Theor Appl Genet (2013) 126:1133–1143
DOI 10.1007/s00122-013-2041-9

ORIGINAL PAPER

Genomic prediction of dichotomous traits with Bayesian logistic models

Frank Technow · Albrecht E. Melchinger

Received: 7 September 2012 / Accepted: 21 December 2012 / Published online: 6 February 2013

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 with 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 biparental 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 GLM 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.

Communicated by M. Sillanpää.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2041-9) contains supplementary material, which is available to authorized users.

F. Technow (✉) · A. E. Melchinger
Department of Applied Genetics, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70599, Germany
e-mail: Frank.Technow@uni-hohenheim.de

Introduction

Genomic prediction (Meuwissen et al. 2001) methodology is now well established for quantitative traits that follow approximately a Gaussian phenotypic distribution. However, many important traits in plant breeding are dichotomous. Especially reproductive traits such as haploid induction ability and spontaneous chromosome duplication rate, two traits important for doubled-haploid (DH) production in maize (Prigge et al. 2012; Kleiber et al. 2012), seed emergence (Yousefabadi and Rajabi 2012; Goggi et al. 2007), male and female fertility (Sellamuthu et al. 2011; Dou et al. 2010) and hybrid sterility (Zhao et al. 2006) fall into this category. Using a Gaussian likelihood function here ignores important features of the data, namely its restriction to positive values and the dichotomous nature of the observations.

In human genetics, dichotomous traits, such as outbreak of a disease or not, are commonly observed (Wray et al. 2008) and genomic prediction methodology was already successfully applied to such data sets (Lee et al. 2011). In plant breeding, however, phenotypic observations are usually made on independent, repeated Bernoulli trials. Thus, the underlying phenotypic distribution of the data can be characterized as being Binomial. An extension of the genomic prediction methodology to Binomial

generalized linear models (GLM) with appropriate link function is therefore needed.

A common choice of link function is the logit link. Logistic GLM predict the log-odds ratio of observing a certain outcome (e.g., a seed being haploid). A great advantage of logistic GLM is that they allow convenient and efficient Gibbs-sampling computations also for Binomial data, when using the auxiliary mixture sampling parametrization developed by Frühwirth-Schnatter et al. (2009).

A phenomenon commonly associated with Binomial and other types of count data is overdispersion (Dey et al. 1997). Overdispersion means that the data observed are more heterogeneous than expected under a Binomial model, in which case the predicted variance is a direct function of the conditional mean (conditional on the set of predictors, e.g., the markers). The linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) is seldom complete. Overdispersion can, therefore, always be a problem in genomic prediction because under incomplete LD, the Binomial sampling process is not the only source of uncertainty. This is especially the case when phenotyping is done in field trials, where in addition to genetic differences also non-genetic sources of variation are present that cannot be fully accounted for. Under overdispersion, a simple Binomial model will not fit the data optimally, which can reduce the prediction accuracy. Genomic prediction methodology for Binomial phenotypes therefore needs to be able to account for overdispersion for successful application to real world data sets.

Bayesian methods are a popular choice for genomic prediction of genotypic values (Kärkkäinen and Sillanpää 2012). They might be coarsely separated into (1) marker effects methods and (2) polygenic or total genetic effects methods, where genetic effects are associated directly with individuals (Kärkkäinen and Sillanpää 2012). The latter might be understood as Bayesian versions of the popular non-Bayesian GBLUP method. Numerous results point to a superiority of GBLUP (both Bayesian and non-Bayesian) when the trait is controlled by a large number of QTL with small effects and to a superiority of marker effects methods for traits with a more oligogenic architecture (Kärkkäinen and Sillanpää 2012; Hayes et al. 2010; Clark et al. 2011; Zhong et al. 2009). However, such a comparison is lacking for traits with a Binomial phenotypic distribution.

Binomial GLM are the only statistically valid way of analyzing Binomial data. However, practitioners applying genomic prediction are mostly interested in identifying superior genotypes for selection purposes. For this, a standard Gaussian GLM might provide a useful ad hoc approximation after an appropriate transformation of the data. Furthermore, Binomial GLM can be associated with considerable implementational and computational overhead and complexity. Therefore, it seems worthwhile to investigate whether and under which circumstances Gaussian GLM are sufficient for practical applications.

Our objectives were to (1) implement logistic GLM for genomic prediction of dichotomous traits with Binomial phenotypic distribution that can account for overdispersion and (2) compare for these traits the performance of Bayesian GBLUP and marker effect-based methods. Thereby we based our investigation on real and simulated plant breeding data sets.

Materials and methods

We consider a segregating population of genotypes such as F_2 individuals or DH lines from a bi-parental cross, the genotypes of which are genotyped and progeny of them are either produced by self-pollination or cross-pollination with a common tester. Let n_i be the number of progeny derived from genotype i. Each offspring was phenotyped for a dichotomous trait, receiving a phenotypic value of either one or zero, depending on which of two possible events occurred (e.g., seed viable or not, seed haploid or not). Thus, each offspring can be viewed as an independent Bernoulli trial. The sum s_i over all n_i offspring is the phenotypic score of genotype i, which consequently follows a Binomial distribution.

Marker effects methods

To analyze this kind of data, we first used a Binomial GLM based on marker effects

$$s_i \sim \mathscr{B}(p_i, n_i)$$
$$p_i = \mathscr{L}(\beta_0 + \boldsymbol{X_i u}), \quad (1)$$

where p_i is the probability of observing the "successful" outcome for the ith individual and $\mathscr{L}(\beta_0 + \boldsymbol{X_i u})$ its linear predictor, with $\mathscr{L}(\cdot)$ denoting the logit link function. The intercept is denoted by β_0. The row vector $\boldsymbol{X}_i$ is a known marker genotype incidence vector of the ith individual, for the additive marker effects in $\boldsymbol{u}$. The whole matrix $\boldsymbol{X}$ has dimensions $N \times M$ (N = number of individuals, M = number of biallelic markers). The marker genotypes were coded with 1 and -1 for the two homozygous genotypes and 0 for the heterozygous genotype and were scaled and centered prior to analysis by subtracting the mean and dividing by the standard deviation, a common practice in regression models as discussed by de los Campos et al. (2012). The term $\mathscr{B}(\cdot)$ denotes the Binomial probability function with n_i being known.

