (Meuwissen et al., 2001).

The shortcomings of MAS

QTL identification is the necessary first step of MAS. QTL might be identified in several ways. For example by linkage mapping in artificial mapping populations. These are most commonly biparental families from parents strongly differing in the trait (e.g., resistant and non-resistant to a disease). Later on, with the availability of denser marker maps, it became possible to exploit historical linkage disequilibrium in established breeding populations. This approach, called association mapping, resolved some shortcomings of linkage mapping. For example, association mapping does not require artificial mapping populations with little resemblance to actual breeding populations. For a review of these and other approaches see Mackay (2001). While all approaches of QTL identification have their advantages and disadvantages, they all have one problem in common: only QTL with large effects can be detected. QTL with small effect fail to pass the stringent significance thresholds in place.

The observed genetic variation of many traits can be described well with Fisher's (1918) infinitesimal model, despite its simplicity (Hill, 2010). This means that most relevant traits, and yield in particular have a polygenic genetic architecture with many QTL of very small effect. MAS can only be superior to phenotypic selection when the utilized QTL explain a considerable portion of the genetic variance (Lande and Thompson, 1990). Theoretical results show that huge sample sizes ($\gg 1,000$) would be required to capture considerable portions of genetic variance of polygenic traits (Lande and Thompson, 1990). Even if the immense resources required for reaching such sample sizes were available, success is still doubtful. For example, in a recent study almost 1,500 maize inbred lines were used for identifying QTL

for flowering time and northern corn leaf blight resistance (Van Inghelandt et al., 2012). Despite the uncommonly large sample size, the few identified QTL explained only marginal amounts of genetic variation. Even in human genetics, where sample sizes can range into the tens of thousands, only negligible amounts of genetic variance could be accounted for by QTL mapped for traits such as body height (Yang et al., 2010). This supposedly paradox phenomena was coined 'missing heritability' by the human genetics community (Maher, 2008). Aggravating the identification problem is that the estimated effects of the QTL that do get detected are often biased and inconsistent across and even within populations (Bernardo, 2008; Utz et al., 2000). Because of these and other shortcomings, MAS in plant breeding was found largely unsuitable for improving polygenic traits, and especially yield (Jannink et al., 2010; Bernardo, 2008; Moreau et al., 2004).

Genomic prediction and selection**

Genomic selection presents a solution to the shortcomings of MAS for polygenic traits. The principle ideas of genomic prediction and selection were laid out in the landmark paper by Meuwissen et al. (2001). The revolutionary novelty of their approach is that there is no QTL identification step involved; predictions are directly obtained from the complete marker profile. Compared to MAS, genomic selection shifts the focus from QTL identification to prediction of genetic values. As a consequence, the effects of minor QTL can also be utilized for prediction.

** The term *genomic prediction* as it is used here encompasses all procedures involved in obtaining the predicted genetic values, especially the applied statistical procedures. The term *genomic selection* will refer to selection procedures based upon those predicted genetic values.

Genomic selection proceeds as follows:

1. a genotyped and phenotyped *training population* is generated,

2. the selection candidates are genotyped (*candidate population*),

3. the training population is used to build models for predicting genetic values or breeding values of the candidates,

4. the candidates are selected according to the predicted values.

Meuwissen et al. (2001) devised genomic prediction for applications in animal breeding, specifically for predicting breeding values of dairy bulls. Dairy cattle breeding is indeed the field where genomic prediction and selection were adopted by practitioners first and where it has the most impact hitherto (Pryce and Daetwyler, 2012; Schefers and Weigel, 2012; VanRaden et al., 2009; Hayes et al., 2009).

The advantage of genomic selection over MAS in plant breeding was demonstrated in several simulation studies (Yabe et al., 2013; Heffner et al., 2010; Wong and Bernardo, 2008; Piyasatian et al., 2007; Bernardo and Yu, 2007). This was confirmed recently by an experimental study, conducted over three cycles of recurrent selection for yield and stover traits in maize (Massman et al., 2013). Genomic prediction methodology was also shown to solve the 'missing heritability' paradox in human genetics (Yang et al., 2010).

Genomic prediction methodology

Genomic prediction methods can be categorized in methods that associate genetic effects with markers ('marker effects methods') and methods that associate genetic effects directly with individuals ('polygenic' or 'total genetic effects' methods) (Kärkkäinen and Sillanpää, 2012). The marker effects methods are typically Bayesian or there exist Bayesian versions of non-Bayesian methods (Kärkkäinen and Sillanpää, 2012). I will focus on Bayesian marker effects methods here. Examples of which are BayesB and BayesA and several Lasso-type methods. They mostly differ in the prior distribution associated with marker effect variance components and thereby in their shrinkage behavior (Gianola, 2013; Kärkkäinen and Sillanpää, 2012). BayesB, developed by Meuwissen et al. (2001), seems to be the most well known and widely used one. The BayesB model is

$$y_i \sim \mathcal{N}(\mu_i, \sigma_e^2)$$
$$\mu_i = \beta_0 + \boldsymbol{X}_i \boldsymbol{u},$$

where y_i denotes the phenotypic observation of the i^{th} individual, μ_i is its linear predictor, β_0 denotes the intercept and σ_e^2 the residual variance component. $\mathcal{N}$ denotes the Gaussian density function and indicates that a Gaussian likelihood is used. The row vector $\boldsymbol{X}_i$ codes the marker genotype of the i^{th} individual (e.g., as 0,1 and 2) and vector $\boldsymbol{u}$ contains the additive marker effects. To β_0 and σ_e^2 uninformative prior distributions are typically assigned. The prior for the marker effects is $p(\boldsymbol{u}_j|\sigma_{u_j}^2) = \mathcal{N}(0, \sigma_{u_j}^2)$. The prior associated with the prior variance of the effect of the j^{th} marker ($\sigma_{u_j}^2$) is specific to BayesB

and equal to

$$p(\sigma^2_{u_j}|\nu, S^2) \begin{cases} = 0 & \text{with probability } \pi \\ = \chi^{-2}(\nu, S^2) & \text{with probability } (1-\pi). \end{cases}$$

The hyperparameters π, ν and S^2 were set to fixed values in the original implementation of Meuwissen et al. (2001). We, however, follow the developments of Yang and Tempelman (2012) and specify prior distributions to these hyperparameters, too. Details can be found in Yang and Tempelman (2012).

Genomic best linear unbiased prediction (GBLUP), i.e., BLUP based on genomic realized relationships, is the most typical and widely used representative of the total genetic effect methods. It was first described by Villanueva et al. (2005) and later shown to be mathematically equivalent to ridge regression BLUP (Strandén and Garrick, 2009; Piepho, 2009), developed by Meuwissen et al. (2001). For sake of a unified treatment, I will present the Bayesian version of GBLUP (Kärkkäinen and Sillanpää, 2012). In my experience this delivers virtually identical prediction results as the more popular frequentist version. The model is

$$\begin{aligned} \mu_i &= \boldsymbol{\beta}_0 + a_i \\ y_i &\sim \mathcal{N}(\mu_i, \sigma^2_e), \end{aligned}$$

where a_i denotes the total genetic value of individual i, and all other terms are as before. The prior for a_i is $\mathcal{MVN}(\mathbf{0}, \boldsymbol{A}\sigma^2_a)$, where $\mathcal{MVN}$ denotes the Multivariate-Gaussian density function and σ^2_a the total genetic variance component. Matrix $\boldsymbol{A}$ is typically an estimate of the realized additive relationship matrix, obtained from marker data (VanRaden, 2008). There are

several different approaches for expressing the genetic similarities between individuals embodied in $\boldsymbol{A}$, apart from additive relationships. Piepho (2009) for example, proposes geostatistical methods for this purpose. The intercept and the variance components are associated with uninformative priors. The frequentist version of GBLUP follows by setting the variance components to fixed values (estimated from the data by REML, for example) and solving the corresponding mixed model equations. We note again that there are several equivalent formulations of (frequentist) GBLUP, including formulations involving marker effects, all having different computational properties (Piepho et al., 2012; Piepho, 2009; Strandén and Garrick, 2009).

