Theor Appl Genet (2013) 126:69–82
DOI 10.1007/s00122-012-1960-1

ORIGINAL PAPER

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

Torben Schulz-Streeck · Joseph O. Ogutu · Hans-Peter Piepho

Received: 11 November 2011 / Accepted: 31 July 2012 / Published online: 19 August 2012

Abstract Genomic selection (GS) is a method for predicting breeding values of plants or animals using many molecular markers that is commonly implemented in two stages. In plant breeding the first stage usually involves computation of adjusted means for genotypes which are then used to predict genomic breeding values in the second stage. We compared two classical stage-wise approaches, which either ignore or approximate correlations among the means by a diagonal matrix, and a new method, to a single-stage analysis for GS using ridge regression best linear unbiased prediction (RR-BLUP). The new stage-wise method rotates (orthogonalizes) the adjusted means from the first stage before submitting them to the second stage. This makes the errors approximately independently and identically normally distributed, which is a prerequisite for many procedures that are potentially useful for GS such as machine learning methods (e.g. boosting) and regularized regression methods (e.g. lasso). This is illustrated in this paper using componentwise boosting. The componentwise boosting method minimizes squared error loss using least squares and iteratively and automatically selects markers that are most predictive of genomic breeding values. Results are compared with those of RR-BLUP using five-fold cross-validation. The new stage-wise approach with rotated means was slightly more similar to the single-stage analysis than the classical two-stage approaches based on non-rotated means for two unbalanced datasets. This suggests that rotation is a worthwhile pre-processing step in GS for the two-stage approaches for unbalanced datasets. Moreover, the predictive accuracy of stage-wise RR-BLUP was higher (5.0–6.1 %) than that of componentwise boosting.

Abbreviations

BLUP	Best linear unbiased prediction
GEBV	Genomic estimated breeding value
GS	Genomic selection
RCBD	Randomized complete block design
REML	Restricted maximum likelihood
RR-BLUP	Ridge regression BLUP
SNP	Single nucleotide polymorphism

Introduction

Genomic selection (GS) is a method for predicting genomic breeding values (GEBV) for plants or animals using dense genetic markers, such as single-nucleotide polymorphisms (SNPs; Meuwissen et al. 2001). A number of approaches have been used for GS including mixed models (Meuwissen et al. 2001; Piepho 2009), machine learning (Long et al. 2007; Ogutu et al. 2011) and Bayesian methods (Meuwissen et al. 2001) and models accounting for polygenic effects (e.g. Calus and Veerkamp 2007; Hayes et al. 2009; Piepho 2009; Schulz-Streeck and Piepho 2010). All these approaches normally undertake GS in two stages. The first stage involves computing adjusted means for genotypes which are then used in the second stage to predict GEBVs based on markers. This is especially important in

Communicated by J. Yu.

T. Schulz-Streeck · J. O. Ogutu · H.-P. Piepho (✉)
Bioinformatics Unit, Institute of Crop Science,
University of Hohenheim, Fruwirthstrasse 23,
70599 Stuttgart, Germany
e-mail: piepho@uni-hohenheim.de

T. Schulz-Streeck
e-mail: torben.schulz-streeck@uni-hohenheim.de

plant breeding where the genotypes are routinely tested in different environments so that it is convenient to first compute adjusted genotype means across the environments of a target region and then use these means in GS. The adjusted means can have variance–covariance structures of varying complexity, depending on the details of the field trial and the pattern of genotype–environment interaction. In plant breeding experiments incomplete blocking and spatial methods of analysis are often used and not all genotypes are typically tested in all environments, leading to heterogeneous variance–covariance between adjusted means. However, all stage-wise models for GS typically assume that means have independent errors. This assumption constitutes an approximation that ignores correlation among adjusted means. For these reasons, single-stage analysis is often regarded as the gold standard because it can fully account for the entire variance–covariance structure of the observed data (Smith et al. 2001b).

Most plant breeding programs involve testing hundreds of genotypes, thus requiring incomplete block or row-column designs (John and Williams 1995), in which not every genotype is tested in each block. The extent of design-induced unbalancedness often increases in series of trials conducted over different environments when not all genotypes can be grown in every environment due to limited availability of seeds or other resources. Compared with balanced designs, unbalanced designs complicate the use of stage-wise approaches due to the increased likelihood of heterogeneity and hence complexity in the variance–covariance between adjusted means. This makes it much harder to account for the error variance–covariance structure and hence to minimize information loss when adjusted means are passed on from the first to the second stage of a stage-wise analysis and thereby widens the difference between the single- and two-stage approaches (Piepho et al. 2011, 2012a). Several weighting schemes have been devised to enhance the performance of the stage-wise relative to the single-stage approach, e.g. weighting the adjusted means by the inverse of their squared standard errors (Cullis et al. 1996) or weighting the adjusted means by the diagonal elements of the inverse of their variance–covariance matrix (Smith et al. 2001a). Modelling the environment and block effects as fixed normally results in relatively small covariances between the adjusted means so that weighting the means by the inverse of their squared standard errors often constitutes a reasonable approximation (Möhring and Piepho 2009). To recover inter-block and inter-environment information, by exploiting information from across blocks and environments, both the block and the environment effects must be modelled as random effects. This increases the covariances between the adjusted means. Consequently, weighting the adjusted means by the reciprocal of the inverse of their variance–covariance matrix may be more efficient than using their squared standard errors (Smith et al. 2001a). Evidence from recent theoretical and simulation studies comparing the approximate two-stage method of Smith et al. (2001a) to single-stage analysis (Welham et al. 2010) suggests that ignoring covariance information can incur substantial loss of information. Although existing weighting schemes have thus far focused on weighting the adjusted means per environment before submission to the analysis across environments, GS, QTL mapping, or association mapping studies normally use adjusted means computed across environments from a target region. This increases the extent of heterogeneity of variances and covariances between the adjusted means and thus the importance of weighting to minimize information loss.

If feasible, a single-stage analysis is therefore preferable to a classical two-stage analysis for GS (Cullis et al. 1998). But a two-stage analysis is well suited to GS due to its simplicity and computational efficiency (Möhring and Piepho 2009). For GS this can be an important advantage because the computing time for the single-stage analysis can be considerable and dependent on the complexity of the dataset, the method used and the number of markers and genotypes involved. Moreover, it is common to use mixed models for phenotypic analysis and exploit the mixed model framework to represent the experimental design for the phenotypic data. Whereas this has its own advantages, some methods suitable for GS do not include mixed modelling options. Since a single-stage analysis is not always feasible or is computationally too burdensome for complex datasets, especially on some computing platforms such as the currently available versions of the SAS software, because of a lack-of-memory problem, the two-stage analysis is often the only feasible option. Piepho et al. (2011, 2012a) presented a new stage-wise method, in which all the information in the variance–covariance structure of the adjusted means in the first stage is passed on to the second stage, making the method fully efficient relative to a single-stage approach. Piepho et al. (2011, 2012a) showed this stage-wise method to be efficient in tests involving analyses of a series of field trials and conjectured that it may be similarly efficient for association mapping and GS. We extend the new stage-wise approach to genomic selection.

Many methods and models suitable for GS, such as componentwise linear least squares boosting (Bühlmann and Hothorn 2007), assume independently and identically normally distributed (i.i.d.) errors. The new stage-wise approach makes these procedures available for GS, by rotating (orthogonalizing) the adjusted means and ensuring that the assumption of i.i.d. errors is satisfied. Boosting is an ensemble machine learning procedure that combines the performance of many "weak" learners each of which performs only slightly better than random guessing to yield a powerful learning algorithm, “weighted committee” or “ensemble” (Friedman and Hastie 2000; Bühlmann and Hothorn 2007; Hastie et al. 2009). Boosting was first introduced in the machine learning literature by Freund and Schapire (1997). Boosting fits stage-wise additive models using basis functions (Friedman and Hastie 2000). The individual basis functions, each of which is a function of the predictors, are called “weak” learners or base learners. The base learners are weak in the sense that they usually have low complexity, high bias (in most cases) and low variance (Buja et al. 2007). The base learners are many and varied and include classification and regression trees, least squares, exponential family models, survival models, splines, etc.