The auxiliary mixture sampling method (Frühwirth-Schnatter et al. 2009) was used to facilitate Gibbs-sampling. Briefly, in auxiliary mixture sampling, an aggregated

latent variable $y_i^* = log(\beta_0 + \boldsymbol{X_i u}) + \boldsymbol{\epsilon_i}$ is introduced for each Binomial observation. The distribution of the residuals ϵ_i is a negative log-gamma distribution, which is approximated by a Gaussian mixture distribution. A second latent variable r_i is introduced as an indicator of the component of the mixture. Then, conditional on y_i^* and r_i, model (1) reduces to a linear model with Gaussian likelihood function, y_i^* as response and heteroscedastic but fixed residual variances, which are determined by r_i. The major advantage of this procedure is that then standard Gibbs-sampling methodology developed for Gaussian models can be used.

The parameters (i.e., weights, means and variances) of the components of the Gaussian mixture distributions were precomputed and remained fixed during Gibbs-sampling. We used the Matlab function "compute_mixture", obtained by request to Frühwirth-Schnatter et al. (2009), for precomputing these parameters. For convenience, we provide them tabulated up to $n_i = 3{,}500$ in supplemental file S1.

The joint hierarchical prior distribution was

$$p(\beta_0) \times p(\boldsymbol{u_j}|\boldsymbol{\sigma^2_{u_j}}) \times \boldsymbol{p}(\boldsymbol{\sigma^2_{u_j}}|\boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{S^2}) \times \boldsymbol{p}(\boldsymbol{\pi}|\boldsymbol{a}, \boldsymbol{b}) \times \boldsymbol{p}(\boldsymbol{S^2}) \times \boldsymbol{p}(\boldsymbol{\nu}).$$

The priors for β_0 and the marker effects were $p(\beta_0) \propto 1$ and $p(\boldsymbol{u}_j|\sigma^2_{u_j}) = \mathcal{N}(0, \sigma^2_{u_j})$.

The prior variance of the effect of the jth marker ($\sigma^2_{u_j}$) was

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

From Eq. (2) follows that our method falls into the class, the priors for the marker effects of which can be parametrized as Student's t distributions with a non-zero probability mass over zero. Specifically, because the prior probability mass over zero is introduced through $\sigma^2_{u_j}$ instead of through an explicit indicator variable, our method belongs to the "BayesB" class (Meuwissen et al. 2001). Therefore, it is referred to as "Binomial BayesB GLM" in the remainder of this treatise.

Following Yang and Tempelman (2012) and Technow et al. (2012), the hyperparameters π, ν and S^2 were associated with prior distributions too and, thus, were estimated from the data. The prior of S^2 was a Gamma distribution with shape and rate parameter equal to 0.1, and for ν we used the uninformative improper prior distribution

$$p(\nu)\begin{cases} \propto (\nu+1)^{-2} & \text{if } 0 < \nu < 100 \\ = 0 & \text{else.} \end{cases} \quad (3)$$

The prior for π was a Beta distribution. Its parameters were different for each data set and are specified together with their description given below.

We employed finite mixture models to account for overdispersion (Frühwirth-Schnatter 2006). Model (1) now generalizes to

$$s_i \sim \sum_{k=1}^{K} \eta_k \mathscr{B}(p_{ik}, n_i)$$
$$p_{ik} = \mathscr{L}(\beta_{0,k} + \boldsymbol{X_i u_k}), \quad (4)$$

where k indexes the mixture component, K is the number of components used and η_k the weight of the kth component. The marker genotypes were not scaled and centered, because the sampling algorithm used would have required a constant re-centering and re-scaling, which would have been computationally prohibitive. The number of components K is a constant, but we fitted the model with K ranging from 2 to 12 and report results for the value of K that gave best results. Further note that for $K = 1$, model (4) reduces to the standard Binomial GLM in (1).

The joint prior distribution then is indexed by k as well and a uniform Dirichlet distribution with concentration parameter $\alpha_1, \ldots, \alpha_K = 4$ was used as prior for the weights η_k. We used the data augmentation technique as described by Fruhwirth-Schnatter (2006, sect. 3.5) for sampling from model (4). Here, a group indicator $S_i \in \{1, 2, \ldots, K\}$, is introduced as missing data. Then, the parameters and hyperparameters for the kth mixture component are sampled conditional on knowing S_i (i.e., from all observations in group $S_i = k$) and S_i conditional on knowing the parameters and hyperparameters. Adopting this data augmentation method has the advantage that conditional on S_i, the parameters can be sampled with standard Gibbs-sampling methodology, while sampling S_i conditional on the parameters reduces to a straightforward classification problem.

As a baseline method for comparison purposes, we also fitted a Gaussian BayesB GLM commonly used for Gaussian data in the literature (e.g., Technow et al. 2012; Yang and Tempelman 2012). The model was

$$w_i \sim \mathcal{N}(p_i, \sigma^2_e)$$
$$p_i = \beta_0 + \boldsymbol{X_i u} \quad (5)$$

Here, $w_i = arcsin(\sqrt{s_i/n_i})$, $\mathcal{N}(\cdot)$ denotes the Gaussian density function and σ^2_e the residual variance. We scaled and centered w_i prior to the analysis. Note that p_i is not confined to the probability scale as for the Binomial GLM above. The same joint hierarchical prior distribution as for the Binomial BayesB GLM was used here. The additional parameter σ^2_e was associated with an uninformative scaled inverse Chi-square prior.

Samples from the joint posterior distribution of the parameters were drawn by Gibbs-sampling, with a single chain of 250,000 iterations, of which the first 100,000 were

discarded as burn-in and only samples from every 50th iteration were stored. Details on the sampling strategy for auxiliary mixture sampling can be found in Frühwirth-Schnatter et al. (2009) and details on data augmentation for sampling from finite mixture models in Frühwirth-Schnatter (2006). The fully conditional distributions (FCD) and Gibbs-sampling strategy for BayesB are described in Yang and Tempelman (2012). For sampling from the FCD of $\sigma^2_{u_j}$ and ν, Metropolis–Hastings algorithms were used as described by Technow et al. (2012). For the standard Binomial BayesB GLM as well as the Gaussian BayesB GLM, the posterior means of β_0 and the marker effects $\boldsymbol{u}_j$ were used as point estimates for predicting genotypic values.