Genomic prediction in maize hybrid breeding

Single-cross hybrids are the predominant cultivar type in maize. They are generated by crossing two homozygous inbred lines. For maximizing heterosis, these lines are taken from genetically distant groups of germplasm, called 'heterotic groups' (Melchinger and Gumber, 1998). In Central Europe these are the Dent and Flint heterotic groups. Because the parental lines are fully homozygous, maize single-cross hybrids are fixed genotypes that can be multiplied *ad libitum* and released as cultivar.

Facilitated by advances in doubled-haploid technology (Wedzony et al., 2009) the arrays of available parental lines increased tremendously. With this, the number of potential hybrids grew enormously. For example, with only 1,000 lines generated in each heterotic group per year, the number of potential hybrids

reaches a staggering number of one million. Only a small fraction of these can be tested in field trials. Prediction of hybrid performance is therefore of tremendous importance for hybrid breeding.

Building onto the success of genomic prediction in other fields, genomic prediction of hybrid performance might be a valuable tool for identifying superior hybrids. However, the genomic prediction models and methods used for predicting additive breeding values in animal breeding or recurrent selection programs in plant breeding must be extended to account for unique characteristics of maize hybrids.

The specific combining ability (SCA) of the parental inbred lines is a major factor determining hybrid performance (Sprague and Tatum, 1942). In the absence of epistasis, the genetic variance pertaining to SCA effects is the sum of the variances due to dominance effects of QTL (Reif et al., 2007). Thus, a maximum amount of the total genetic variance can only be captured when incorporating dominance marker effects into genomic prediction models.

Another complication comes from the fact that the parents of a hybrid are taken from different heterotic groups. It was shown that the Dent and Flint groups have been separated for at least 500 years (Rebourg et al., 2003). Because of the many generations of differentiation between them, the linkage phases between marker and QTL can be different in Dent and Flint (Charcosset and Essioux, 1994). Further, random drift and mutations could have led to the presence of different QTL alleles in both groups. Thus, models that estimate a single effect for each marker, regardless from which heterotic group the marker allele was derived, did not seem adequate.

The presence of separated heterotic groups also requires the execution of parallel breeding programs for inbred line development. The limited resources available for phenotyping of training populations, for the goal of genomic selection of inbred line parents for either testcross or *per se* performance, then have to be allocated to all heterotic groups. This hampers the construction of sufficiently sized training populations within each heterotic group. Augmenting the training population of one group with individuals from the other would be a cost neutral way of increasing the training population size. However, whether this would also increase prediction accuracy was doubtful, because of the 500 years separation of Dent and Flint.

Non-gaussian data in plant breeding

In plant breeding, typically multiple phenotypic records are available per individual (e.g., from multiple locations). A special case of this are repeated phenotypic records for a dichotomous trait, i.e., a Binomial phenotypic distribution. Examples of dichotomous traits are disease resistance (disease outbreak or not), germination (seed germinates or not) and haploid induction and spontaneous chromosome doubling in maize (seedling haploid or diploid). The latter two traits are of immense importance for economic production of doubled haploid lines in maize (Prigge et al., 2012; Kleiber et al., 2012). Genomic prediction methodology was originally developed for Gaussian traits, such as yield. Later, extensions for Bernoulli distributed phenotypes (i.e., a single observation of a dichotomous trait) were proposed (Lee et al., 2011). However, generalized linear model extensions of BayesB and GBLUP for binomially distributed phenotypic data were unavailable.

Objectives

The main goal of this thesis was to adapt and extend genomic prediction methods and approaches to cover unique aspects of maize hybrid breeding in particular and plant breeding in general. Specifically, the goals were to

1. provide extension of prediction models to dominance and heterotic group specific marker effects,
2. investigate the merit of augmenting training populations with individuals from different heterotic groups,
3. provide generalized linear model extensions of BayesB and GBLUP for genomic prediction of traits with Binomial phenotypic distribution, and
4. identify the circumstances in which our extensions have the greatest impact on prediction accuracies.

Literature cited

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci 48*, 1649–1664.

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

Charcosset, A. and L. Essioux (1994). The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theor Appl Genet 89*, 336–343.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc Edin 52*, 399–433.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci 92*, 433–443.

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

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Phil Trans R Soc 365*, 73–85.

Jannink, J.-L., A. J. Lorenz, and H. Iwata (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics & proteomics 9*, 166–177.

Kärkkäinen, H. P. and M. J. Sillanpää (2012). Back to basics for bayesian model building in genomic selection. *Genetics 191*, 969–987.

Kleiber, D., V. Prigge, A. E. Melchinger, F. Burkard, F. San Vicente, G. Palomino, and G. A. Gordillo (2012). Haploid fertility in temperate and tropical maize germplasm. *Crop Sci 52*, 623–630.

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

Lee, S. H., N. R. Wray, M. E. Goddard, and P. M. Visscher (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet 88*, 294–305.

Mackay, T. (2001). The genetic architecture of quantitative traits. *Annu rev genet 35*, 303–339.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature 456*, 18–21.

Massman, J. M., H.-J. G. Jung, and R. Bernardo (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci 53*, 58–66.

Melchinger, A. E. and R. K. Gumber (1998). *Concepts and Breeding of Heterosis in Crop Plants edited by K.R. Lamkey and J.E. Staub*, Chapter Overview of heterosis and heterotic groups in agronomic crops, pp. 29–44. CSSA, Madison, WI.

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

Moreau, L., A. Charcosset, and A. Gallais (2004). Experimental evaluation of several cycles of marker-assisted selection in maize. *Euphytica 137*, 111–118.

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

Piepho, H. P., J. O. Ogutu, B. Estaghvirou, A. Gordillo, and F. Technow (2012). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Sci 52*, 1093–1104.

Piyasatian, N., R. L. Fernando, and J. C. M. Dekkers (2007). Genomic selection for marker-assisted improvement in line crosses. *Theor Appl Genet 115*, 665–674.

Prigge, V., X. Xu, L. Li, R. Babu, S. Chen, G. N. Atlin, and A. E. Melchinger (2012). New insights into the genetics of in vivo induction of maternal haploids, the backbone of doubled haploid technology in maize. *Genetics 190*, 781–93.

Pryce, J. and H. Daetwyler (2012). Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim Prod Sci 52*, 107–114.

Rebourg, C., M. Chastanet, B. Gouesnard, C. Welcker, P. Dubreuil, and A. Charcosset (2003). Maize introduction into Europe: the history reviewed in the light of molecular data. *Theor Appl Genet 106*, 895–903.

Reif, J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger (2007). Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics 176*, 1931–1934.

Schefers, J. M. and K. a. Weigel (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. *Animal Frontiers 2*, 4–9.

Sprague, G. F. and L. A. Tatum (1942). General vs. specific combining ability in single crosses of corn. *J. Amer. Soc. Agron. 34*, 923–932.

Strandén, I. and D. J. Garrick (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci 92*, 2971–2975.

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

Van Inghelandt, D., A. E. Melchinger, J.-P. Martinant, and B. Stich (2012). Genome-wide association mapping of flowering time and northern corn leaf blight (Setosphaeria turcica) resistance in a vast commercial maize germplasm set. *BMC plant biology 12*, 56.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J Dairy Sci 91*, 4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci 92*, 16–24.