There are many boosting algorithms for classification and regression, including adaptive, stochastic gradient and componentwise least squares boosting. However, all boosting algorithms exploit two basic ideas, namely very flexible fitting functions that capture local characteristics of the data and then averaging over many iterations (Berk 2008). As a result, boosting can also be usefully viewed, partly, as a regularization (penalty-based) procedure, similar to such shrinkage procedures as the lasso or ridge regression, which enforce less shrinkage with each successive iteration (Berk 2008). At each iteration in boosting misclassified or poorly fitted observations are given more relative weight which forces the boosting algorithm to concentrate more on the incorrectly classified or fitted observations in the next iteration. The final fitted values from boosting are weighted combinations of the many values from previous fitting attempts; hence the term “weighted committee” or “ensemble”. Mathematical and technical details on boosting algorithms can be found elsewhere (Friedman and Hastie 2000; Bühlmann and Hothorn 2007; Hastie et al. 2009). Bühlmann and Hothorn (2007) also present mathematical and algorithmic details of componentwise linear least squares boosting used in this study.

Our main aim in this paper is to compare different stage-wise approaches for GS. We expect a robust performance of the new stage-wise method because it passes all the information contained in the variance–covariance matrix of the adjusted means from the first to the second stage and ensures that the errors are approximately i.i.d., in contrast to other two stage-wise methods, which either ignore or approximate correlations among the means by a diagonal matrix. Since the single-stage analysis is more accurate based on theoretical and simulation evidence (Welham et al. 2010), we used it as the gold standard for assessing the performance of the stage-wise methods. We used ridge regression best linear unbiased prediction (RR-BLUP) (Piepho 2009) for this comparison because it is commonly used and can be integrated into a mixed model framework. In fact, for some other methods used in GS it would be difficult, if not impossible, to devise a reasonable single-stage analysis, because this requires a mixed model framework to represent the experimental design for the phenotypic data. Additionally, we show that the stage-wise method of Piepho et al. (2011, 2012a) is not limited to RR-BLUP alone but can be implemented using various methods, such as boosting, and we use fivefold cross-validation to compare the performance of the stage-wise RR-BLUP and componentwise boosting with linear least squares as base learners or basis functions (Bühlmann and Hothorn 2007) in GS.

Materials and methods

Dataset

We used two separate datasets (called A and B) each containing 177 un-replicated double haploid maize (*Zea mays* L.) lines (testcross genotypes) each derived from a different biparental cross. However, one parent is common to the two populations.We also combined both populations to enable a combined analysis. The hybrid performance for dry grain yield (tons/ha) for both datasets was tested with the same common tester. The testcross genotypes were tested in six locations in the same target region in 1 year, but not every testcross genotype was tested in each location. An augmented trial design with incomplete blocks was used in each location to test the testcross genotypes. In each location three to five incomplete blocks, each containing a single column of plots, were used. The two standard varieties but not the testcross genotypes were replicated (i.e. planted in all blocks) in each location. As a result, the standard varieties enable estimation of the inter-block variance and separation of the block from the error variance and are therefore said to connect the different blocks. The standard varieties are intended solely to facilitate the analysis of the testcross genotypes and are themselves not used in predicting GEBVs.

Genotyping of all the genotypes was done by 768 SNP markers spaced equally throughout the genome and the information stored in a matrix $M = \{m_{ik}\}$. The marker covariate m_{ik} for the i-th genotype ($i = 1, 2,\ldots, G$) and the k-th marker ($k = 1, 2,\ldots, M$) for biallelic SNP markers with alleles A_1 and A_2 was set to 1 for A_1A_1, -1 for A_2A_2 and to 0 for A_1A_2, A_2A_1 and missing values. Heterozygous markers were treated the same way as missing information, because double haploids are completely homozygous. Markers with more than 20 % missing values, or more than 5 % heterozygous genotypes, or with minor allele frequency less than 2.5 % were excluded, resulting in 275 markers for dataset A, 201 for dataset B and 298 for the combined dataset (A + B).

Ridge regression BLUP

Genomic selection was done using RR-BLUP, where the genotypic value for the i-th genotype (u_{ai}) was predicted by the following regression on the makers:

$$u_{ai} = \sum_{k=1}^{M} v_k m_{ik}, \tag{1}$$

where m_{ik} is the regressor variable for the i-th genotype and the k-th marker, while v_k are the regression coefficients. The regression coefficients are assumed to be random sample from a common normal distribution, $v_k \sim N(0, \sigma_a^2)$. The linear model in (1) can be rewritten in matrix form as $u_a = Mv$, where $u_a^{\mathrm{T}} = (u_{a1}, u_{a2}, \ldots, u_{aG})$ and $v^{\mathrm{T}} = (v_1, v_2, \ldots, v_M)$. For single (mean-centred) observation y_i per genotype with independent residual errors e_i having zero mean and variance σ_e^2, the model for the observed data is $y = u_a + e$, where $y^{\mathrm{T}} = (y_1, y_2, \ldots, y_G)$ and $e^{\mathrm{T}} = (e_1, e_2, \ldots, e_G)$. Many authors (e.g. Ruppert et al. 2003; Piepho 2009) have shown that for RR-BLUP the genotypic variance given the covariates (markers) is $\mathrm{var}(g|M) = \sigma_a^2 MM^T$, where σ_a^2 is estimated by REML. When the observations are mean-centred, GEBVs are the predicted genotypic values. Otherwise the model is extended by an intercept and GEBVs then are estimated by the sum of intercept plus the predicted genotypic values.

Overview of single-stage and two-stage approaches

We used different approaches to estimate GEBVs. First, we used a single-stage analysis where the prediction of GEBVs and the computation of adjusted genotypic means across locations were done in one stage. Additionally, we used three different two-stage approaches involving the prediction of adjusted means for the testcross genotypes across the different locations in the first stage and then using these means for GS in the second stage. The adjusted means were correlated because of the details of the field trial design. In the first two-stage approach we simply ignored this correlation. In the second approach, we used a weighting scheme based on the diagonal matrix extracted from the inverse of the variance–covariance matrix of the adjusted means to approximate the error structure of the adjusted means (Smith et al. 2001a). In the third, we rotated the adjusted means so that the variance–covariance matrix of the adjusted means was orthogonal (Piepho et al. 2011, 2012a).

Single-stage approach

The following linear mixed model was used for the single-stage analysis,

$$y = 1_N \alpha + Z_a u_a + Z_b u_b + Z_c u_c + Z_d u_d + e_e, \tag{2}$$

where y is the observed data vector of yield per plot, 1_N is an N-dimensional vector with all elements equal to 1 and N is the number of observation; α is the common intercept; Z_a, Z_b, Z_c and Z_d are design matrices for the random effects; u_a is a vector of random genotypic main effects, with $\mathrm{var}(u_a) = \sigma_a^2 MM^{\mathrm{T}}$, M representing the matrix with the marker information and M^{T} its transpose (Piepho 2009). Thus, $\mathrm{var}(u_a)$ is the marker-based variance–covariance matrix of the genetic main effects. u_b is a vector of random location effects with $\mathrm{var}(u_b) = G_b = I\sigma_b^2$, u_c is a vector of random genotype-location effects with $\mathrm{var}(u_c) = G_c = I\sigma_c^2$, u_d is a vector of random within-location incomplete block effects, with $\mathrm{var}(u_d) = G_d = I\sigma_d^2$, and e_e is a vector of plot errors with $\mathrm{var}(e_e) = R_e = I\sigma_e^2$.

We did not model heterogeneous variances between the different locations due to lack of replicates of the testcross genotypes within locations. For simplicity of presentation, we ignore the fact that standard varieties were used in model (2). The standard varieties were used simply to aid the analysis of the field trials and did not contribute to (i.e. were blocked out from) the prediction of the genotypic main effect, or the population mean. Therefore, we fitted a fixed effect with a different level for each standard variety and one level for all the testcross genotypes (Piepho et al. 2006). Implicit in this model are the assumptions that the standard varieties have different means, all testcross genotypes belonging to the same population have the same population mean and that the standard varieties are independent from the testcross genotypes and from the other standard varieties. For the combined analysis of both populations, a fixed population effect was fitted. Note that the testcross genotypes are modelled as random effects. The standard varieties were blocked out when estimating genotypic main effects through the use of a dummy variable equal to zero for all the standard varieties and one for all the testcross genotypes. Since the dummy variable was defined as a quantitative variable, the genetic variance conditional on the markers was zero for the standard varieties but was $\mathrm{var}(u_a) = \sigma_a^2 MM^{\mathrm{T}}$ for the testcross genotypes.

The GEBVs for the testcross genotypes were predicted by

$$\mathrm{GEBV} = 1_C \hat{\alpha} + \hat{u}_a \tag{3}$$

where 1_C is a C-dimensional vector of ones and C is the number of testcross genotypes, $\hat{\alpha}$ is the predicted mean for all testcross genotypes, excluding the standard varieties, and $\hat{u}_a$ is the vector of the predicted genotypic main effects.