In finite mixture models, the estimates of the posterior means might be sensitive to "label switching" (Frühwirth-Schnatter 2006). Therefore, we used the mean of the posterior predictive distribution for predicting the genotypic values of new observations. The posterior predictive distribution is robust against label switching (Frühwirth-Schnatter 2006).

All BayesB algorithms were implemented as C routines compatible with the R software environment (R Development Core Team 2011). The source code is provided in supplemental file S2. The R package "coda" (Plummer et al. 2010) was used to estimate effective sample sizes (ESS) of marker effects for the standard Binomial and Gaussian GLM. Because of "label switching", there seems to be no straightforward way to estimate an ESS for our finite mixture models.

Bayesian GBLUP methods

The Binomial GBLUP GLM was

$$s_i \sim \mathscr{B}(p_i, n_i)$$
$$p_i = \mathscr{L}(\beta_0 + a_i), \tag{6}$$

with a_i denoting the total genetic effect of the ith individual.

We used the JAGS Gibbs-sampling environment (Plummer 2003) for GBLUP, which allows for convenient specification and implementation of standard models. The JAGS environment is very similar to the BUGS/OpenBUGS environment (Thomas et al. 2006), but platform independent and designed for integration with R. JAGS uses auxiliary mixture sampling as well, however, its implementation might differ somewhat from ours used for BayesB.

The joint hierarchical prior distribution was

$$p(\beta_0) \times p(a_i|\sigma^2_a) \times p(\sigma^2_a) \tag{7}$$

Again we used an uniform prior for β_0. The prior for the total genetic effects was $\mathcal{MVN}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\sigma}^2_{\boldsymbol{a}})$, where $\boldsymbol{A}$ is the genomic relationship matrix, computed according to Method 1 of VanRaden (2008) and σ^2_a the additive genetic variance component. We associated $1/\sigma^2_a$ with an uninformative Gamma prior with shape and rate parameters equal to 0.01.

In this case too, we considered a Gaussian GBLUP GLM version as

$$w_i \sim \mathcal{N}(p_i, \sigma^2_e)$$
$$p_i = \beta_0 + a_i. \tag{8}$$

The set-up for the joint hierarchical prior distribution was identical to the Binomial GBLUP GLM, with $1/\sigma^2_e$ also associated with an uninformative Gamma prior with shape and rate parameters equal to 0.01.

Sampling was done by running three independent Gibbs-sampling chains for 10,000 iterations each. The first 5,000 iterations of each chain were discarded as burn-in and only samples from every 3rd iteration stored afterwards. The JAGS source code is provided in supplemental file S3.

Wheat female fertility data set

The data comprises a bi-parental wheat (*Triticum aestivum* L.) F_2 population of size 243, genotyped with 28 markers. The Binomial phenotypic scores are the number of seeded spikelets (s_i) from a total number of spikelets per plant (n_i). The average s_i was 19.13, and the average n_i was 25.15. For the majority of plants, the ratio s_i/n_i was around 0.9, some plants had a ratio close to or exactly zero and there were only few intermediate observations. We therefore performed the analysis additionally for a subset of the data with plants for which $s_i/n_i > 0.75$ ("high fertility subset"). The number of remaining observations was 186, with an average s_i of 23.88 and an average n_i of 25.04. The phenotypic and the marker data were obtained by request to the authors of Che and Xu (2012), who previously analyzed the data. A more detailed description of the data set is given in Dou et al. (2009).

Given the rather small number of markers, the Beta distribution prior for π was $Beta(a = 1, b = 9)$, meaning that almost all markers are expected to have an effect a priori. We used fivefold cross-validation (CV) for assessing the performance of our models. Here, the whole data set is split into five distinct subsets and each subset is in turn predicted by a model fitted using the data from the remaining four subsets as training set. The statistics of interest are determined each time and averaged at the end over the five runs. This whole process was repeated 25 times, with independent random splits each time. All models were fitted on the same subsets and the differences between them assessed for significance using standard

frequentist paired t tests. The recorded statistic was the Pearson correlation coefficient of predicted and observed phenotypic values in the prediction set (predictive ability). We refrained from calculating prediction accuracies commonly obtained by dividing the predictive ability by the square root of the trait heritability (h^2) because obtaining sensible estimates of h^2 for dichotomous traits is not trivial if n_i varies.

Maize haploid induction rate data set

In maize (*Zea mays* L.), DH lines are generated in-vivo by pollinating source germplasm with pollen from so-called "inducer" genotypes. A proportion of the so obtained seeds then contain embryos with haploid genome of maternal origin. The proportion of haploid seeds in the total number of seeds produced is commonly referred to as the haploid induction rate (HIR).

Our example data set comprised a bi-parental experimental maize F_2 population used in a recently published study on QTL mapping for HIR (Prigge et al. 2012). The parents of the population were the European inducer line UH400 and the Chinese inducer line CAUHOI. UH400 has a HIR of around 0.08 (Prigge et al. 2012), CAUHOI of around 0.02 (Li et al. 2009). The $N = 185$ F_2 individuals were genotyped with 90 polymorphic simple sequence repeat (SSR) markers. The HIR was determined by pollinating a tester genotype with pollen from each F_2 individual and counting the number of haploid seeds in the progeny. The average number of pollinated test cross seeds per F_2 individual was 1,108 but ranged from below 166 to 3,279. The average HIR was 0.053. The marker data were retrieved from the supporting material of Prigge et al. (2012), and the phenotypic data were obtained by request to the authors of cited publication.

As previous research suggests an oligogenic or even monogenic inheritance of HIR (Barret et al. 2008; Lashermes et al. 1988), we chose $Beta(a = 8, b = 2)$ as prior for π, which concentrates most of the probability mass around 0.8. Here, we used 25 times repeated tenfold CV, which results in larger training sets as compared with fivefold CV.