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

Wedzony, M., B. Forster, I. Zur, E. Golemiec, M. Szechyńska-Hebda, E. Dubas, G. Gotebiowska, and M. Wedzony (2009). Progress in doubled haploid technology in higher plants. In A. Touraev, B. Forster, and S. Jain (Eds.), *Advances in Haploid Production in Higher Plants*, pp. 1–33. Springer Netherlands.

Wong, C. K. and R. Bernardo (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet 116*, 815–824.

Yabe, S., R. Ohsawa, and H. Iwata (2013). Potential of genomic selection for mass selection breeding in annual allogamous crops. *Crop Sci 53*, 95–105.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat genet 42*, 565–569.

Yang, W. and R. J. Tempelman (2012). A Bayesian antedependence model for whole genome prediction. *Genetics 190*, 1491–1501.

Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects

F. Technow, C. Riedelsheimer, T.A. Schrag, A.E. Melchinger

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

Theor Appl Genet. 125:1181–1194 (2012)
The original publication is available at link.springer.com

Abstract Identifying high performing hybrids is an essential part of every maize breeding program. Genomic prediction of maize hybrid performance allows to identify promising hybrids, when they themselves or other hybrids produced from their parents were not tested in field trials.

Using simulations, we investigated the effects of marker density (10, 1, 0.3 marker per mega base pair, Mbp^{-1}), convergent or divergent parental populations, number of parents tested in other combinations (2, 1, 0), genetic model (including population specific and/or dominance marker effects or not) and estimation method (GBLUP or BayesB) on the prediction accuracy. We based our simulations on marker genotypes of Central European flint and dent inbred lines, from an ongoing maize breeding program. To simulate convergent or divergent parent populations, we generated phenotypes by assigning QTL to markers with similar or very different allele frequencies in both pools, respectively. Prediction accuracies increased with marker density and number of parents tested and were higher under divergent compared to convergent parental populations. Modeling marker effects

as population specific slightly improved prediction accuracy under lower marker densities (1 and 0.3 Mbp^{-1}). This indicated that modeling marker effects as population specific will be most beneficial under low linkage disequilibrium. Incorporating dominance effects improved prediction accuracies considerably for convergent parent populations, where dominance results in major contributions of SCA effects to the genetic variance among inter-population hybrids. While the general trends regarding the effects of the above mentioned influence factors on prediction accuracy were similar for GBLUP and BayesB, the latter method produced significantly higher accuracies for models incorporating dominance.

Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups

F. Technow, A. Bürger, A.E. Melchinger

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

G3. 3:197–203 (2013)

The original publication is available at g3journal.org

Abstract Northern corn leaf blight (NCLB), a severe fungal disease causing yield losses worldwide, is most effectively controlled by resistant varieties. Genomic prediction could greatly aid resistance breeding efforts. But the development of accurate prediction models requires large training sets of genotyped and phenotyped individuals. Maize hybrid breeding is based on distinct heterotic groups that maximize heterosis (the dent and flint groups in Central Europe). The resulting allocation of resources to parallel breeding programs challenges the establishment of sufficiently sized training sets within groups. Therefore, using training sets combining both heterotic groups might be a possibility of increasing training set sizes and thereby prediction accuracies. The objectives of our study were to assess the prospect of genomic prediction of NCLB resistance in maize and the benefit of a training set which combines two heterotic groups. Our data comprised 100 dent and 97 flint lines, phenotyped for NCLB resistance *per se* and genotyped with high density single nucleotide polymorphism marker data. A genomic BLUP model was used to predict genotypic values. Prediction accuracies reached a maximum of 0.706 (dent) and 0.690 (flint) and

there was a strong positive response to increases in training set size. Using combined training sets led to significantly higher prediction accuracies for both heterotic groups. Our results encourage the application of genomic prediction in NCLB resistance breeding programs and the use of combined training sets.

Genomic prediction of dichotomous traits with Bayesian logistic models

F. Technow, A.E. Melchinger

Institute of Plant Breeding, Seed Science, and Population Genetics,
University of Hohenheim, 70593 Stuttgart, Germany

Theor Appl Genet. 126:1133–1143 (2013)
The original publication is available at link.springer.com

Abstract Bayesian methods are a popular choice for genomic prediction of genotypic values. The methodology is well established for traits with approximately Gaussian phenotypic distribution. However, numerous important traits are of dichotomous nature and the phenotypic counts observed follow a Binomial distribution. The standard Gaussian generalized linear models (GLM) are not statistically valid for this type of data. Therefore, we implemented Binomial GLM with logit link function for the BayesB and Bayesian GBLUP genomic prediction methods. We compared these models to their standard Gaussian counterparts using two experimental data sets from plant breeding, one on female fertility in wheat and one on haploid induction in maize, as well as a simulated data set. With the aid of the simulated data referring to a bi-parental population of doubled haploid lines, we further investigated the influence of training set size (N), number of independent Bernoulli trials for trait evaluation (n_i) and genetic architecture of the trait on genomic prediction accuracies and abilities in general and on the relative performance of our models. For BayesB, we in addition implemented finite mixture Binomial GLM to account for overdispersion. We found that prediction accuracies increased with increasing N and n_i. For

the simulated and experimental data sets, we found Binomial generalized linear models to be superior to Gaussian models for small n_i, but that for large n_i Gaussian models might be used as *ad hoc* approximations. We further show with simulated and real data sets that accounting for overdispersion in Binomial data can markedly increase the prediction accuracy.

5. General Discussion

Accounting for presence of heterotic groups

The popular genomic prediction methods GBLUP and BayesB were extended to heterotic group specific and dominance marker effects (Technow et al., 2012). A proof of concept was provided with a simulation study, based on the marker profiles of actual Dent and Flint inbred lines from the hybrid maize breeding program of the University of Hohenheim. We showed that the extensions can lead to higher prediction accuracies than simple additive models. However, the differences in prediction accuracy of the extended models and the basic additive model were usually rather moderate and depended on the particular scenario simulated. An investigation of the consistency of marker linkage phases between the Dent and Flint lines in this study revealed a remarkably high consistency in the linkage phase of markers in close proximity, i.e., with a physical distance of less than 0.5 Mbp (Technow et al., 2012, 2013). This was surprising, given the more than 500 years of separation between the Dent and Flint germplasm (Rebourg et al., 2003). The unexpectedly

high linkage phase consistency explained why the model that included population specific marker effects did not increase prediction accuracy over the conventional unspecific marker effects model under high marker densities. Only when the marker density was decreased to about one marker per Mbp, (i.e., to a point where across heterotic group linkage phase consistency was low), the increase in prediction accuracy was substantial. However, achieving the required marker densities should not be a problem anymore with modern genotyping techniques such as the 50k SNP chips (Ganal et al., 2011) or genotyping by sequencing (Elshire et al., 2011). The true effects of the QTL were simulated to be the same in both heterotic groups. The presence of heterotic group specific QTL alleles would increase the necessity of estimating heterotic group specific marker effects, irrespective of the linkage phase consistency between markers and QTL.

For genomic prediction of inbred line *per se* performance for resistance against northern corn leaf blight, we showed that augmenting the training population of one heterotic group with individuals from the other did increase the prediction accuracy considerably (Technow et al., 2013). Thereby, we compared the prediction accuracy increase achieved by adding a certain number x of individuals from the opposite heterotic group to the training population to the increase achieved by adding x individuals from the same heterotic group. We found that the prediction accuracy increase in the latter case was considerably larger than the increase in the former case. Further, attempting to predict individuals from one heterotic group with a training population solely consisting of individuals from the other heterotic group resulted in very low prediction accuracies. Thus, the information contributed by individuals from the opposite

heterotic group is considerably lower than that contributed by individuals from the same one as the predicted candidates. This was expected because of the centuries of separation of the Dent and Flint heterotic groups. However, the fact that augmentation of training populations with individuals from the opposite heterotic group increased prediction accuracy at all is remarkable. It seems possible only, because Dent and Flint share a common genetic basis in terms of QTL alleles and linkage phases between markers and QTL. Otherwise, Dent lines would not convey any useful information for prediction of Flint lines and vice versa. This hypothesized shared genetic basis, together with our simulation results, indicated that estimating heterotic group specific marker effects might not be essential for accurate prediction of hybrid performance. As will be discussed later, the increased dimensionality of heterotic group specific models might even have negative effects on prediction accuracy.