Two-stage approaches

The analysis can also be done in two stages. To this end, model (2) can be re-cast as

$$y = 1_N\alpha + Z_a u_a + f \tag{4}$$

where $f = Z_b u_b + Z_c u_c + Z_d u_d + e_e$ and $\mathrm{var}(f) = \Sigma_f = Z_b G_b Z_b^{\mathrm{T}} + Z_c G_c Z_c^{\mathrm{T}} + Z_d G_d Z_d^{\mathrm{T}} + R_e$.

Model (4) is equivalent to model (2) but all random effects except the genetic part are combined in one random effect called f. Model (4) can be represented in two stages (Piepho et al. 2011, 2012a). The first stage is given by

$$y = X_1\mu_1 + f \tag{5}$$

where μ_1 are the genotype means across locations and X_1 is an associated design matrix, which in this case is equal to Z_a. At the second stage the adjusted means (μ_1) in (5) can be calculated using the equation

$$\mu_1 = 1_C\alpha + u_a. \tag{6}$$

This means that upon replacing μ_1 in (5) with its counterpart in (6), model (5) becomes equivalent to the single-stage model (2).

The first stage of the two-stage approaches

At the first stage adjusted means for the testcross genotypes are computed across the different locations ($\hat{\mu}_1$) using model (5) and submitted to the second stage. Note that the adjusted means of the standard varieties are excluded from the dataset before submission to the second stage.

The second stage of the two-stage approaches

At the second stage, the adjusted means ($\hat{\mu}_1$) from the first stage are used to predict GEBVs. For RR-BLUP the GEBVs were estimated in the second stage using the linear mixed model

$$\hat{\mu}_1 = 1_C\alpha + u_a + e_a, \tag{7}$$

where $e_a = \left(X_1^{\mathrm{T}}\Sigma_f^{-1}X_1\right)^{-1} X_1^{\mathrm{T}}\Sigma_f^{-1}f$ and $\mathrm{var}(e_a) = \left(X_1^{\mathrm{T}}\Sigma_f^{-1}X_1\right)^{-1}$. Note that the expression for the error term e_a results from generalized least squares estimation in stage one (Piepho et al. 2011, 2012a).

In the second stage, we either used “rotated means”, which will be explained below, or we simply used “unrotated means”. For the latter approach, we followed two options: The variance–covariance matrix of residual errors was either approximated as $\mathrm{var}(e_a) = I\sigma_{ea}^2$ (unweighted analysis) or as $\mathrm{var}(e_a) = D_{ea}$ (weighted analysis), where D_{ea} is a diagonal matrix, whose elements are equal to those of the inverse of the variance–covariance matrix of the adjusted means $\left(X_1^{\mathrm{T}}\Sigma_f^{-1}X_1\right)$ (Smith et al. 2001a).

Rotation of the adjusted means $\hat{\mu} \sim N(\mu, W_1)$ was done using the spectral decomposition $W_1^{-1} = \left(W_1^{-1/2}\right)^2$, where $W_1^{-1/2}$ is a square symmetric matrix of full rank (Rao et al. 2008). More precisely, we used the spectral decomposition of $\mathrm{var}(e_a) = \left(X_1^{\mathrm{T}}\Sigma_f^{-1}X_1\right)^{-1} = S\Lambda S^{\mathrm{T}}$, where S is a matrix of eigenvectors and Λ is the corresponding matrix of eigenvalues. From this decomposition, we compute $W_1 = S\Lambda^{-1/2}S^{\mathrm{T}}$, which can then be used to compute rotated (orthogonalized) means as $\tilde{\hat{\mu}}_1 = W_1^{-1/2}\hat{\mu}_1 \sim N\left(W_1^{-1/2}\mu, I\right)$ (Piepho et al. 2011, 2012a). For these rotated means the following mixed model holds (Piepho et al. 2011, 2012a):

$$\tilde{\hat{\mu}}_1 = \tilde{1}_C\alpha + \tilde{u}_a + \tilde{e}_a, \tag{8}$$

where $\tilde{\hat{\mu}}_1 = W_1^{-1/2}\hat{\mu}_1$, $\hat{\mu}_1 = \left(X_1^{\mathrm{T}}\Sigma_w^{-1}X_1\right)^{-1}X_1^{\mathrm{T}}\Sigma_w^{-1}y$, $\tilde{1}_C = W_1^{-1/2}1_C$, $\tilde{u}_a = \tilde{M}v = W_1^{-1/2}Mv$ with $\mathrm{var}(\tilde{u}_a) = \sigma_a^2\tilde{M}\tilde{M}^{\mathrm{T}}$ and $\tilde{e}_a = W_1^{-1/2}e_a$ with $\mathrm{var}(\tilde{e}_a) = I$,

Hence, the rotated means have i.i.d. standard normal errors. It is important to note that the rotation in (8) affects only the elements of the design matrix of markers, but not the estimated effects of the individual markers. This facilitates fitting the effects using the model for RR-BLUP. It can be shown (Piepho et al. 2011, 2012a) that the mixed model (8) is equivalent to the mixed model (2) for the single-stage analysis for known variance components. The rotation approach is available in the R-Package rrBlupMethod6 (Piepho et al. 2012b; Schulz-Streeck 2012).

GEBVs for the nonrotated means were estimated as

$$\mathrm{GEBV} = 1_C\hat{\alpha} + \hat{u}_a \tag{9}$$

and for the rotated means as

$$\mathrm{GEBV} = \tilde{1}_C\hat{\alpha} + \hat{\tilde{u}}_a \tag{10}$$

where the terms are defined analogously as above.

Predicting non-phenotyped genotypes using the rotation method

A major concern of genomic selection is to predict the performance of the non-phenotyped testcross genotypes. This is possible for all the methods we used but is shown here only for the two-stage rotation method as an illustrative example. Genotypic main effects ($\hat{\tilde{u}}_a$) of the phenotyped testcross genotypes were first predicted using the rotation method (8) and the marker information and the resulting relationship used to predict the genotypic main effects of the non-phenotyped testcross genotypes. Estimates of the marker effects ($\hat{v}$) were then obtained using the following estimator (Henderson 1977):

$$\hat{v} = \tilde{M}^{\mathrm{T}}\left(\tilde{M}\tilde{M}^{\mathrm{T}}\right)^{-1}\hat{\tilde{u}}_a \tag{11}$$

where $\hat{\tilde{u}}_a$ is the vector of predicted genotypic main effects of phenotyped testcross genotypes and $\tilde{M}$ is the corresponding matrix containing the rotated marker information. The genotypic main effects of the non-phenotyped testcross genotypes were predicted using the estimated marker effects as

$$\hat{u}_{a\mathrm{NP}} = M_{\mathrm{NP}}\hat{v} \tag{12}$$

where $\hat{u}_{\mathrm{aNP}}$ is the predicted genotypic main effects of the non-phenotyped, and M_{NP} is the corresponding design matrices containing the marker information. Since the design matrix of markers for the non-phenotyped testcross genotypes (M_{NP}) in (12) is not rotated, predictions for the non-phenotyped testcross genotypes ($\hat{u}_{a\mathrm{NP}}$) are made on the original scale. Furthermore, assuming the matrix $\tilde{M}\tilde{M}^{\mathrm{T}}$ is invertible, GEBVs can be calculated as

$$\mathrm{GEBV}_{\mathrm{NP}} = 1_H\hat{\alpha} + \hat{u}_{a\mathrm{NP}}, \tag{13}$$

where 1_H is a H-dimensional column vector of ones, H is the number of non-phenotyped testcross genotypes and $\hat{\alpha}$ is a vector of the predicted mean for all phenotyped testcross genotypes.

However, if $\tilde{M}\tilde{M}^{\mathrm{T}}$ is not invertible the following method can be used to estimate genotypic values for the non-phenotyped testcross genotypes. First, the marker effects are estimated by (Searle et al. 1992):

$$\hat{v} = \hat{\sigma}_a^2\tilde{M}^{\mathrm{T}}\hat{V}^{-1}\left(\tilde{\hat{\mu}}_1 - \tilde{1}_C\hat{\alpha}\right), \tag{14}$$

where $\hat{v}$ is the vector of BLUPs of the marker effects, $\hat{\sigma}_a^2$ is the variance of the marker effects, $\hat{V}$ is the estimated variance–covariance matrix of the response variable of the rotated adjusted means ($\tilde{\hat{\mu}}_1$), and $\tilde{1}_C$ is a C-dimensional column vector obtained by rotating the vector with all elements equal to one.