Simulated data set

For investigating the influence of factors such as genetic architecture, population size and number of Bernoulli trials, we simulated a bi-parental DH population with size 500. The genome consisted of 10 chromosomes, each of 100 cM length. There were 50 equally spaced markers and 15 QTL per chromosome. The meiosis events for generating the DH lines in silico were simulated according to the Haldane mapping function, using the R package "hypred" (Technow 2011). The polygenic trait architecture was simulated by assigning additive effects, defined according Falconer and Mackay (1996), drawn from a standard Gaussian distribution to all 150 QTL. To simulate an oligogenic architecture with small and large QTL, we assigned additive effects drawn from a Gamma distribution with parameters scale = 1.66 and shape = 0.4 (Meuwissen et al. 2001) to a random subset of five of the QTL per chromosome and effects of zero to the remaining QTL. The QTL effects were summed according to the QTL genotypes of each DH line to create a raw genotypic score g_i, which was additionally scaled and centered. These scores were then transformed to the probability scale by computing $p_i = \phi(\phi(g_i/max(g_i)_{i=1\ldots500}) - 2.12)$, where ϕ is the standard Gaussian cumulative distribution function. This transformation assured that the true probability parameters of the genotypes were within the interval [0, 0.1]. Phenotypes were simulated by drawing the number of observed events s_i from a Binomial distribution with probability parameter equal to p_i. The number of independent Bernoulli trials n_i was $n_i^* + 1$, where n_i^* is a random number from a Poisson distribution. The rate parameter λ of this distribution was set to 25, 50, 100, 250 and 500. For each combination of trait architecture and λ, we created 50 independent subdivisions of the data into training and prediction set. The training set was used for fitting the model. As sizes of the training set, we considered $N = 100$, 200 and 300. The true genotypic values p_i were known from simulation. Thus, we could compute the prediction accuracy of the models as the Pearson correlation between predicted and true p_i values for the prediction set individuals.

To avoid incorporating information on the genetic architecture that might be unavailable in practice, we used $Beta(a = 0.9, b = 0.9)$ as prior for π. This prior has its probability peaks close to zero and close to one. It suggests that we either expect a polygenic or oligogenic trait architecture, without completely ruling out anything in between. We observed better convergent properties with this prior than with a completely uniform prior. To facilitate computations, the number of independent Gibbs-sampling chains for GBLUP was reduced to 1.

The marker and QTL genotypes of the 500 individuals, their phenotypes and true probability parameters as well as the simulated QTL effects are provided as supplemental file S4.

We further performed a simulation to investigate the effects of overdispersion on the prediction accuracies and to compare the standard Binomial BayesB GLM with the finite mixture implementation. For this we used the marker data described above together with the n_i values corresponding to $\lambda = 25$ and the p_i values from the oligogenic

trait. Overdispersion was simulated according to a Beta-Binomial sampling process by drawing p_i^* from a Beta distribution with parameters $\alpha = \kappa\ p_i$ and $\beta = \kappa(1 - p_i)$. Then we redrew s_i from a Binomial distribution with probability parameter equal to p_i^* and size parameter equal to n_i, as shown above. Thus, p_i^* was drawn from a Beta distribution with mean equal to p_i and a standard deviation (SD) that is determined largely by κ. The smaller κ, the higher the SD and thereby the degree of extra-binomial variation or overdispersion. For the simulations, we considered $\kappa = 10$ (100, 1,000). The exact value of the SD also depends on p_i, but to give an example, for $p_i = 0.05$, $\kappa = 10$ results in a SD of 0.066, $\kappa = 100$ in a SD of 0.022 and $\kappa = 1{,}000$ in a SD of just 0.007. The simulation was repeated 50 times for each combination of κ and N (for which we again considered $N = 100$ (200, 300)).

Identifiability is always a problem in GLM when the number of parameters is larger than the number of observation, i.e., when $N < M$. However, by using proper, informative prior distributions, parameter estimation, Bayesian learning and posterior prediction is still possible (Gelfand and Sahu 1999). The huge success of Bayesian marker effects methods for genomic prediction (Kärkkäinen and Sillanpää 2012) shows that especially posterior prediction seems not to be affected by the lack of identifiability in these cases. However, as discussed by Frühwirth-Schnatter (2006), finite mixture regression models suffer especially from nonidentifiability due to the added complexity and flexibility. Consequently, trying to fit model (4) with the full set of $M = 500$ markers led to non-convergent Gibbs-sampling chains and non-sense results. We therefore reduced the number of markers to 250 for $N = 100$ to 300 for $N = 200$ and to 350 for $N = 300$ by randomly sampling from the full marker set. For direct comparison, the standard Binomial GLM (1) was fitted with these reduced marker sets as well and in addition with the full set of $M = 500$ markers.

Results

Wheat female fertility data set

For the full data set, a finite mixture Binomial BayesB GLM with $K = 12$ gave the highest average predictive ability of 0.555. The predictive abilities observed for the other models were significantly lower at around 0.500 (Table 1).

For the "high fertility subset", the highest average predictive ability of 0.202 was observed for the standard Binomial BayesB GLM (Table 1). The Binomial GLM had generally significantly higher predictive abilities than their Gaussian counterparts, for both the BayesB as well as the GBLUP methods. For both the standard Binomial and the Gaussian GLM, method BayesB had a significantly higher predictive ability than GBLUP.

Maize haploid induction rate data set

Here, the highest average predictive ability was 0.684 observed for the Gaussian BayesB GLM (Table 1). For this data set, the Gaussian GLM had significantly higher predictive abilities than their Binomial counterparts. BayesB had again a significantly higher predictive ability than GBLUP, for both the Binomial and the Gaussian GLM. A finite mixture Binomial BayesB GLM with $K = 2$ yielded a slightly but significantly higher predictive ability than the standard Binomial BayesB GLM.

The ESS for marker effects (BayesB) and total genetic effects (GBLUP) for the Gaussian models were in most cases close to the actual sample sizes (3,000 for BayesB and 5,000 for GBLUP) (Table 2). The ESS for the Binomial models, however, were considerably lower than the actual sample sizes but always above 400.