Merit of genomic hybrid prediction

The performance of a single-cross hybrid is the sum of the general combining ability (GCA) of its parents and the specific combining ability effect (SCA) of the parental combination (Sprague and Tatum, 1942). It is therefore possible to predict the performance of a hybrid from the GCA effects of the parents only, with the SCA effects becoming a source of prediction error.
The GCA effects of the parental lines are typically obtained from field evaluation of testcross progeny with testers from the opposite heterotic group. Generating testcross progeny, for example in a top-cross nursery, is much less resource intensive than producing specific single crosses by hand pollination. Genomic prediction could obviously be used to predict the parental GCA

effects, too (Albrecht et al., 2011). Prediction of hybrid performance based only on parental GCA effects, henceforth termed 'GCA based prediction', was in fact the traditional approach for identifying superior hybrids practiced over the last decades. It is still common practice in less progressive breeding programs. In such a scheme, field testing of a few promising experimental hybrids, produced according to a complete or partial factorial mating desing, takes place only as a very last step. Only then could the SCA variance be exploited for selection. The degree in which GCA based prediction of hybrid performance would decrease prediction accuracy depends on (i) the importance of SCA variance relative to GCA variance (Reif et al., 2007) and (ii) whether SCA effects can be predicted accurately.

A comparison between the prediction accuracy of purely additive models and models incorporating dominance effects is not only relevant for model choice but provides also useful hints on answering the previous points. If models with dominance effects fail to increase prediction accuracy, one might conclude that the increased phenotyping efforts required for genomic hybrid prediction are not worthwhile because GCA based prediction already achieves the maximum degree of prediction accuracy.

In the simulation study, inclusion of dominance marker effects increased prediction accuracy considerably when the SCA variance was substantial, but the increase was only moderate when the contribution of SCA to the total genetic variance among hybrids was low (Technow et al., 2012). This demonstrated that dominance effects can be estimated and, consequently SCA effects

predicted to a certain degree by genomic prediction approaches. After the publication of Technow et al. (2012), several other studies on genomic hybrid prediction appeared in maize (Guo et al., 2013; Massman et al., 2013), sunflower (Reif et al., 2013) and wheat (Zhao et al., 2013). All of these studies reported high accuracies for genomic prediction of hybrid performance. However, Reif et al. (2013) and Zhao et al. (2013) found that genomic prediction with models including dominance effects were not superior to prediction based on purely additive models. The authors explained their observation with the low contribution of SCA variance to the total genetic variance in their experiments. This is in line with the results of Guo et al. (2013), who also found that superiority of models incorporating dominance effects over purely additive models depends on the ratio of the SCA variance to the total genetic variance. Thus, a consensus seems to emerge that for genomic hybrid prediction to be advantageous over GCA based prediction, SCA variance must be relatively important. It is still an open question how much higher the accuracy of genomic hybrid prediction compared to GCA based prediction must be in order to economically justify the increased resource requirements. It should be noted, however, that single-cross hybrids have to be phenotyped in any breeding program at some point. Therefore, a certain number of phenotyped single-cross hybrids is generated anyway and could form the basis of a training population. For example, in the maize hybrid breeding program of the University of Hohenheim, more than 1,000 single cross hybrids were phenotyped on a routine basis over the last decade. This would already constitute a sufficiently sized training population, given that all the studies cited above achieved surprisingly high prediction accuracies with much smaller training populations.

Inclusion of dominance effects into genomic prediction models was successfully attempted in animal breeding, too (Wellmann and Bennewitz, 2012; Wittenburg et al., 2011; Toro and Varona, 2010). It was shown that models with dominance can increase both the accuracy of genomic breeding value prediction as well as prediction of genetic values (Wellmann and Bennewitz, 2012). Thereby prediction of genetic values of individuals profited considerably more than prediction of their breeding values. In plant breeding, individual genotypes, such as single-cross hybrids, can be multiplied *ad libitum* and can be of tremendous economic value if successful as a variety. In contrast, genotypes in animal breeding are confined to a single individual of comparatively low economic value, at least in respect to their own performance. Thus, genomic prediction of genetic values might be of less importance in animal breeding. Mate allocation emerged as a particularly promising application involving estimated dominance effects. Here, male and female parents of a paring are chosen such that the contribution of favorable dominance combinations are maximized (Wellmann and Bennewitz, 2012; Toro and Varona, 2010). This concept closely resembles the concept of specific combining ability in hybrid breeding. However, because of Mendelian sampling, mate allocation is limited to increasing the average performance of the resulting full-sib families, from which individual members can deviate.

Estimation of dominance effects of markers is possible only if all relevant individuals have a recorded phenotypic value and are genotyped (Wellmann and Bennewitz, 2012; Wittenburg et al., 2011; Toro and Varona, 2010). This is associated with increased data recording efforts and costs, because also females (e.g., dairy

cows) have to be genotyped. This would not be required for prediction of additive breeding values of bulls. Thus, as was the case for hybrid prediction in maize, the potential increase in prediction accuracy needs to be weighed against the increased resource requirements.

Choice of the statistical model

Implementing my extensions to population specific and dominance marker effects and (overdispersed) binomially distributed data, involved a considerable programming effort. The computational efforts required for fitting of the extended models were also greater than for simple models, because of their higher dimensionality and because of non-standard Gibb-sampling procedures.

Despite the increased effort, using the more sophisticated models resulted in mostly moderate gains in prediction accuracy. This raises the question of whether the increased efforts are worthwhile. An important argument in favor is that any gains come virtually at no extra costs, at least when compared to costs associated with other means of increasing prediction accuracy, such as increasing the sample size or heritability. Implementing the software does neither require very advanced computer skills, nor are expensive proprietary software systems necessary. All computations involved in this thesis were carried out with programs implemented in the freely available programming languages and environments C, R (R Core Team, 2012) and JAGS (Plummer, 2003). Once a working program is available, it can be used with

little or no adjustments for years. The maintenance efforts are thus very minimal.

As long as computations are feasible at all, the increased computing time is an issue only for elaborate simulation or cross-validation studies. In such studies the algorithms have to be run hundreds or even thousands of times. In practical applications the goal is usually not to test an algorithm or hypothesis but rather just to predict genetic values for selection purposes. Here, computations have to be run only once per trait and season. Furthermore, technical advances continue to make the remaining computational burden more manageable. The computing power of even standard desktop systems is increasing by the year. With cloud computing services like 'Amazon EC2'[††], even small breeding companies and research institutions gain access to high performance computing facilities.

Thus, I strongly advocate to employ the model that delivers highest prediction accuracies, even when the differences to simpler models are small. The advantage of more sophisticated models will increase the more the modeled features (e.g., dominance, overdispersion ...) influence the distribution of the data. The more negligible these feature are, the more appropriate a simpler model will be. It is impossible to decide beforehand whether this point is reached for a specific data set. Therefore, using simple models by default runs the risk of missing important features of the data and thereby a serious loss in prediction accuracy.