Note that we can always predict random effects of the markers (v) because $\hat{V}$ is generally positive-definite even when $\tilde{M}\tilde{M}^{\mathrm{T}}$ is not. The genotypic value of non-phenotyped testcross genotypes can then be estimated by

$$\mathrm{GEBV}_{\mathrm{NP}} = 1_H\hat{\alpha} + M_{\mathrm{NP}}\hat{v} \tag{15}$$

where the terms are defined similarly as for the preceding models (11, 12 and 13).

Boosting

After rotating the adjusted means to obtain approximately i.i.d. errors, GS is no longer limited to methods able to account for correlated error structures such as RR-BLUP alone but can be implemented using various other methods such as boosting. We illustrate this here for componentwise boosting.

We used the same regression model for componentwise linear least squares boosting as for RR-BLUP, but L_2-boosting for linear models is more akin to ordinary least squares and thus to fixed effects modelling for marker effects v (Bühlmann and Hothorn 2007). Our stage-two model for boosting is

$$\hat{\mu}_1 = 1_C\alpha + Mv + e_a \tag{16}$$

Note that in fitting (16) and other boosted models in this paper only the best supported predictor is selected and fitted at each iteration of the boosting algorithm thus enabling automatic variable selection because some predictors will have estimated coefficients for the markers exactly equal to zero in the final model (Bühlmann and Hothorn 2007; Boulesteix and Hothorn 2010). Moreover, from regression theory the least squares loss function is best when errors are i.i.d.

We used componentwise linear least squares as a base procedure. Boosting componentwise linear least squares within a generalized linear model framework enabled automatic selection of predictor variables with the greatest influence on the response variable (Bühlmann and Hothorn 2007). Since boosting iteratively adds basis functions (base learners) in a greedy fashion such that each additional base learner further reduces the selected loss (error) function (Hastie et al. 2009), prediction is achieved using regression coefficients for predictors retained in the model at the end of the boosting iterations. We used a gradient boosting algorithm, assuming the Gaussian distribution for minimizing squared-error loss in the R package *mboost* (Hofner 2010). We determined the main tuning parameter, the optimal number of iterations, using cross-validation and the step-size length using a grid search.

For boosting independent genotypic means ($\hat{\mu}_1$), e.g. adjusted means form randomized complete block designs, the adjusted means are mean-centred (Bühlmann and Hothorn 2007), corresponding to the model

$$(\hat{\mu}_1 - \bar{\hat{\mu}}_1) = (M - \bar{M})v + (e_a - \bar{e}_a). \tag{17}$$

where $\hat{\mu}_1$ is the vector of the adjusted means with a mean vector $\bar{\hat{\mu}}_1$; M is the design matrix of marker covariates, with a matrix of means for each marker $\bar{M}$ and v is a vector of the regression coefficients of markers.

As a result of the centring, the intercept drops out of the model. For simplicity (17) can be rewritten as

$$\hat{\mu}_{1(\mathrm{centre})} = M_{(\mathrm{centre})}v + e_{a(\mathrm{centre})} \tag{18}$$

where the subscript centre indicates that the vector or matrix has been centred as explained above.

We boosted the following rotated model, where the rotation was done as for the RR-BLUP model:

$$\tilde{\hat{\mu}}_1 = \tilde{1}_C\alpha + \tilde{M}v + \tilde{e}_a \tag{19}$$

where $\tilde{1}_C$ is the C-dimensional column vector obtained by rotating the vector with all elements equal to one. Note that the elements of $\tilde{1}_C$ do not usually equal one. To boost this model using componentwise linear least squares, we combined the standard linear model with the boosting algorithm as proposed by Boulesteix and Hothorn (2010) as follows. We first regressed the rotated observed values ($\tilde{\hat{\mu}}_1$) against the rotated intercept vector ($\tilde{1}_C$) and the usual intercept (γ) to obtain predicted values for each instance of $\tilde{\hat{\mu}}_1$ as well as estimates of the regression coefficients ($\hat{\gamma}, \hat{\alpha}$) using the regression machinery for standard generalized linear models. We then supplied the vector of predicted values for $(1_C\hat{\gamma} + \tilde{1}_C\hat{\alpha})$ as an offset in the boosting algorithm. Each predictor variable in the design matrix $\tilde{M}$ in (19) was mean-centred (i.e. $\tilde{M}_{(\text{centre})}$) in the boosting algorithm (Bühlmann and Hothorn 2007). Hence the full boosted model was

$$\tilde{\hat{\mu}}_1 = (1_C\hat{\gamma} + \tilde{1}_C\hat{\alpha}) + \tilde{M}_{(\text{centre})}v + \tilde{e}_{a(\text{centre})} \tag{20}$$

where $1_C\hat{\gamma} + \tilde{1}_C\hat{\alpha}$ is the offset and the subscript centre indicates that the vector or matrix has been mean centred. Note that the rotated model in (19) does not require an intercept as such, so the added regression on a common intercept γ in (20) corresponds to mean-centreing, which is generally recommended but is also hard to avoid in contemporary implementations of component linear least squares boosting, such as the algorithm used in the *mboost* package (Hofner 2010), that require an intercept as an integral part of the boosted model and do not permit blocking out the intercept during model fitting as an option. Boosting was used to obtain estimates of v (i.e. $\hat{v}$) only, since the offset was held fixed.

Comparison of single-stage and two-stage approaches

We compared the single- and two-stage approaches using RR-BLUP only. The single-stage approach served as the gold standard for evaluating the performance of the two-stage methods as follows, thus obviating the need to first split testcross genotypes into training and validation sets. The Pearson and Spearman rank correlation coefficients between the GEBVs predicted by the single- and two-stage methods were used for the evaluation. GEBVs predicted using rotated means were back-transformed through pre-multiplication by $W_1^{1/2}$, so that all predictions were compared on the same scale. As a further evaluation criterion, we calculated the absolute deviation between the values predicted from the single- and two-stage analyses. The minimum, lower and upper quartiles, median, and maximum values of the absolute deviations of predictions of the single-stage approach from those of the non-rotated, weighted and rotated two-stage approaches were compared using box plots. Furthermore, we compared the two-stage approaches to the single-stage approach to assess the extent to which they selected the same n ($n = 1,\ldots,100$) best testcross genotypes as the single-stage approach did.

Comparison of RR-BLUP and boosting using cross-validation

We compared the predictive accuracies of RR-BLUP and boosting using cross-validation (CV) only for the stage-wise analysis of the rotated means, which was identified as the stage-wise method of choice based on comparing the single- and two-stage approaches. A fivefold CV involving randomly splitting the rotated adjusted means into five subsamples, one of which was held out as a validation set (Val) at a time, was undertaken to evaluate predictive accuracy. The remaining four subsamples were combined into one training set (T). This was repeated five times, thus allowing each subsample to be the validation set once. This entire process was repeated ten times, yielding a total of 50 replicate GS predictions. For boosting, tuning parameters were optimized using the training set only.

An assumption integral to the proper conduct of the k-fold CV, where k is the number of random subsamples (e.g. 5 in our case), is that the errors are i.i.d. and hence that the training and validation sets are independent (Arlot and Celisse 2010). To satisfy this assumption, the adjusted means of all the testcross genotypes were rotated prior to splitting the dataset into training and validation sets.

For RR-BLUP the model (8) was specialized for the training set only as

$$\tilde{\hat{\mu}}_\text{T} = \tilde{1}_\text{T}\alpha + \tilde{u}_{a\text{T}} + \tilde{e}_{a\text{T}}, \tag{21}$$

where the subscript T indicates that only the training set is being used. The corresponding predictions for the validation set were obtained from

$$\tilde{\hat{u}}_{a\text{Val}} = \tilde{M}_\text{Val}\tilde{M}_\text{T}^\text{T}\left(\tilde{M}_\text{T}\tilde{M}_\text{T}^\text{T}\right)^{-1}\tilde{\hat{u}}_{a\text{T}}, \tag{22}$$

where the subscript Val indicates that only the validation set is being used.