Simulated data sets

Results for the polygenic trait architecture showed that the average prediction accuracy increased with increasing N, from 0.686 at $N = 100$, to 0.781 at $N = 200$ and to 0.819 at $N = 300$ and with increasing λ, from 0.567 at $\lambda = 25$, to 0.801 at $\lambda = 100$ and to 0.898 at $\lambda = 500$ (Table 3). The increase in prediction accuracy with increasing λ thereby depended on the size of N. For example, the average increase from $\lambda = 25$ to $\lambda = 100$ amounted to 0.282 at $N = 100$ and 0.230 at $N = 200$, but just 0.190 at $N = 300$. Analogously, the increase in prediction accuracy with increasing N depended on the level of λ. For example, the increase of N from 100 to 200 increased the prediction accuracy by 0.156 at $\lambda = 25$, by 0.105 at $\lambda = 100$, but only by 0.050 at $\lambda = 500$.

Binomial GLM tended to have higher prediction accuracies than Gaussian GLM for $\lambda < 250$, but for greater values of λ the prediction accuracy of the Gaussian GLM was equal or higher (Table 3). The superiority of Binomial over the Gaussian GLM was highest for $\lambda = 25$. GBLUP had in most cases a slightly higher prediction accuracy than BayesB, but there were no obvious trends regarding the relative superiority of BayesB and GBLUP with respect to N or λ.

Virtually, the same trends with regard to N and λ as for the polygenic trait architecture were observed for the oligogenic trait architecture (Table 3). In the case of BayesB, the Binomial GLM had always significantly higher prediction accuracy compared with Gaussian GLM, with greater differences for lower values of λ. For GBLUP,

Table 1 Average predictive ability of the wheat female fertility data set, the high fertility subset of this data set and maize haploid induction data from 25 replications of fivefold (wheat) and tenfold (maize) cross-validation for the Binomial and Gaussian BayesB and GBLUP generalized linear models (GLM)

Data set	BayesB			GBLUP	
	Binomial ($K = 1$)	Binomial ($K > 1$)	Gaussian	Binomial	Gaussian
Wheat female fertility	0.511[a]	0.555[b]	0.515[ad]	0.505[ed]	0.495[e]
Wheat high fertility subset	0.202[a]	0.180[b]	0.177[bc]	0.168[c]	0.134[d]
Maize haploid induction	0.654[a]	0.667[b]	0.684[c]	0.649[d]	0.666[b]

For BayesB, the Binomial GLM was fitted both without ($K = 1$) and with accounting for overdispersion ($K > 1$). For the wheat female fertility data, $K = 12$, for the high fertility subset from this data set, $K = 2$, and for the maize haploid induction data set, $K = 2$. Values within a row with common letters are not statistically different at an α level of 0.05 in paired t tests

Table 2 Mean effective sample size of marker effects (BayesB) and total genetic values (GBLUP), for the standard Binomial and Gaussian BayesB and GBLUP GLM, averaged over cross-validation runs for the wheat female fertility and maize haploid induction data sets

Data set	BayesB		GBLUP	
	Binomial	Gaussian	Binomial	Gaussian
Wheat female fertility	870	2,741	532	3,847
Wheat high fertility subset	1,188	2,872	408	2,399
Maize haploid induction	433	3,000	407	4,882

Binomial GLM had higher prediction accuracy until $\lambda = 100$, after which Gaussian GLM had prediction accuracies of equal or slightly higher size. On average (across Binomial and Gaussian GLM), BayesB was only slightly superior to GBLUP. However, the differences between the Binomial BayesB GLM, which had the highest prediction accuracy in all cases, and the Binomial GBLUP GLM were considerably greater than this.

For the simulated overdispersion data sets, by far the lowest prediction accuracies were observed for $\kappa = 10$, i.e., for strong overdispersion (Table 4). The highest prediction accuracies were observed for $\kappa = 1{,}000$, but the average difference to $\kappa = 100$ was with 0.052 marginal compared with the average difference between $\kappa = 10$ and $\kappa = 100$, which was 0.288. The average prediction accuracies increased with increasing N, from 0.435 at $N = 100$ to 0.549 at $N = 200$ and to 0.609 at $N = 300$ (Table 4).

The standard Binomial BayesB GLM fitted with $M = 500$ tended to have slightly higher prediction accuracies than the same model fitted with the reduced marker sets (Table 4). Because the differences were marginal, the comparison between models will be focused on the comparison between the standard GLM with $M = 500$ and the finite mixture Binomial GLMs (with $K = 2$). For $\kappa = 10$, the highest prediction accuracies were observed for the finite mixture Binomial GLM (Table 4). The difference thereby increased with increasing N, from 0.044 at $N = 100$, to 0.130 at $N = 200$, to 0.171 at $N = 300$. The increase in difference thereby mostly came from an increased prediction accuracy of the finite mixture Binomial GLM. The prediction accuracy of the standard Binomial GLM changed only marginally. For $\kappa = 100$ and $\kappa = 1{,}000$, however, the standard Binomial GLM had higher (at $N = 100$ and $N = 200$) or virtually equal (at $N = 300$) prediction accuracies (Table 4).

Discussion

Influence of training set size and number of Bernoulli trials

Prediction accuracies increased with increasing N, as expected and already observed in previous research for Gaussian traits (Zhong et al. 2009). We also observed a considerable increase in prediction accuracy with increasing λ, i.e., n_i. The higher n_i, the better will s_i represent the true probability parameter p_i of the individual. Thus, an increase in λ will have the same effect as an increase in h^2, which was previously recognized as major factor influencing prediction accuracy (Villumsen et al. 2009).