[††] http://aws.amazon.com/ec2/

Choice of the prediction method

The choice between prediction methods can be seen as a model choice as well. Hereby marker effects methods like BayesB represent more complex models which can represent a genetic architecture consisting of many QTL with very small and a few QTL with very large effects. GBLUP in turn is based on the simplifying and more restrictive assumptions of Fisher's (1918) infinitesimal model, i.e., a genetic trait architecture consisting of a very large number of QTL, all with very small effects. It is generally agreed that marker effects methods are superior under an oligogenic trait architecture and total genetic effects methods under a polygenic architecture (Kärkkäinen and Sillanpää, 2012; Clark et al., 2011; Hayes et al., 2010; Zhong et al., 2009). This was confirmed in this thesis for dichotomous traits, too (Technow and Melchinger, 2013).

However, as was the case for models of different complexity, we observed that the differences between GBLUP and BayesB were usually small, compared to other factors (Technow and Melchinger, 2013; Technow et al., 2012). Heslot et al. (2012) compared the performance of a wide array of genomic prediction methods. They found for many different traits and crops only small differences between methods. In a recent study, Hu et al. (2013) compared the performance of Bayesian GBLUP and BayesB for several stalk bending traits in maize. They also found only small and inconsistent differences between the two methods.

The theoretical concept of independent chromosome segments (M_e), developed by Goddard (2009), helps to understand why

the differences between marker effects methods like BayesB and total genetic effects methods like GBLUP are not expected to be large in typical plant breeding populations. M_e, computed as

$$M_e = \frac{2N_eL}{\log(4N_eL)}$$

is a function of the genome length in Morgan (L) and the effective population size (N_e). The lower N_e, the lower is M_e. Based on M_e, the expected prediction accuracy can be estimated (Daetwyler et al., 2010). Estimating M_e in this way assumes absence of a complex family structure. This assumption is obviously unrealistic. Wientjes et al. (2012) used a method for computing M_e that takes into account the family structure by using observed genomic and pedigree relationships. They found that both methods for computing M_e delivered similar results. Thus, $M_e = {}^{2N_eL}/_{\log(4N_eL)}$ seems to provide a useful approximation. A good agreement between observed and expected prediction accuracy (based on $M_e = {}^{2N_eL}/_{\log(4N_eL)}$) was found in this thesis (Technow et al., 2013). This further demonstrates the validity of the concept.

The GBLUP model assumes equal importance of all segments and weights them equally in the computation of the realized relationships. BayesB and the other marker effects methods are more flexible in this regard. They can adapt to a situation where some segments are more important than others and some might not influence the trait at all (Daetwyler et al., 2010). Thus, when the number of QTL (N_{QTL}) is lower than M_e and/or the distribution of QTL effects is such that a few QTL explain a large part of the genetic variance, BayesB and similar methods should have an advantage over GBLUP.

When N_{QTL} is large (so that all segments carry QTL) and the QTL have effects of similar magnitude, then the assumption of GBLUP is reasonable. This does not mean that GBLUP is necessarily superior to BayesB in this case. In fact, BayesB is expected to deliver approximately the same prediction accuracy as GBLUP, provided that certain hyperparameters that control the amount of sparseness and shrinkage were associated with prior distributions and, thus, estimated from the data (Yang and Tempelman, 2012).

In plant breeding populations, N_e, and thereby also M_e, is typically low (Guzman and Lamkey, 2000). With $L \approx 16\,\mathrm{M}$, as was observed for maize (Martin et al., 2011), and with $N_e = 25$, a typical value for maize populations under recurrent selection (Guzman and Lamkey, 2000), $M_e \approx 108$. For bi-parental populations, M_e might be as low as 28 (Lorenz, 2013). Consequently, for many traits, including yield, $M_e \ll N_{QTL}$. Thus, BayesB and similar methods are not expected to have any substantial advantage over GBLUP. It might be noted that even with $M_e \ll N_{QTL}$, the importance of individual segments might differ. Unrealistically high population sizes, however, would be required for BayesB type models to capture these small differences. Therefore, choice of the prediction method is not expected to lead to dramatic differences in prediction accuracy for typical population types encountered in plant breeding. However, as discussed above for the model choice, potential gains in prediction accuracy come virtually at no costs and therefore should be exploited.

The curse of dimensionality

From the derivation of expected prediction accuracy (Daetwyler et al., 2010) and also from the argumentation of Yang and Tempelman (2012), BayesB and similar marker effects methods should in fact reach the same accuracy level as GBLUP. In other words, GBLUP represents a simplified model compared to BayesB, which, in some cases ($M_e \ll N_{QTL}$) is sufficient. The same could be said for statistical models of different complexity; a more sophisticated and complex model can be expected to be at least as good as the simpler and restrained model. In this part I will discuss reasons why this is often not the case. To simplify the discussion, the choice between marker effects methods and total genetic effects methods will be subsumed under 'model complexity', too.

Many studies observed that GBLUP actually outperformed BayesB and similar methods significantly. In this thesis, this was observed for the special case of binomially distributed data when the simulated trait architecture was highly polygenic (Technow and Melchinger, 2013). We also observed that a highly complex model able to fit overdispersion led to significantly lower prediction accuracies than a less complex model when the sample size was low (Technow and Melchinger, 2013). Other authors observed that a marker effects model including dominance terms led to lower prediction accuracies than a simple additive model (Zhao et al., 2013). These observations contradict the assumption that complex models should always be at least as good as simple ones.

Overfitting

Complex and flexible models are more prone to overfitting than more simple models (Hawkins, 2004). An overfitted model is one that violates the well established principle of parsimony. This principle is tied to Ockham's razor principle. It states that an explanation or theory should not make any unnecessary assumptions. However, the principle might be best understood with the following proverb: "When you hear hoofbeats, think horses, not zebras. (Unless you are in Africa)".

One of the risks of overfitting is that predictors capture spurious features of the data ('artifacts') and noise. This could lead to erroneous interpretations when the goal is inference about parameters. The coefficients of predictors representing artifacts or noise would add random variation to the predictions of future observations (Hawkins, 2004). Therefore, overfitting leads to a reduction in prediction accuracy, too. Thus, overfitting is one reason why complex models might result in lower prediction accuracies than simple models, especially if the sample size is small.

Non-identifiability

Models where the number of parameters (p) exceeds the sample size (n) are not *likelihood-identified*. This means that an infinite number of possible parameter vectors would fit the data equally well and result in the same likelihood value (Gianola, 2013; Gelfand and Sahu, 1999). Hence, in the $p \gg n$ scenarios, maximum likelihood or least squares methods can not be used for parameter estimation or for prediction purposes (Gianola,

2013). Instead, some kind of regularization mechanism has to be introduced, which is automatically provided by the prior distribution of Bayesian models (Gianola, 2013). Gelfand and Sahu (1999) showed that Bayesian learning (*i.e.*, the data influencing the posterior distribution) is still possible in non-identified models. However, they caution that the prior in this case will always be influential and may determine the posterior distribution of the parameters to a large extend. Gianola (2013) discussed non-identifiability explicitly in the context of genomic prediction. He strongly questions the validity and meaningfulness of attempts to infer upon QTL effects and genetic architecture from the posterior distributions of marker effects, when models are not likelihood-identified. However, this does not question the value of models like BayesB for prediction purposes, since the posterior predictive distribution $p(y_f|\boldsymbol{y})$, where y_f denotes a future observation and $\boldsymbol{y}$ the current data, is unique (Gianola, 2013). The expression $p(y_f|\boldsymbol{y})$ shows that the parameters "do not necessarily play an 'existential role'" (Gianola, 2013) for prediction purposes and Gianola (2013) argues that they should be viewed as "tools enabling one to go from past to future observations".

Gelfand and Sahu (1999) mention another problem of more practical nature: Non-identifiable models can lead to convergence problems of the Gibbs-samplers used for drawing samples from the posterior distribution. The consequence of this is inaccurate and unstable parameter estimation, which necessarily affects prediction accuracy. We observed severe convergence problems in our very complex BayesB models for fitting overdispersed binomially distributed data when $p \gg n$ (Technow and Melchinger,

2013). The models employed were finite mixture models with several effects for each marker, one per mixture component. Such models are especially prone to non-identifiability problems (Frühwirth-Schnatter, 2006). We solved these problems simply by reducing the number of markers. When overdispersion was strong, the model that fitted overdispersion then outperformed the simple model (fitted with the reduced or full set of markers) that did not model overdispersion. Interestingly, it was not necessary to reduce the number of markers below the training population size for the convergence problems to disappear.