GEBVs were then calculated by

$$\text{GEBV}_\text{Val} = \tilde{1}_\text{Val}\alpha + \tilde{\hat{u}}_{a\text{Val}}. \tag{23}$$

For boosting we used model (20), specialized for the training set only as

$$\tilde{\hat{\mu}}_{1\text{T}} = (1_\text{T}\hat{\gamma} + \tilde{1}_\text{T}\hat{\alpha}) + \tilde{M}_{\text{T}(\text{centre})}v + \tilde{e}_{a\text{T}} \tag{24}$$

The predictions of $\tilde{\hat{\mu}}$ were then obtained from

$$\text{GEBV}_\text{Val} = (1_\text{Val}\hat{\gamma} + \tilde{1}_\text{Val}\hat{\alpha} - \bar{\tilde{M}}_T\hat{v}) + \tilde{M}_\text{Val}\hat{v} \tag{25}$$

where $\bar{\tilde{M}}_\text{T}$ is a matrix of means for each marker for the training set. The term $\bar{\tilde{M}}_\text{T}\hat{v}$ is subtracted from the intercept because in (24) the design matrix for the effect v is mean centred. Thus the design matrix $\tilde{M}_{\text{T}(\text{centre})}$ in (24) can be rewritten as $\tilde{M}_{\mathrm{T(centre)}} = \tilde{M}_{\mathrm{T}} - \bar{\tilde{M}}_{\mathrm{T}}$. In (25) the non-centred design matrix of the marker for the validation set is used thus $\bar{\tilde{M}}_{\mathrm{T}}\hat{v}$ must be subtracted.

Measures of accuracy via cross-validation

The means of the Pearson correlation coefficients ($n = 50$ replicates) between the rotated adjusted means from the first stage ($\tilde{\mu}_{1\mathrm{Val}}$) and their corresponding predicted values (GEBV$_V$) using models (23) and (25) from the validation sets were used as measures of prediction accuracy. In each validation set 35 or 36 testcross genotypes were used for dataset A and B, but 70 or 71 for the combined dataset. A *t* test was used for head-to-head comparison of RR-BLUP and boosting based on the 50 replicate Pearson correlations derived from the 50 cross-validation runs.

Results

Comparison of single-stage and two-stage approaches

The variances and covariances of the adjusted means were heterogeneous within each dataset (Fig. 1), contrary to expectation for balanced data, thus indicating a departure from the common assumption of i.i.d. errors for the adjusted means.

Fits of the non-rotated and the rotated two-stage analyses were quite similar to each other and to the single-stage analyses for the combined and each of the two datasets (Table 1; Figs. 2, 3). This means that violation of the i.i.d. assumption had relatively small adverse impact on performance of the classical stage-wise approach. However, the predictions of the rotated two-stage analysis, for which

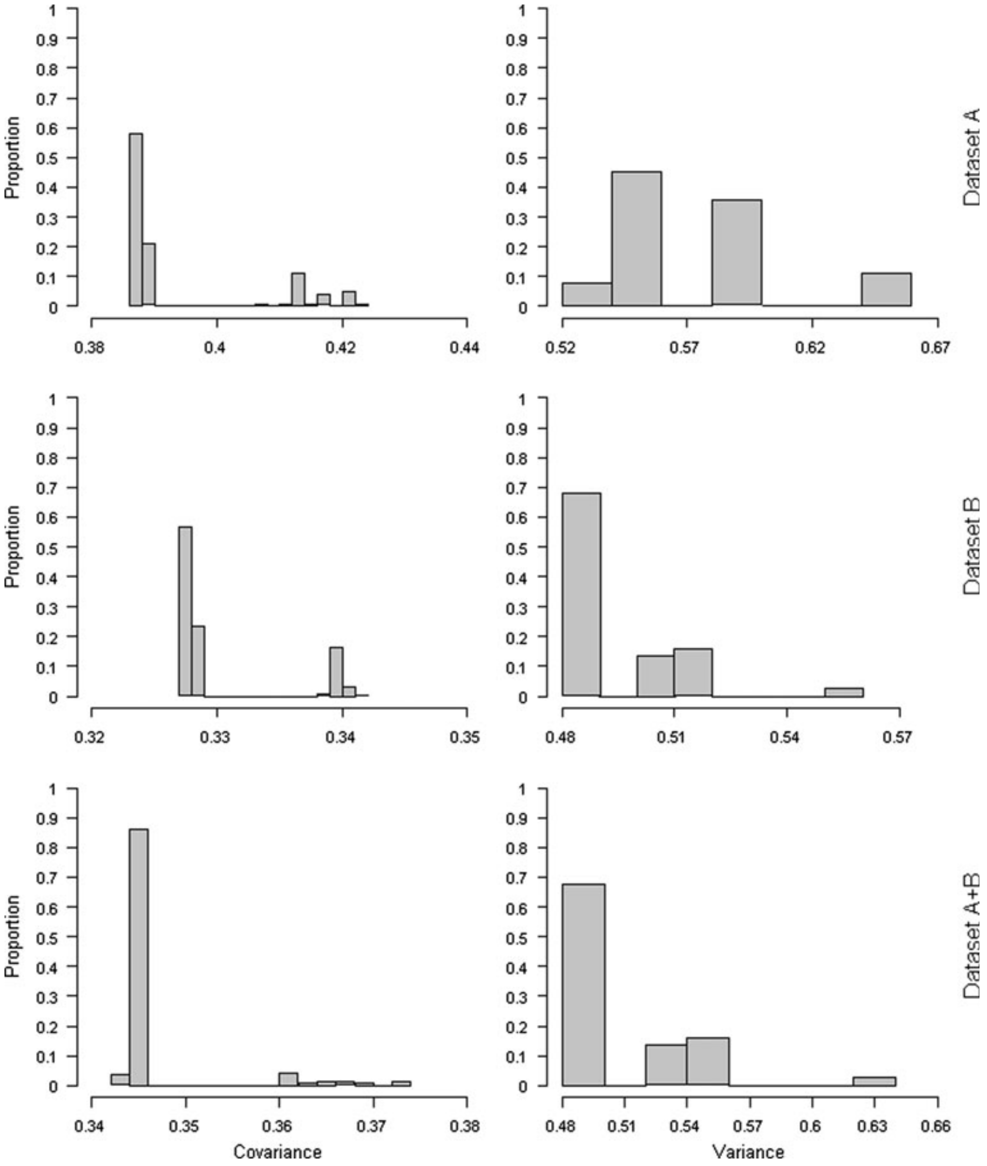

Fig. 1 *Histograms* of the variances and covariances between the adjusted means. *Left panels* covariances, *right panels* variances, *top panels* dataset A, *middle panels* dataset B, *bottom panels* dataset A + B. The relative proportions of the variances and covariances are shown on the *vertical axes*

Table 1 Pearson (lower triangle) and Spearman (upper triangle) correlations between predictions of single- and two-stage analyses

	Single-stage	Two-stage non-rotated (unweighted analysis)	Two-stage non-rotated (weighted analysis)	Two-stage rotated
Dataset A				
Single-stage	–	0.99619	0.99669	0.99997
Two-stage non-rotated (unweighted analysis)	0.99576	–	0.99807	0.99603
Two-stage non-rotated (weighted analysis)	0.99602	0.99759	–	0.99688
Two-stage rotated	0.99987	0.99546	0.99620	–
Dataset B				
Single-stage	–	0.99283	0.99363	0.99999
Two-stage non-rotated (unweighted analysis)	0.99132	–	0.99908	0.99329
Two-stage non-rotated (weighted analysis)	0.99209	0.99848	–	0.99411
Two-stage rotated	0.99995	0.99180	0.99258	–
Dataset A + B				
Single-stage	–	0.99397	0.99584	0.99996
Two-stage non-rotated (unweighted analysis)	0.99567	–	0.99817	0.99392
Two-stage non-rotated (weighted analysis)	0.99717	0.99862	–	0.99575
Two-stage rotated	0.99999	0.99559	0.99710	–

approximate i.i.d. normal errors can be assumed (Piepho et al. 2011, 2012a), were more similar to those of the single-stage analysis for all comparison criteria, namely Pearson and Spearman correlation coefficients and absolute deviations (Table 1; Fig. 3). The median absolute deviation for the non-rotated methods was 14–25 times larger than that for the rotated method. Moreover, the absolute deviation also revealed some substantial differences between predictions of the single-stage and the non-rotated two-stage methods, and attained a maximum value within the range 0.121–0.181, depending on the non-rotated method and dataset used. This difference is non-negligible when considered relative to the recorded range of variation for predictions of the single-stage analysis. Using a diagonal weighting matrix (Smith et al. 2001a) did not produce any discernible improvement over unweighted analysis for the classical two-stage method. Furthermore, the two non-rotated two-stage methods tended to not select fewer than the first n ($n = 1,\ldots,100$) best genotypes selected by the single-stage approach in the majority of cases than the rotation approach (Fig. 4). The results for the single-stage and rotation approaches were very similar but differed from those for the two-stage non-rotated approaches. For instance, of the first 20 % of the genotypes that the single-stage approach selected as the best for datasets A and B the rotation method missed none, whereas both the unweighted and weighted two-stage approaches missed three. Similarly, of the first 20 % testcross genotypes selected by the single-stage approach as the best for the combined dataset, the rotation, weighted and unweighted approaches missed zero, five and three, respectively. Overall, the prediction of the rotated method and the single-stage analysis were nearly identical, suggesting that rotation is a worthwhile pre-processing step in GS for the two-stage approach.