Interestingly, a high λ could almost compensate for low N. For example, for $N = 100$ and $\lambda = 500$ the prediction accuracies were as high or even higher than at $N = 300$ and $\lambda = 100$. An elaborate study about the optimal allocation of resources, taking the relative costs of increasing N and λ into account, should be conducted to investigate the optimal combination of N and λ under a restricted total budget. We speculate that λ should always be raised to high values, because this would be absolutely neutral with respect to genotyping costs. However, we recognize that λ is biologically constrained in many cases, for example by the number of progeny seeds of a plant. We limited ourselves to p_i within the interval [0, 0.1]. The effect of λ will likely be smaller for values of p_i closer to 0.5, where a Gaussian distribution will be approached already for much

Table 3 Average prediction accuracy over 50 replications for the simulated data set with polygenic and oligogenic trait architecture, for the Binomial and Gaussian BayesB and GBLUP generalized linear models

N	λ	BayesB		GBLUP	
		Binomial	Gaussian	Binomial	Gaussian
Polygenic architecture					
100	25	0.452^{a}	0.420^{b}	0.452^{a}	0.432^{a}
	50	0.590^{a}	0.558^{b}	0.623^{c}	0.592^{a}
	100	0.726^{a}	0.694^{b}	0.736^{c}	0.727^{a}
	250	0.810^{ab}	0.800^{a}	0.820^{b}	0.829^{c}
	500	0.863^{a}	0.855^{b}	0.865^{a}	0.868^{d}
200	25	0.609^{a}	0.583^{bc}	0.607^{a}	0.582^{c}
	50	0.715^{a}	0.696^{b}	0.729^{c}	0.713^{a}
	100	0.829^{a}	0.809^{b}	0.836^{c}	0.830^{a}
	250	0.865^{a}	0.856^{b}	0.866^{a}	0.875^{c}
	500	0.909^{a}	0.906^{b}	0.908^{a}	0.911^{c}
300	25	0.687^{a}	0.658^{b}	0.674^{c}	0.650^{d}
	50	0.762^{a}	0.739^{b}	0.774^{c}	0.749^{d}
	100	0.858^{a}	0.848^{b}	0.861^{c}	0.857^{a}
	250	0.894^{a}	0.891^{b}	0.893^{a}	0.900^{c}
	500	0.925^{a}	0.923^{b}	0.924^{b}	0.926^{a}
Oligogenic architecture					
100	25	0.613^{a}	0.537^{b}	0.607^{a}	0.555^{c}
	50	0.704^{a}	0.637^{b}	0.693^{c}	0.654^{d}
	100	0.803^{a}	0.768^{b}	0.787^{c}	0.790^{c}
	250	0.866^{a}	0.805^{b}	0.838^{c}	0.833^{d}
	500	0.917^{a}	0.872^{b}	0.878^{c}	0.877^{c}
200	25	0.746^{a}	0.706^{b}	0.738^{a}	0.700^{bc}
	50	0.778^{a}	0.731^{b}	0.770^{c}	0.738^{b}
	100	0.874^{a}	0.841^{b}	0.844^{bc}	0.846^{c}
	250	0.927^{a}	0.879^{b}	0.891^{c}	0.892^{c}
	500	0.947^{a}	0.918^{b}	0.918^{bc}	0.919^{c}
300	25	0.796^{a}	0.750^{b}	0.780^{c}	0.740^{d}
	50	0.819^{a}	0.790^{b}	0.814^{a}	0.783^{c}
	100	0.915^{a}	0.880^{b}	0.880^{b}	0.881^{b}
	250	0.946^{a}	0.910^{b}	0.914^{c}	0.915^{d}
	500	0.958^{a}	0.936^{b}	0.936^{bc}	0.935^{c}

The training set size is denoted by N, the parameter of the Poisson distribution used for drawing n_i is denoted as λ. Values within a row with common letters are not statistically different at an α level of 0.05 in paired t tests

lower values of λ. However, while increasing λ has the effect of better representing the information in the data provided by the N biological replicates, the actual amount of information can only be increased by increasing N. Conversely, if N and thereby the information content of the data becomes too low, the prediction accuracy will inevitably deteriorate strongly, regardless of how high λ is. Therefore, the most critical factor remains N. This is apparent from the fact that even at very low λ, accurate prediction is possible when N is high. This is in fact the typical scenario encountered in human genetic studies, where $n_i = 1$ but often $N \gg 1{,}000$ (Wray et al. 2008; Lee et al. 2011).

Table 4 Average prediction accuracy over 50 replications for the simulated data set with overdispersion using the Binomial BayesB GLM

N	κ	$K = 1$		$K = 2$
		$M = 500$	$M < 500$	
100	10	0.278^{ab}	0.267^{a}	0.322^{b}
	100	0.505^{a}	0.490^{b}	0.420^{c}
	1,000	0.577^{a}	0.571^{a}	0.485^{b}
200	10	0.263^{a}	0.266^{a}	0.393^{b}
	100	0.665^{a}	0.659^{a}	0.634^{b}
	1,000	0.708^{ac}	0.699^{bd}	0.661^{cd}
300	10	0.310^{a}	0.315^{a}	0.481^{b}
	100	0.706^{a}	0.703^{a}	0.706^{a}
	1,000	0.757^{a}	0.752^{b}	0.751^{ab}

The training set size is denoted by N, the parameter controlling the overdispersion strength is κ. The number of mixture components is denoted by K, with $K = 1$ indicating that overdispersion is not modeled and $K = 2$ that it is.The value of the rate parameter of the Poisson distribution used for drawing n_i was $\lambda = 25$. For the $K = 2$ models, the number of markers was reduced to $M = 250$ (300, 350) at $N = 100$ (200, 300). The $K = 1$ models were fitted with the full data set of $M = 500$ markers as well as with the reduced set of markers used for the $K = 2$ models ($M < 500$). Values within a row with common letters are not statistically different at an α level of 0.05 in paired t tests