Usage of less vague priors can also improve convergence properties (Gelfand and Sahu, 1999). Indeed, we observed that using slightly informative priors for certain hyperparameters of the BayesB model improved convergence (Technow et al., 2012; Technow and Melchinger, 2013). This was also observed by Yang and Tempelman (2012), who developed the specific BayesB parametrization we took as starting point for our extensions. Yang and Tempelman (2012) argue that specifying prior distributions for certain key hyperparameters allows Bayesian inference about these. This would be interesting because these hyperparameters can be given quantitative genetic interpretations, which would in turn facilitate inferences on the genetic architecture of complex traits (e.g., about N_{QTL}). To what extend the estimation of these hyperparameters is affected by non-identifiability, remains to be investigated.

Our extensions of BayesB for hybrid predictions involved estimation of up to three effects per marker (one additive effect for each heterotic group and one dominance effect). This greatly increased the dimensionality of the models and set them up for non-identifiability issues as well. We did not notice any se-

vere convergence problems, though. However, non-identifiability might have prevented these models from accurately capturing small differences between the effects of markers in one vs. the other heterotic group or small dominance effects. This might be one reason for the only moderate gains in prediction accuracy from using these models.

Resume and outlook

The results of this thesis clearly showed that more complex and sophisticated statistical models and prediction methods can lead to cost free gains in prediction accuracy. These gains will be most pronounced when the modeled features have a substantial impact on the properties of the data. In some cases, sophisticated and specialized models are in fact required to obtain decent prediction accuracies at all, e.g., with severely overdispersed binomially distributed data. However, their successful application requires a certain level of expertise from the user. For example, the user needs to be able to detect and solve convergence problems of elaborate Gibbs-sampling algorithms, like BayesB. Their application without proper understanding of the underlying methodology and of Bayesian statistics in general, is not recommended.

It is worthwhile to emphasize again that the potentially considerable gains in prediction accuracy associated with more sophisticated prediction methods and models come at no extra costs and usually do not require changes in the breeding methodology. This makes method and model choice a critical factor of success

for a breeding program in a very competitive market, where even small advantages can be decisive.

Further improvements of methods and models are possible. Future advances will be facilitated by better understanding of the genetic factors that drive prediction accuracy. For example, the relationship between linkage disequilibrium and genomic relationship, and how both affect prediction accuracy, still seems elusive. Important first steps for elucidating the role of these factors have been made recently (Habier et al., 2013).

In my view, however, improvements will not come from further extending the 'Bayesian Alphabet' (Gianola, 2013) or the already large assembly of regularized regression (Ogutu et al., 2012) and machine learning methods (Ogutu et al., 2011) with methods that are all supposed to do the same thing: predicting yield in ordinary populations. For this purpose, the differences between methods can be expected to be negligible indeed, as Gianola (2013) argues for the methods of the 'Bayesian Alphabet'. Instead, I believe that optimization of prediction methods can contribute the most by providing tailored solutions for cases in which standard methods will not work optimally. In this thesis, this approach was successfully demonstrated for phenotypic data from an (overdispersed) Binomial distribution (Technow and Melchinger, 2013).

The stabilizing effect that slightly informed priors have on computations was already mentioned. I argue that informative priors, far from being a liability, are actually an asset of Bayesian statistics. This point was largely neglected by the genomic prediction community hitherto. Valid prior information, for example obtained from past experiments, could mitigate the influence of artifacts and sampling effects in the typically small training

populations of plant breeding programs. Thus, incorporation of prior information could prove to be a novel strategy for improving the accuracy of genomic prediction in plant breeding.

Despite the importance of prediction methods, aspects of breeding methodology are of great importance, too. They must not be overlooked when implementing genomic breeding programs. The optimal allocation of resources for maximizing the response to genomic selection is of major importance in this regard. Only a few examples of parameters that have to be optimized will be given here, namely the training populations size and heritability. Prediction accuracy can be increased by increasing the training population size or the heritability. However, both constrain each other under a fixed budget. One can either produce a large training population phenotyped with low heritability or a small training population phenotyped with high heritability. Maximizing prediction accuracy with respect to these two parameters under the constrains of a fixed budget will be a vital step in planning of genomic selection programs. As long as genotyping costs are not negligible, they have to be factored into the optimization process as well. The larger the candidate population, the higher the selection intensity. However, resources spend on genotyping candidates can not be spend on increasing size and heritability of the training population. Thus, the size of the candidate population constrains the prediction accuracy and vice versa and the merits of both have to be weighted against each other. The ability to perform selection on the current candidate population based on training populations of previous generations is of great interest to plant breeders. Removing the need for phenotyping shortens the breeding cycle considerably. This could lead to a dramatic increase in the response to selection over time. Un-

fortunately, the more generations training and candidate population are removed, the lower the prediction accuracy becomes. A pressing issue therefore is, for how many generations genomic selection can be performed, before a new training population has to be generated. Some of these issues were addressed in recent studies, based on stochastic simulations and analytic approaches (Lorenz, 2013; Yabe et al., 2013). Such studies can provide valuable hints for practitioners. Ultimately, however, the short- and long-term success of a breeding program over others will determine the 'best' strategy under the circumstances given.
The reward of optimizing prediction methods might be smaller than that of optimizing selection strategies. One does not exclude the other, however. The full potential of genomic selection can be realized only when combining innovative breeding strategies with state of the art prediction methods.

Conclusions

For genomic hybrid prediction, we found that inclusion of dominance into genomic prediction models can lead to considerable gains in prediction accuracy, when the SCA variance is relatively large. Estimation of heterotic group specific effects was found to be of less importance under higher marker density. We explained this with the high linkage phase consistency between Dent and Flint heterotic groups, for markers in close proximity. We showed that combining individuals from different heterotic groups can increase prediction accuracy considerably. This encouraging result, together with the high linkage phase consistency, points the shared genetic base of the Dent and Flint heterotic groups. Using Binomial generalized linear models for genomic prediction

is strongly recommended, when the phenotypic distribution of the data is decidedly non-Gaussian. If in addition the data is heavily overdispersed, the models need to be able to fit overdispersion to attain decent levels of prediction accuracy. Despite the difficulties involved in their use, we argue that tailored and sophisticated prediction methods and models are often needed to exploit the full potential of genomic selection.

Literature cited

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*, 339–50.

Clark, S., J. M. Hickey, and J. H. van der Werf (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet Sel Evol 43*, 18.

Daetwyler, H. D., J. M. Hickey, J. M. Henshall, S. Dominik, B. Gredler, J. H. J. van der Werf, and B. J. Hayes (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim Prod Sci 50*, 1004–1010.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a. Poland, K. Kawamoto, E. Buckler, and S. E. Mitchell (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE 6*, e19379.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc Edin 52*, 399–433.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, US.

Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, A. Charcosset, J. D. Clarke, E.-M. Graner, M. Hansen, J. Joets, M.-C. Le Paslier, M. D. McMullen, P. Montalent, M. Rose, C.-C. Schön, Q. Sun, H. Walter, O. C. Martin, and M. Falque (2011). A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PloS One 6*, e28334.

Gelfand, A. E. and S. K. Sahu (1999). Identifiability, improper priors and gibbs sampling for generalized linear models. *J Am Stat Assoc 94*, 247–253.

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

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica 136*, 245–257.