Comparison of boosting and RR-BLUP

The mean Pearson correlation coefficient between the GEBVs and the observed values in the validation set ranged from 0.476 to 0.710 for the rotated means, depending on the dataset and method used (Table 2). For both methods prediction accuracy was higher for dataset A than B, whereas RR-BLUP marginally outperformed boosting on both datasets by between 5.0, 6.1 and 6.0 % for the combined dataset (Table 2). The differences were highly significant for dataset A ($P = 0.0002$) and the combined dataset A + B ($P \leq 0.0001$), but only marginally significant for dataset B ($P = 0.0506$) based on the t test. On average, 61 % of the 20 % best testcross genotypes were selected by RR-BLUP and 59 % by boosting for dataset A in each validation set. For dataset B only 44 % of the 20 % best testcross genotypes were selected by RR-BLUP and 47 % by boosting. For the combined dataset 51 % of the 20 % best testcross genotypes were selected by RR-BLUP compared with 49 % by boosting. Componentwise boosting selected widely different numbers of markers as the most relevant and predictive of GEBVs, depending on both the dataset (A, B or A + B) and the training subset used. Across the 50 replicate training sets the selected number of markers ranged between 23 and 86 for dataset A, 5 and 88 for dataset B and 53 and 125 for the combined dataset.

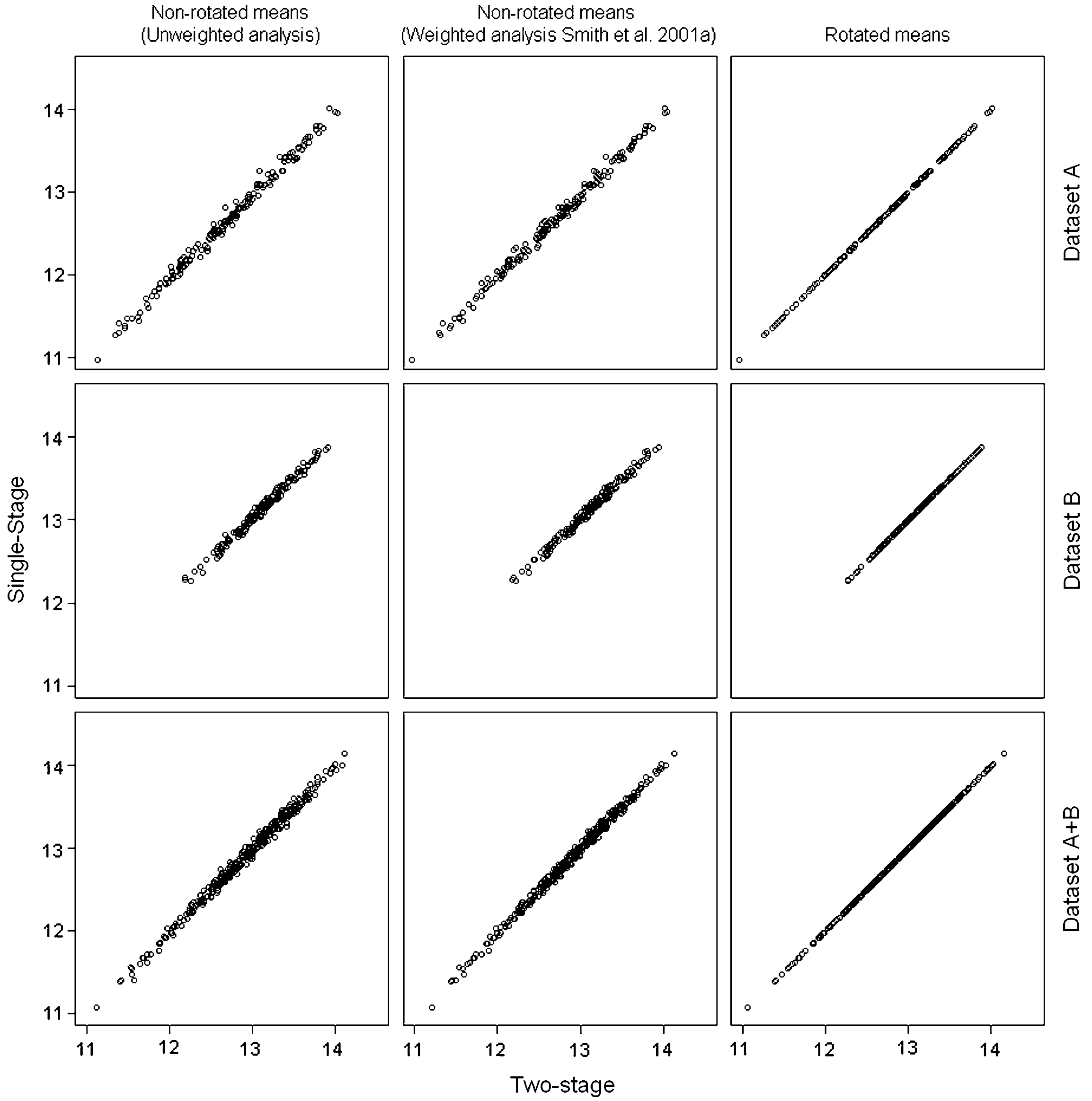

Fig. 2 Comparisons of predictions of single- and two two-stage approaches. *Left panels* non-rotated means (unweighted analysis), *middle panels* non-rotated means (weighted analysis), *right panels* rotated means across the two different datasets (*top panels* dataset A, *middle panels* dataset B, *bottom panels* dataset A + B)

Discussion

The new two-stage approach based on rotating the means was nearly perfectly correlated with and slightly more similar to the single-stage analysis than the two classical two-stage approaches based on non-rotated means, similar to the findings of Piepho et al. (2011, 2012a), from their analysis of a series of field trials. The benefit of rotation was greater in this study than in that of Piepho et al. (2011, 2012a), in which rotation was done prior to the analysis across environments. This is because, unlike the usual first stage of stage-wise analyses which involves computation of adjusted means within environments, our first stage involved computation of adjusted means across environments. This is more likely to result in a more complex variance–covariance structure due to unbalancedness in the trial design arising from not all testcross genotypes being tested in all environments. Hence, even though the rotated and non-rotated stage-wise approaches had similar correlations between their predictions and that of the single-stage approach, the absolute deviation revealed a decidedly higher accuracy for the rotated than for both the non-rotated stage-wise approaches. Moreover, the two-stage approaches tended to select fewer testcross genotypes than those selected by the single stage-stage analysis as the best. Even so, results of the rotation approach were more similar

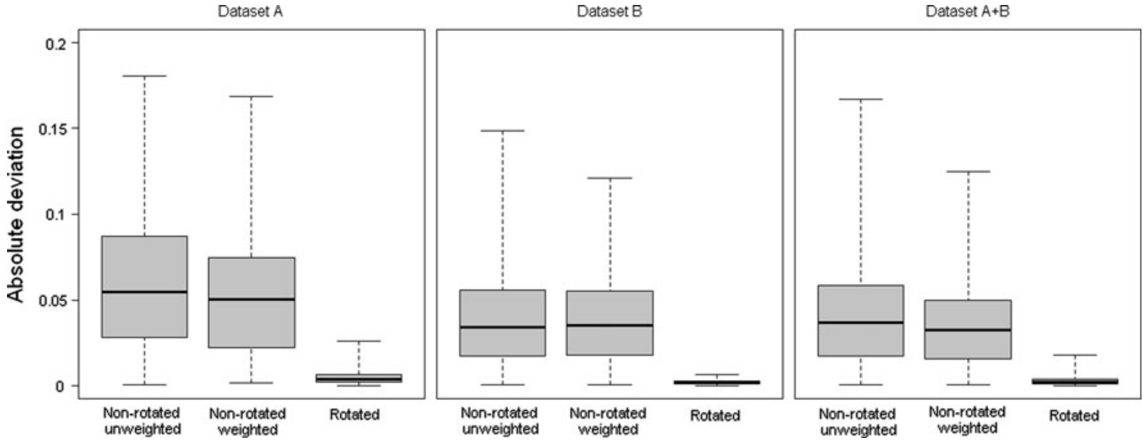

Fig. 3 *Box plots* of absolute deviations between predictions of single- and two-stage analyses. *Left panels* dataset A, *middle panels* dataset B, *right panels* dataset A + B. The *box plots* display the minimum, lower and upper quartiles, median and maximum of the absolute deviations