Model comparison based on simulated data sets

As expected, BayesB tended to be superior under an oligogenic trait architecture and GBLUP under a polygenic trait architecture. Numerous researchers found similar results for traits displaying a Gaussian phenotypic distribution (Kärkkäinen and Sillanpää 2012; Hayes et al. 2010; Clark et al. 2011; Zhong et al. 2009). Kärkkäinen and Sillanpää (2012) previously also reported that Bayesian marker effect models are superior to GBLUP models under an oligogenic trait architecture for binary traits (i.e., with $n_i = 1$ for all i). Thus, we can confirm these observations for Binomial phenotypic distributions, too. However, the differences between BayesB and GBLUP were rather small, especially under the polygenic trait architecture. The choice between BayesB and GBLUP might therefore be driven by convenience considerations regarding computation and implementation. However, computational requirements of Binomial GBLUP GLM were not necessarily lower than those of Binomial BayesB GLM in our study. We used MCMC algorithms also for the Gaussian models. We are aware that approximate but fast expectation–maximization

algorithms are available, for which computation times are almost negligible (Kärkkäinen and Sillanpää 2012). Nevertheless, it seems that our BayesB method, with hyperpriors on hyperparameters, cannot be fitted by them (Kärkkäinen and Sillanpää 2012). We furthermore decided against their use to be able to compare Binomial and Gaussian GLM on exactly the same terms. Recently, an improved auxiliary mixture sampler for logistic GLM was developed, which has the potential of substantially decreasing the computation times for the Binomial GLM due to increased efficiency (Fussl et al. 2012).

The greatest gains in prediction accuracy by using Binomial GLM were observed for smaller values of λ, where the distribution of the data is definitely non-Gaussian, and for p_i values far removed from 0.5. For higher values of λ, the distribution of the data will rapidly approach a Gaussian distribution. Consequently, results from Gaussian and Binomial GLM converged for the highest values of λ considered, at least under the polygenic trait architecture. Under the oligogenic trait architecture, the best Binomial model (BayesB) remained consistently superior over its Gaussian counterparts, albeit with reduced differences. We speculate that this is so because usage of the correct model and likelihood function is more important when estimating effects of single markers than when estimating total genetic effects of individuals, because the latter might be more robust with regard to the distribution underlying the data. With true probability parameters p_i closer to 0.5, the distribution of the data would approach a Gaussian distribution earlier, i.e., for lower values of λ. However, the range of p_i chosen by us reflects the most interesting situation of traits deviating substantially from a Gaussian distribution, where finding better alternatives than the standard Gaussian GLM is most important. This range seems also of greatest relevance in practice. For example, both experimental data sets used in this study exhibit probability parameters close to 1.0 or 0.0. This will also be the case for traits such as seed emergence, where the emergence rate of reasonably well-adapted material almost always exceeds 0.90 (Goggi et al. 2007).

The standard Gaussian GLM used commonly in genomic prediction, which we used as baseline for comparison with our Binomial GLM, make the simplifying assumption of homogeneous residual variances. Thus, they ignore that some individuals will have been phenotyped more precisely than others, depending on the size of n_i. Our Binomial GLM automatically incorporate the differences in n_i, via heteroscedastic residual variances (Frühwirth-Schnatter et al. 2009). The resulting weighing of the observations in the training set by n_i is therefore another advantage of Binomial GLM over the standard Gaussian GLM. However, in Gaussian GLM as well, records could be explicitly weighted by $1/n_i$, which could alleviate this disadvantage.

Our results suggest that a Gaussian GLM might indeed provide a useful approximation on an ad hoc basis when the phenotypic distribution approaches a Gaussian distribution. However, real-world data are too complex to know in advance when exactly this will be the case. Therefore, when dealing with Binomial data, the performance of a Binomial GLM should always be evaluated before relying on a Gaussian approximation.

Modeling overdispersion

Our simulations clearly showed that accounting for overdispersion with finite mixture models is vital and improves prediction accuracy considerably, when there is strong overdispersion present in the data. Nonetheless, nonidentifiability still was an issue, especially for $N = 100$, where the finite mixture models performed significantly worse than the standard model under higher values of κ, i.e., when overdispersion was less pronounced. The better performance for $\kappa = 10$, however, showed that under strong overdispersion the greater flexibility of the finite mixture models overcompensates for problems due to nonidentifiability.

Comparing the results for $\kappa = 1{,}000$ with the results of the corresponding scenario in Table 3 (for which conceptually $\kappa = \infty$) shows that even low degrees of overdispersion can depress prediction accuracies. Thus, finite mixture models could still be of advantage under low degrees of overdispersion, but presumably only with very high N.

How much nonidentifiability depresses prediction accuracy will depend on the size of N compared with M. At higher N, we were able to fit more markers without severe nonidentifiability problems, as long as the increase in M was under proportional to the increase in N.

Owing to extensive long-range LD in bi-parental populations, a high marker density is not required for accurate genomic predictions. This is apparent from the only very marginal difference in prediction accuracy between the standard Binomial GLM fitted with the full and reduced marker sets. Reducing the number of markers for improving identifiability of finite mixture GLM therefore does not reduce the LD between markers and QTL to such an extent as to negatively affect the performance of the model. In most other types of populations encountered in animal or plant breeding, the marker density will be much more critical. How finite mixture Binomial GLM can be applied under such scenarios remains to be studied.

Wheat female fertility data set

The typical value of n_i observed in this data set corresponds to our $\lambda = 25$ scenario in the simulated data. Thus, in line

with our findings on the clear superiority of Binomial GLM under low λ, we found that Binomial GLM performed better than the Gaussian alternatives. For the full data set, the best model was the finite mixture Binomial BayesB GLM, indicating that there indeed was overdispersion present in this data set. The standard Binomial BayesB GLM was the best model for the "high fertility subset", however. Thus, either there was no overdispersion, or it could not be modeled due the nonidentifiability problems mentioned above.

Che and Xu (2012), who previously analyzed this data set, also found that using a Binomial GLM (with probit link, though) delivers considerably better QTL detection results than a Gaussian GLM. They also strongly argued in favor of using Binomial GLM for Binomial data, if only for the sake of statistical rigor, regardless of the quality of the approximation by Gaussian GLM. The authors also reported the presence of major QTL, explaining why BayesB tended to outperform GBLUP.

The low level of the predictive ability generally observed for the "high fertility subset" is most likely attributable to the low marker density, which left some of the chromosomes completely uncovered. The fertility rates of the full data set are concentrated at very high and very low values. Therefore, the model mostly captures the differentiation between these two groups. Thus, what is predicted is mainly whether a new observation has a very low or very high fertility. Doing this correctly is obviously easier than to predict the right order within any of these two groups (e.g., within the "high fertility subset") and may be done with lower marker coverage, which may explain the higher predictive ability observed for the full data set.