Guo, T., H. Li, J. Yan, J. Tang, J. Li, Z. Zhang, L. Zhang, and J. Wang (2013). Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines. *Theor Appl Genet 126*, 189–201.

Guzman, P. and K. Lamkey (2000). Effective population size and genetic variability in the BS11 maize population. *Crop Sci 40*, 338–346.

Habier, D., R. L. Fernando, and D. J. Garrick (2013). GenomicBLUP decoded: a look into the black box of genomic prediction. *Genetics 194*, 597–607.

Hawkins, D. M. (2004). The problem of overfitting. *J. Chem. Inf. Comput. Sci. 44*, 1–12.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. Goddard (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet 6*, e1001139.

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

Hu, H., W. Liu, Z. Fu, L. Homann, F. Technow, H. Wang, C. Song, S. Li, A. E. Melchinger, and S. Chen (2013). QTL mapping of stalk bending strength in a recombinant inbred line maize population. *Theor Appl Genet.*.

Kärkkäinen, H. P. and M. J. Sillanpää (2012). Back to basics for bayesian model building in genomic selection. *Genetics 191*, 969–987.

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3 3*, 481–491.

Martin, M., T. Miedaner, B. S. Dhillon, U. Ufermann, B. Kessel, M. Ouzunova, W. Schipprack, and a. E. Melchinger (2011). Colocalization of QTL for gibberella ear rot resistance and low mycotoxin contamination in early european maize. *Crop Sci 51*, 1935–1945.

Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo

(2013). Genomewide predictions from maize single-cross data. *Theor Appl Genet 126*, 13–22.

Ogutu, J. O., H.-P. Piepho, and T. Schulz-Streeck (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC proceedings 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. *BMC proceedings 6 Suppl 2*, S10.

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

Piepho, H. P., J. O. Ogutu, B. Estaghvirou, A. Gordillo, and F. Technow (2012). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Sci 52*, 1093–1104.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rebourg, C., M. Chastanet, B. Gouesnard, C. Welcker, P. Dubreuil, and A. Charcosset (2003). Maize introduction into Europe: the history reviewed in the light of molecular data. *Theor Appl Genet 106*, 895–903.

Reif, J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger (2007). Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics 176*, 1931–1934.

Reif, J. C., Y. Zhao, T. Würschum, M. Gowda, and V. Hahn (2013). Genomic prediction of sunflower hybrid performance. *Plant Breeding 132*, 107–114.

Sprague, G. F. and L. A. Tatum (1942). General vs. specific combining ability in single crosses of corn. *J. Amer. Soc. Agron. 34*, 923–932.

Strandén, I. and D. J. Garrick (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science 92*, 2971–2975.

Technow, F., A. Bürger, and A. E. Melchinger (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3 3*, 197–203.

Technow, F. and A. E. Melchinger (2013). Genomic prediction of dichotomous traits with Bayesian logistic models. *Theor Appl Genet 126*, 1133–1143.

Technow, F., C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet 125*, 1181–1194.

Toro, M. a. and L. Varona (2010). A note on mate allocation for

dominance handling in genomic selection. *Genet Sel Evol 42*, 33.

Wellmann, R. and J. Bennewitz (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet Res Camb 94*, 21–37.

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus (2012). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics 193*, 621–631.

Wittenburg, D., N. Melzer, and N. Reinsch (2011). Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC genetics 12*, 74.

Yabe, S., R. Ohsawa, and H. Iwata (2013). Potential of genomic selection for mass selection breeding in annual allogamous crops. *Crop Sci 53*, 95–105.

Yang, W. and R. J. Tempelman (2012). A Bayesian antedependence model for whole genome prediction. *Genetics 190*, 1491–1501.

Zhao, Y., J. Zeng, R. Fernando, and J. C. Reif (2013). Genomic prediction of hybrid wheat performance. *Crop Sci 53*, 802–810.

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

6. Summary

Marker assisted selection (MAS) was a first attempt to exploit molecular marker information for selection purposes in plant breeding. The MAS approach rested on the identification of quantitative trait loci (QTL). Because of inherent shortcomings of this approach, MAS failed as a tool for improving polygenic traits, in most instances. By shifting focus from QTL identification to prediction of genetic values, a novel approach called 'genomic selection', originally suggested for breeding of dairy cattle, presents a solution to the shortcomings of MAS. In genomic selection, a training population of phenotyped and genotyped individuals is used for building the prediction model. This model uses the whole marker information simultaneously, without a preceding QTL identification step. Genetic values of selection candidates, which are only genotyped, are then predicted based on that model. Finally, the candidates are selected according their predicted genetic values.

Because of its success, genomic selection completely revolutionized dairy cattle breeding. It is now on the verge of revolutionizing plant breeding, too. However, several features set apart plant breeding programs from dairy cattle breeding. Thus, the methodology has to be extended to cover typical scenarios in

plant breeding. Providing such extensions to important aspects of plant breeding are the main objectives of this thesis.

Single-cross hybrids are the predominant type of cultivar in maize and many other crops. Prediction of hybrid performance is of tremendous importance for identification of superior hybrids. Using genomic prediction approaches for this purpose is therefore of great interest to breeders. The conventional genomic prediction models estimate a single additive effect per marker. This was not appropriate for prediction of hybrid performance because of two reasons. (1) The parental inbred lines of single-cross hybrids are usually taken from genetically very distant germplasm groups. For example, in hybrid maize breeding in Central Europe, these are the Dent and Flint heterotic groups, separated for more than 500 years. Because of the strong divergence between the heterotic groups, it seemed necessary to estimate heterotic group specific marker effects. (2) Dominance effects are an important component of hybrid performance. They had to be included into the prediction models to capture the genetic variance between hybrids maximally.

The use of different heterotic groups in hybrid breeding requires parallel breeding programs for inbred line development in each heterotic group. Increasing the training population size with lines from the opposite heterotic group was not attempted previously. Thus, a further objective of this thesis was to investigate whether an increase in the accuracy of genomic prediction can be achieved by using combined training sets.

Important traits in plant breeding are characterized by binomially distributed phenotypes. Examples are germination rate, fertility rates, haploid induction rate and spontaneous chromosome doubling rate. No genomic prediction methods for such

traits were available. Therefore, another objective was to provide methodological extensions for such traits.

We found that incorporation of dominance effects for genomic prediction of maize hybrid performance led to considerable gains in prediction accuracy when the variance attributable to dominance effects was substantial compared to additive genetic variance. Estimation of marker effects specific to the Dent and Flint heterotic group was of less importance, at least not under the high marker densities available today. The main reason for this was the surprisingly high linkage phase consistency between Dent and Flint heterotic groups. Furthermore, combining individuals from different heterotic groups (Flint and Dent) into a single training population can result in considerable increases in prediction accuracy. Our extensions of the prediction methods to binomially distributed data yielded considerably higher prediction accuracies than approximate Gaussian methods.

In conclusion, the developed extensions of prediction methods (to hybrid prediction and binomially distributed data) and approaches (training populations combining heterotic groups) can lead to considerable, cost free gains in prediction accuracy. They are therefore valuable tools for exploiting the full potential of genomic selection in plant breeding.