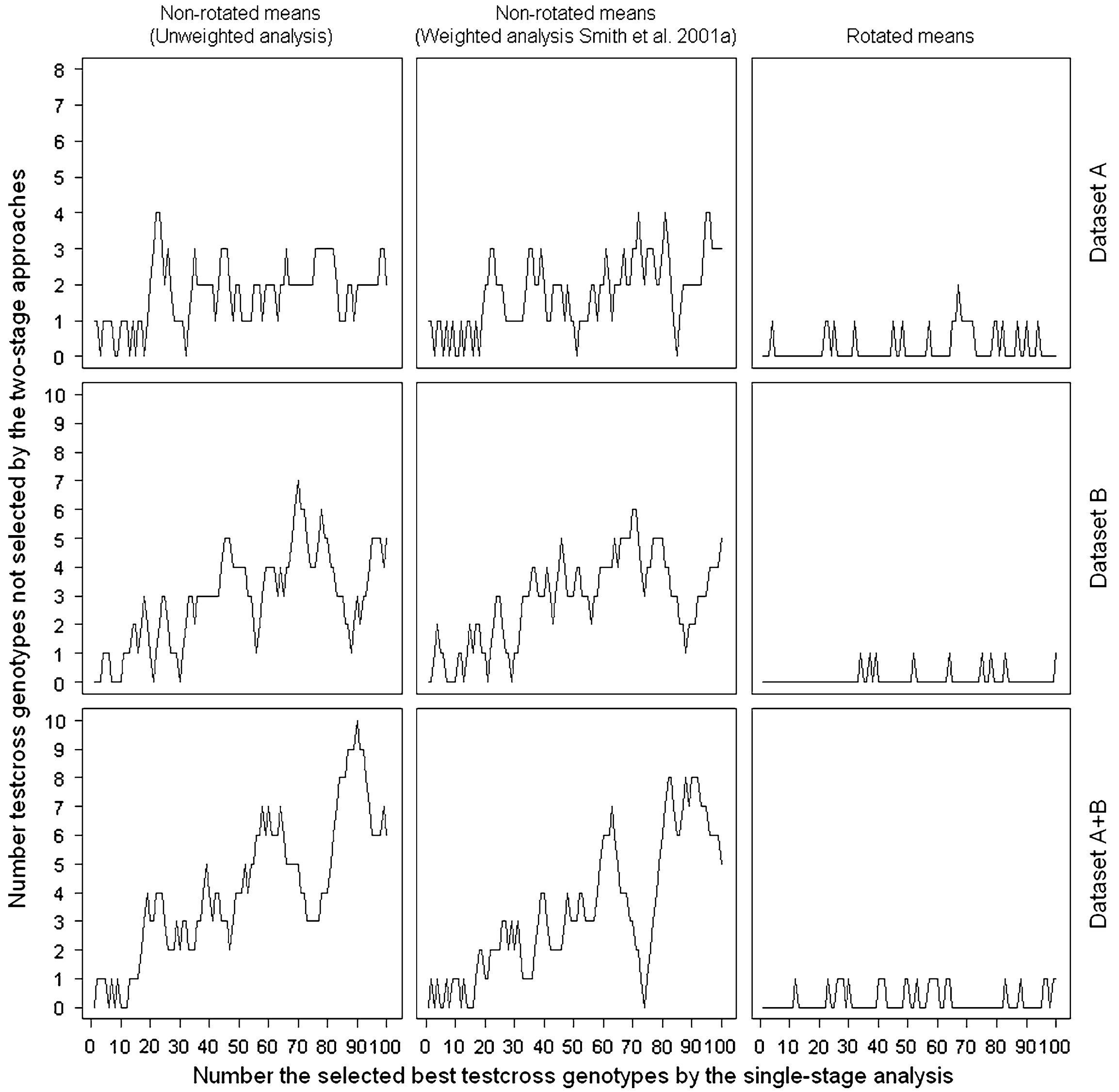

Fig. 4 Number of the first n ($n = 1,\ldots,100$) testcross genotypes selected as best by the single-stage analysis that were missed by each of the three two-stage approaches for datasets A, B and A + B. *Top panels* dataset A, *middle panels* dataset B, *bottom panels* dataset A + B

Table 2 Prediction accuracies of ridge regression BLUP and boosting

Dataset	Mean Pearson correlation		*P* value of two-sided *t* test
	RR-BLUP	Boosting	
A	0.710	0.649	0.0002
B	0.526	0.476	0.0506
A + B	0.673	0.614	<0.0001

Prediction accuracy is calculated as the Pearson correlation between rotated observed and predicted values in the validation set for each replicate and then averaging the correlations over the 50 replicates using fivefold cross-validation. The different methods were compared by a two-sided *t* test

to those for the single-stage approach than were the results for the non-rotated two-stage approaches. This finding also underlines the value of using multiple accuracy metrics in GS studies.

The slight difference between the new two-stage and single-stage analyses occurred because the variance components were unknown and so had to be estimated from the data. Results would be identical with known variance components. The difference between the single-stage and the new two-stage method would be inflated if too few degrees of freedom were available to estimate the variance components. For our dataset lack of replication of testcross genotypes within locations precluded modelling of heterogeneous variances among different locations based on the testcross genotypes. However, the standard varieties each of which had three to five replicates within each location enabled modelling of variance heterogeneity among locations. As expected, the estimated heterogeneous error variances were imprecise due to few replicates within locations, thus amplifying differences between the predictions of the single- and two-stage approaches. Even so, the rotated approach remained more similar to the single-stage approach than the other stage-wise approaches were (results not shown). Weighting the adjusted means in the second stage by the diagonal elements of the inverse of their variance–covariance matrix (Smith et al. 2001a) only slightly improved model performance, similar to expectations based on findings from earlier analyses of a series of field trials (Möhring and Piepho 2009; Smith et al. 2001a). Consequently, weighting will not always substantially improve prediction.

Our results raise the question as to what are the implications of the choice of experimental design for GS. If the trial design at each location is a randomized complete block design (RCBD) and all genotypes are tested in all locations, then all pairs of genotypes are equally correlated. In this case, single-stage and most two-stage methods coincide, provided that variance components are known. In particular, covariance among adjusted means can be accounted for by the location main effect (Möhring and Piepho 2009). The RCBD is the only practically relevant design that would allow ignoring the covariances among adjusted means. This might suggest that RCBD should be the preferred design in genomic selection, but such a suggestion is unwarranted for several reasons. In most plant breeding programs, the number of testcross genotypes to be tested is in the hundreds. With such a large number of testcross genotypes, complete blocking is well known to be utterly inefficient, and it is for this reason that most breeders will use some form of incomplete blocking. If trials are fully replicated, some kind of resolvable incomplete block design or row-column design is usually preferred (John and Williams 1995). Moreover, efficiency of analysis is often improved by the use of spatial methods (Qiao et al. 2000). Both incomplete blocking and spatial methods of analysis cause heterogeneity of covariance between adjusted means. Thus, with the currently preferred experimental designs and analysis methods, heterogeneity of covariance will remain an issue in plant breeding programs. The method proposed in this paper provides a means to efficiently confront this challenge in the GS context.

We decided to take the location main effect as random. In series of experiments, location main effects are typically very large, as they were for both the datasets we analysed, so it hardly makes any difference whether location main effects are taken as fixed or random, even when the number of locations is small (Piepho and Möhring 2006).

We have focused on results of the RR-BLUP for the single-stage analysis because it is much easier to implement than componentwise linear least squares boosting. For single-stage boosting, one would need to add a mixed-model component that reflects the field-trial design, and this mixed-model component would need to be included in boosting iterations, which is possible in principle (Tutz and Reithinger 2007), but computationally prohibitive in practice.