The presence of a number of observations far from the bulk of the data, with numbers of seeded spikelets very close to or exactly zero, might be an indication of zero-inflation, i.e., the presence of more zeros in the data than expected from the (overdispersed) Binomial model. Thus, incorporating zero-inflation into the models, as is often done for Poisson data (Meng 1997), might be worthwhile.

Maize haploid induction data set

Our results for the maize data set indicated that in bi-parental populations an average of about ten markers per chromosome is sufficient for obtaining decent levels of predictive ability. Results of our simulations showed that Binomial and Gaussian GLM converged for the highest value of $\lambda = 500$. Therefore, we did not expect the Binomial GLM to perform notably better than the Gaussian GLM for this data set, where the typical value of n_i was greater than 1,000. That the finite mixture Binomial BayesB GLM yielded a higher prediction accuracy than the standard Binomial GLM is again an indication of overdispersion in the data.

Prigge et al. (2012) detected several major QTL within this data set, again explaining why BayesB outperformed GBLUP significantly. To generate a sufficient number of DH lines and to exploit the entire genetic variance present in the source germplasm, a high HIR of the inducers is desired, especially because many of the haploid seedlings will not survive the subsequent chromosomal doubling process. Currently, the HIR of known inducers rarely exceeds 8 %. Therefore, breeding efforts are underway for improving HIR and based on our results, genomic prediction of HIR could be a valuable tool in this process, given the generally high predictive abilities observed.

In summary, we found that Binomial GLM, based either on marker effects or on total genetic values, can increase the accuracy of the predictions considerably as compared with Gaussian GLM. We further found that accounting for overdispersion can increase prediction accuracy and is vital under strong overdispersion.

Acknowledgements This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr Synbreed—Synergistic plant and animal breeding (FKZ: 0315528d).

References

Barret P, Brinkmann M, Beckert M (2008) A major locus expressed in the male gametophyte with incomplete penetrance is responsible for in situ gynogenesis in maize. Theor Appl Genet 117:581–94

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2012) Whole genome regression and prediction methods applied to plant and animal breeding. Genetics. doi:10.1534/genetics.112.143313

Che X, Xu S (2012) Generalized linear mixed models for mapping multiple quantitative trait loci. Heredity 109:41–49

Clark S, Hickey JM, van der Werf JH (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18

Dey D, Gelfand A, Peng F (1997) Overdispersed generalized linear models. J Stat Plan Infer 64:93–107

Dou B, Hou B, Xu H, Lou X, Chi X, Yang J, Wang F, Ni Z, Sun Q (2009) Efficient mapping of a female sterile gene in wheat (*Triticum aestivum* L.). Genetics res 91:337–43

Dou B, Hou B, Wang F, Yang J, Ni Z, Sun Q, Zhang YM (2010) Further mapping of quantitative trait loci for female sterility in wheat (*Triticum aestivum* L.). Genetics res 92:63–70

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer series in statistics. Springer, New York

Frühwirth-Schnatter S, Frühwirth R, Held L, Rue Hv (2009) Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Stat Comput 19:479–492

Fussl A, Frühwirth-Schnatter S, Frühwirth R (2012) Efficient mcmc for binomial logit models. ACM T Model Comput S (special issue on Monte Carlo methods in statistics forthcoming)

Gelfand AE, Sahu SK (1999) Identifiability, improper priors and gibbs sampling for generalized linear models. J Am Stat Assoc 94:247–253

Goggi A, Pollak L, Golden J (2007) Impact of early seed quality selection on maize inbreds and hybrids. Maydica 52:223–233

Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard M (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:e1001, 139

Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for bayesian model building in genomic selection. Genetics 191:969–987

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

Lashermes P, Beckert M, Crouelle DD (1988) Genetic control of maternal haploidy in maize (*Zea mays* L.) and selection of haploid inducing lines. Theor Appl Genet 76:405–410

Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294–305

Li L, Xu X, Jin W, Chen S (2009) Morphological and molecular evidences for DNA introgression in haploid induction via a high oil inducer CAUHOI in maize. Planta 230:367–376

Meng X (1997) The EM algorithm and medical studies: a historical linik. Stat Methods Med Res 6:3–23

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

Plummer M (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling

Plummer M, Best N, Cowles K, Vines K (2010) coda: output analysis and diagnostics for MCMC. http://CRAN.R-project.org/package=coda,rpackageversion0.14-2

Prigge V, Xu X, Li L, Babu R, Chen S, Atlin GN, Melchinger AE (2012) New insights into the genetics of in vivo induction of maternal haploids, the backbone of doubled haploid technology in maize. Genetics 190:781–793

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. ISBN: 3-900051-07-0

Sellamuthu R, Liu GF, Ranganathan CB, Serraj R (2011) Genetic analysis and validation of quantitative trait loci associated with reproductive-growth traits and grain yield under drought stress in a doubled haploid line population of rice (*Oryza sativa* L.). Field Crops Res 124:46–58

Technow F (2011) hypred: simulation of genomic data in applied genetics. R package version 0.1

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

Thomas A, OHara R, U L, Sturtz S (2006) Making bugs open. R News 6:12–17

VanRaden PM (2008) Efficient methods to compute genomic predictions. J dairy Sci 91:4414–4423

Villumsen TM, Janss L, Lund MS (2009) The importance of haplotype length and heritability using genomic selection in dairy cattle. J Anim Breed Genetics 126:3–13

Wray NR, Goddard ME, Visscher PM (2008) Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 18:257–263

Yang W, Tempelman RJ (2012) A Bayesian antedependence model for whole genome prediction. Genetics 190:1491–1501

Yousefabadi V, Rajabi A (2012) Study on inheritance of seed technological characteristics in sugar beet. Euphytica 186:367–376

Zhao Z, Wang C, Jiang L, Zhu S, Ikehashi H, Wan J (2006) Identification of a new hybrid sterility gene in rice (bi *Oryza sativa* L.). Euphytica 151:331–337

Zhong S, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics 182:355–364