7. Zusammenfassung

Die markergestütze Selektion (MGS) war ein erster Versuch die Information aus molekularen Markern für Selektionszwecke in der Pflanzenzüchtung nutzbar zu machen. Der MGS Ansatz basierte auf der Identifikation von "quantitative trait loci" (QTL, zu deutsch: Loci mit Effekt auf ein quantitatives Merkmal). Auf Grund inhärenter Defizite schlug der Versuch, MGS für die Verbesserung poligener Merkmale zu verwenden, meistens fehl. Mit einem völlig neuen Ansatz, genomische Selektion genannt und ursprünglich für die Milchrinderzüchtung entwickelt, gelang es, die Defizite der MGS zu überwinden, indem der Schwerpunkt weg von der Identifikation von QTL und hin zur Vorhersage von genetischen Werten gelegt wurde. Für die genomische Selektion wird mit Hilfe einer Kalibrierungspopulation, bestehend aus phenotypisierten und genotypisierten Individuen, ein Vorhersagemodell erstellt. Für dieses Modell wird die Information aller molekularer Marker simultan verwendet, ohne vorhergehende Identifikation von QTL. Mit Hilfe des Vorhersagemodells werden anschließend die genetischen Werte der Selektionskandidaten, die nur genotypisiert wurden, vorhergesagt. Abschließend erfolgt dann die Selektion der Kandidaten anhand der vorhergesagten genetischen Werte.

Aufgrund ihres Erfolges revolutionierte die genomische Selektion bereits die Milchrinderzüchtung. Dieser Prozess hat auch in der Pflanzenzüchtung begonnen. Pflanzenzüchtung und Milchrinderzüchtung unterscheiden sich aber in einigen grundlegenden Aspekten. Auf Grund dessen war es notwendig, die Methodik zu erweitern, um die genomische Selektion für die in der Pflanzenzüchtung typischen Szenarien einsetzen zu können. Es war das Hauptziel dieser Dissertation, eben solche Erweiterungen bereitzustellen.

Einfachkreuzungen sind der dominierende Sortentyp in Mais und vielen anderen Kulturen. Um überlegene Hybriden zu identifizieren, ist die Vorhersage der Hybridleistung von zentraler Bedeutung. Der Einsatz von genomischen Vorhersageverfahren ist daher von großem Interesse für die Pflanzenzüchtung. Die herkömlichen genomischen Vorhersagemodelle schätzen nur einen einzigen, additive Effekt pro Marker. Aus zwei Gründen war dies nicht adäquat für die Vorhersage der Hybridleistung. (1) Die Elternlinien einer Hybride entstammen üblicherweise genetisch sehr verschiedenen Genpools, auch heterotische Gruppen genannt. In der Maishybridzüchtung in Mitteleuropa, sind dies zum Beispiel der Dent- und Flintpool, die nun schon seit mindestens 500 Jahren getrennt sind. Wegen dieser ausgeprägten Divergenz schien es notwendig, spezifische Markereffekte für jede heterotische Gruppe zu schätzen. (2) Dominanzeffekte sind eine wesentliche Komponente der Hybridleistung. Sie mussten daher in die Vorhersagemodelle aufgenommen werden, um die genetische Varianz zwischen den Hybriden so vollständig wie möglich zu erfassen.

Die Verwendung verschiedener heterotischer Gruppen in der Hybridzüchtung erfordert es, für die Linienentwicklung innerhalb der heterotischer Gruppen, parallele Zuchtprogramme zu unterhalten. Es wurde allerdings noch nicht versucht, die Größe der

Kalibrierungspopulation mit Linien der jeweils anderen heterotischen Gruppe zu erhöhen. Ein weiteres Ziel dieser Dissertation war es deshalb, zu untersuchen, ob die Vereinigung verschiedener heterotischer Gruppen in einer Kalibrierungspopulation zu einer Erhöhung der Vorhersagegenauigkeit führen kann.

Einige für die Pflanzenzüchtung wichtige Merkmale sind dadurch gekennzeichnet, dass die phenotypischen Daten einer Binomialverteilung folgen. Beispiele dafür sind Keim-, Fruchtbarkeits- und Haploideninduktionsraten und die Rate der spontanen Chromosomenaufdopplung. Da für diese Art von Merkmal bisher keine Vorhersagemethodik zur Verfügung stand, sollte diese in der vorliegenden Arbeit entwickelt werden.

Unsere Ergebnisse zeigten, dass die zusätzliche Schätzung von Dominanzeffekten die Genauigkeit der vorhergesagten Hybridleistung deutlich erhöhen konnte, wenn die Dominanzvarianz einen wesentlichen Anteil an der gesamten genetischen Varianz darstellt. Wenigstens unter den heute leicht erreichbaren Markerdichten schien es weniger ausschlaggebend, ob für heterotische Gruppen spezifische Markereffekte geschätzt wurden oder nicht. Der Hauptgrund dafür war die überraschend hohe Übereinstimmung in den Kopplungsphasen der heterotischen Gruppen Dent und Flint. Des weiteren konnten wir zeigen, dass die Vereinigung von Linien aus Dent und Flint in einer einzigen Kalibrierungspopulation zu einer beträchtlichen Steigerung der Vorhersagegenauigkeit führen kann. Unsere Erweiterungen der Vorhersagemethodik auf binomialverteilte Daten erzielten im Vergleich zu approximativen gaussianischen Methoden eine deutlich höhere Vorhersagegenauigkeit.

Insgesamt zeigen die erzielten Ergebnisse, dass die in dieser Dissertation entwickelten Erweiterungen der Vorhersagemethoden

(für Vorhersage der Hybridleistung und für binomialverteilte Daten) und -ansätze (Vereinigung von heterotischen Gruppen in einer Kalibrierungspopulation), zu einer beträchtlichen, kostenfreien Erhöhung der Vorhersagegenauigkeit in der genomischen Selektion im pflanzenzüchterischen Kontext führen können. Sie stellen daher ein wertvolles Mittel dar, um das Potential der genomischen Selektion in der Pflanzenzüchtung voll auszuschöpfen.

8. Acknowledgments

First of all I want to thank Prof. Albrecht E. Melchinger for giving me the opportunity to pursue my PhD in his prestigious working group. I am indebted to him for the constant support and guidance I received. I am honored that he trusted me with important tasks since early on and deeply appreciate that he allowed me to find my own way.

I am grateful for the opportunity to meet and work with Prof. Baldev Singh Dhillon. I can only hope to get close to the example of hard, dedicated work he set. Thank you for your patience with a "young scientist"!

I want to thank all my colleagues here at the Institute for creating such a productive and friendly work atmosphere. In particular, I want to thank Christian Riedelsheimer for inspiring discussions during our close interaction throughout the PhD program and Tobias Schrag for allot of work behind the scenes. I am also grateful for the support received from Vilson Mirdita and Manuel Montes during my first steps at the Institute.

Over the years I had the chance to work with more than a dozen Bachelor and Master students doing their thesis with us. I prof-

ited from this experience in many ways and maybe they learned a thing from me or two. I want to thank them for their trust. I especially want to thank my former student Anna Bürger for her close corporation on one of the publications included in this dissertation.

I want to thank Prof. Hans-Peter Piepho and Prof. Jörn Bennewitz for their lectures on statistics, statistical genetics and animal breeding. These lectures had a profound influence on me and helped me to prepare for this PhD and for whatever comes next. I am honored that both agreed to be part of the examination committee.

I further want to acknowledge the financial support from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr Synbreed—Synergistic plant and animal breeding (FKZ: 0315528d).

Curriculum vitae

Name	Frank Technow
Birth	15 July 1984 in Bad Langensalza
School Education	1991–1995, elementary school (Brentano-Grundschule, Bad Langensalza)
	1995–2003, high school (Herman-von-Salza-Gymnasium, Bad Langensalza) Abitur June 2003
Civil service	2003–2004, Schwäbischer Albverein, Stuttgart
Stay abroad	2004–2005, Kibbutz volunteer, Kibbutzim Yiftach and Neot-Smardar, Israel
University education	2005–2008, Agricultural Science, University of Hohenheim, Stuttgart Bachelor of Science June 2008
	2008–2010, Plant Breeding, University of Hohenheim, Stuttgart Master of Science June 2010
	2010–2013, Doctoral student, Plant Breeding and Applied Genetics, University of Hohenheim, Stuttgart

.....
Frank Technow

Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit von mir selbst verfasst wurde 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.

.....
Frank Technow