A comparison of the prediction accuracies of the rotated and non-rotated methods using CV was not feasible due to the absence of true breeding values. It turned out that we could not use the same benchmark with both the rotated and non-rotated means, because the appropriate benchmark for the rotated approach are the rotated adjusted means whereas for the non-rotated approaches it is the non-rotated adjusted means. Besides, for non-rotated means the basic assumption of the *k*-fold CV and its leave-one-out variant that the errors are i.i.d. and hence that the training and validation sets are independent (Arlot and Celisse 2010) is not met. The dependency between the adjusted means arises from the unbalancedness of the field trial design and use of the standard varieties. Thus, provided the same standard varieties are used with the training and validation sets, separately estimating adjusted means for both datasets will not eliminate the dependency. Additionally, for field trial designs with replications, splitting the raw data into training and validation sets prior to estimating adjusted means separately for each set is not always possible nor advisable because the trial design structure will not be preserved as desired. The ideal situation would be to have an independent validation dataset. Rotating the entire dataset before splitting it into the training and the validations is thus necessary to ensure that the training and validation sets are independent as required for valid CV.

Besides prediction, results of hypothesis tests can also be badly biased if the covariance structure of the adjusted means is ignored in a stage-wise analysis. For GS, often a pre-selection of markers is done, where the effect of each marker is tested (Hayes et al. 2009; Macciotta et al. 2009; Schulz-Streeck et al. 2011). Such tests assume i.i.d errors, which would be fulfilled if the means are rotated, but not if correlations between the adjusted means are ignored, or weighted.

In meta-analysis of individual patient data in medical trials, a closely related field, it is a standard procedure to carry the full variance–covariance forward in stage-wise analysis because this is fully efficient. Rotation, which is mainly a trick to speed up computations, is not an issue in medical trials because the number of treatments is typically small. But our rotated analysis is essentially identical to what is commonly done in meta-analysis (Mathew and Nordström 2010; Van Houwelingen et al. 2002). In two-stage analysis of plant breeding trials, however, covariance information has been either ignored or has been accounted for by approximative methods such as those considered in Smith et al. (2001a) and Möhring and Piepho (2009). The simulation results by Welham et al. (2010), where single-stage analysis is compared with the approximate two-stage method of Smith et al. (2001a), indicate that ignoring covariance information may entail a substantial loss of information.

Although more preferable, at least in theory (Cullis et al. 1998; Welham et al. 2010), the single-stage analysis can be hard to implement with massive datasets due to its high computational demands. It is therefore noteworthy that the results of the computationally more efficient new two-stage method were nearly perfectly correlated with those for the single-stage analysis and that all the two-stage approaches also had similar predictions. Moreover, the rotation approach slightly outperformed the classical two-stage approaches. Nonetheless, further simulation studies and empirical analyses are needed to determine to what extent rotation of the data can improve predictive accuracy, in particular for more complex and unbalanced datasets in which testcross genotypes from many different crosses are used and are tested in many different environments, so that the variance–covariance structures for the adjusted means may show much more complex patterns than those in our datasets.

Componentwise boosting can be used in stage-wise analyses for GS, QTL mapping and association studies with large numbers of markers because its performance on the two particular datasets and the combined dataset considered here was competitive with that of RR-BLUP; its iterative algorithm is computationally efficient as it obviates the need to invert massive matrices of marker covariates, and it automatically selects the most relevant and predictive marker subsets. Since componentwise boosting involves automatically selecting the most significant and relevant markers, its performance on our two data sets was somewhat poorer than that of RR-BLUP probably because of the low number of markers, but would be expected to improve with increasing number of markers. Compared with RR-BLUP, an infinitesimal model in which each marker is assumed to have a small effect on the trait of interest, componentwise boosting selects the most relevant markers for a given trait and may, therefore, be expected to be more precise for traits controlled by a fewer QTL with larger effects. Yet, even though relatively low (ca. 50 %) in absolute terms, both methods selected similar proportions of the 20 % best testcross genotypes in the validation sets. Also, using simulated data with about 10,000 markers, Ogutu et al. (2011) showed that the prediction accuracy for GS using boosted regression trees that differ from componentwise boosting was similar to that for RR-BLUP. Whereas Ogutu et al. (2011) used boosted regression trees, which use all the SNP markers and are able to account for arbitrarily high-order interactions to perform genomic prediction, componentwise linear least squares boosting does simultaneous automatic marker selection. Componentwise boosting also required more time as RR-BLUP did, even though the pre-processing step, which is identical, took identical times.

For both methods prediction accuracy was higher for dataset A than for B despite the similarity in the numbers of genotypes and markers in both datasets. This inconsistency in prediction accuracy between the two populations is similar to findings of other studies (e.g. Albrecht et al. 2011; Heslot et al. 2012).

Overall, the new stage-wise approach with rotated (orthogonalized) means was slightly more similar to the single-stage analysis than the classical two-stage approaches based on non-rotated means. This suggests that rotation is a promising step in GS. RR-BLUP showed slightly higher prediction accuracy than componentwise boosting on the two particular datasets, but the more important point to note is that rotation enables methods that assume i.i.d errors such as boosting to be applied to GS.

Acknowledgments We thank AgReliant Genetics for providing the datasets. This research was funded by AgReliant Genetics and the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed—Synergistic plant and animal breeding" (Grant ID: 0315526). Three anonymous referees are thanked for very useful and constructive comments.

Conflict of interest The authors declare that they have no competing interests.

References

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

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

Berk RA (2008) Statistical learning from a regression perspective. Springer, New York

Boulesteix AL, Hothorn T (2010) Testing the additional predictive value of high-dimensional molecular data. BMC Bioinforma 11:78

Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. Stat Sci 22:477–505

Buja A, Mease D, Wyner AJ (2007) Comment: boosting algorithms: regularization, prediction and model fitting. Stat Sci 22:506–512

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

Cullis BR, Thomson FM, Fisher JA, Gilmour AR, Thompson R (1996) The analysis of the NSW wheat variety database. 1. Modelling trial error variance. Theor Appl Genet 91:21–27

Cullis BR, Gogel BJ, Verbyla AP, Thompson R (1998) Spatial analysis of multi-environment early generation trials. Biometrics 54:1–18

Freund Y, Schapire R (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139

Friedman JH, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion). Ann Stat 38:367–378

Hastie TJ, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci 60:783–787

Heslot N, Yang HP, Sorrels ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner, B (2010) mboost: model-based boosting. R package version 2.0-6. http://cran.r-project.org/web/packages/mboost/

John JA, Williams ER (1995) Cyclic and computer generated designs, 2nd edn. Chapman and Hall, London

Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S (2007) Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J Anim Breed Genet 124:377–389

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

Mathew T, Nordström K (2010) Comparison of one-step and two-step meta-analysis models using individual patient data. Biom J 52:271–287

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

Möhring J, Piepho HP (2009) Comparison of weighting in two-stage analyses of series of experiments. Crop Sci 49:1977–1988

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

Piepho HP (2009) Ridge regression and extensions for genome-wide selection in maize. Crop Sci 49:1165–1176

Piepho HP, Möhring J (2006) Selection in cultivar trials—is it ignorable? Crop Sci 146:193–202

Piepho HP, Williams ER, Fleck M (2006) A note on the analysis of designed experiments with complex treatment structure. Hortic Sci 41:446–452

Piepho HP, Schulz-Streeck T, Ogutu JO (2011) A stage-wise approach for analysis of multi-environment trials. Biuletyn Oceny Odmian 33:7–20

Piepho HP, Möhring J, Schulz-Streeck T, Ogutu JO (2012a) A stage-wise approach for analysis of multi-environment trials. Biom J (in press)

Piepho HP, Ogutu JO, Schulz-Streeck T, Estaghvirou B, Gordillo A, Technow F (2012b) Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci 52:1093–1104

Qiao CG, Basford KE, DeLacy IH, Cooper M (2000) Evaluation of experimental designs and spatial analysis in wheat breeding trials. Theor Appl Genet 100:9–16

Rao CR, Toutenburg H, Shalabh, Heumann C (2008) Linear models and generalizations least squares and alternatives. Springer, Berlin

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

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

Schulz-Streeck T, Ogutu JO, Piepho HP (2011) Pre-selection of markers for genomic selection. BMC Proc 5(Suppl 3):S12

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

Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York

Smith AB, Cullis BR, Gilmour AR (2001a) The analysis of crop variety evaluation data in Australia. Aust N Z J Stat 43:129–145

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

Tutz G, Reithinger F (2007) A boosting approach to flexible semiparametric mixed models. Stat Med 26:2872–2900

Van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 21:589–624

Welham S, Gogel BJ, Smith AB, Thompson R, Cullis BR (2010) A comparison of analysis methods for late-stage evaluation trials. Aust N Z J Stat 52:125–149