

# Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction

Steven Maenhout · Bernard De Baets ·  
Geert Haesaert

Received: 31 March 2009 / Accepted: 22 October 2009 / Published online: 11 November 2009  
© Springer-Verlag 2009

**Abstract** Accurate prediction of the phenotypic performance of a hybrid plant based on the molecular fingerprints of its parents should lead to a more cost-effective breeding programme as it allows to reduce the number of expensive field evaluations. The construction of a reliable prediction model requires a representative sample of hybrids for which both molecular and phenotypic information are accessible. This phenotypic information is usually readily available as typical breeding programmes test numerous new hybrids in multi-location field trials on a yearly basis. Earlier studies indicated that a linear mixed model analysis of this typically unbalanced phenotypic data allows to construct  $\varepsilon$ -insensitive support vector machine regression and best linear prediction models for predicting the performance of single-cross maize hybrids. We compare these prediction methods using different subsets of the phenotypic and marker data of a commercial maize breeding programme and evaluate the resulting prediction accuracies by means of a specifically designed field experiment. This balanced field trial allows to assess the reliability of the cross-validation prediction accuracies reported here and in earlier studies. The limits of the predictive capabilities of

both prediction methods are further examined by reducing the number of training hybrids and the size of the molecular fingerprints. The results indicate a considerable discrepancy between prediction accuracies obtained by cross-validation procedures and those obtained by correlating the predictions with the results of a validation field trial. The prediction accuracy of best linear prediction was less sensitive to a reduction of the number of training examples compared with that of support vector machine regression. The latter was, however, better at predicting hybrid performance when the size of the molecular fingerprints was reduced, especially if the initial set of markers had a low information content.

## Introduction

The prediction of phenotypic performance from molecular marker data receives increasing attention from plant breeders, as the cost of phenotyping is gradually overtaking the cost of genotyping (Bernardo 2008). In this field of research, plant species for which it is relatively easy to create and cross almost fully homozygous inbred lines, are particularly useful as they allow to study the effect of a single gamete in different genetic backgrounds. Maenhout et al. (2007) use data that were generated in a commercial maize breeding programme to compare the phenotypic prediction accuracy of  $\varepsilon$ -insensitive support vector machine regression ( $\varepsilon$ -SVR) to that of the method advocated by Bernardo (1994, 1995, 1996a, b) based on best linear prediction (BLP). The reported prediction accuracies, determined by means of a leave-one-out cross-validation routine, indicate that both methods are equally good at predicting phenotypes for three important agronomic traits. In this study, we further examine several key aspects of

---

Communicated by M. Cooper.

---

Contribution to the special issue "Heterosis in Plants".

---

S. Maenhout (✉) · G. Haesaert  
Department of Biosciences and Landscape Architecture,  
University College Ghent, Voskenslaan 270,  
9000 Ghent, Belgium  
e-mail: Steven.Maenhout@hogent.be

B. De Baets  
Department of Applied Mathematics,  
Biometrics and Process Control, Ghent University,  
Coupure links 653, 9000 Ghent, Belgium

hybrid prediction by means of  $\varepsilon$ -SVR and BLP which allows to clarify the strengths and weaknesses of both methods.

Field trial data originating from commercial hybrid breeding programmes are typically very unbalanced. Tester lines are parents of many hybrids, while other inbred lines may appear only once in the company's pedigree. Furthermore, there is usually quite a substantial difference in the number of field trials in which a promising hybrid is tested compared with the often single trial results of the lesser candidates. Both  $\varepsilon$ -SVR and BLP require a set of hybrids for which a molecular fingerprint and a single response value for each trait are available. Such a phenotypic response value or score can be obtained by means of a linear mixed model analysis of the unbalanced phenotypic data, but different model assumptions and prediction approaches can lead to very different results. We study the impact of these assumptions by comparing three different data preparation methods. In the linear mixed models described by Bernardo (1994, 1995, 1996a, b) and Maenhout et al. (2008), the non-genetic effects of growing seasons, locations and blocks are assumed to be fixed while the genotypic and  $G \times E$  effects are assumed to be random. Bernardo (1994, 1995, 1996a, b) obtains a single phenotypic score for a particular hybrid by taking the average of all its phenotypic measurements, after correcting them by means of the estimated fixed effects. Maenhout et al. (2008) on the other hand, aggregate the BLUPs of the genotypic components directly to obtain a single score for each hybrid. Besides these two data preparation methods, we also study a third approach in which the genotypic effects are assumed to be fixed while the non-genetic nuisance parameters are treated as random.

Maenhout et al. (2007) use all hybrids that are represented in the available unbalanced phenotypic data and the entire set of genotyped molecular markers to compare the prediction accuracy of  $\varepsilon$ -SVR and BLP. The sensitivity of both methods to a reduction in the number of training examples or genotyped molecular markers is, however, left unexamined. To assess the impact of the training sample size and marker information content on the prediction accuracy, we apply both methods to selected subsets of the training sample and molecular marker fingerprint. The results allow to identify minimum sample size requirements of  $\varepsilon$ -SVR and BLP models that are trained using comparable, unbalanced data sets.

The accuracy of hybrid prediction techniques is generally measured by some form of cross-validation strategy (Bernardo 1994, 1995, 1996a, b; Charcosset et al. 1998; Schrag et al. 2007, 2009). Schrag et al. (2007) argue that an assessment of prediction accuracy by means of a leave-one-out cross-validation routine does not reflect practical

breeding circumstances where a new inbred line would be crossed with only a few tester lines from the opposite heterotic group. They propose a modified cross-validation sampling scheme that requires a mating design in which every inbred line from one heterotic group is crossed with all lines belonging to the complementary heterotic group. To allow for such a realistic assessment of prediction accuracy in an unbalanced setting, Bernardo (1996b) and Maenhout et al. (2008) use cross-validation schemes that simulate a lack of prior information on one or both parental inbred lines of a newly created hybrid. Although these schemes represent an improvement, they do not solve the fundamental problem of cross-validation-based accuracy measures. As the training examples are predicted marginal to the effects of growing seasons, test locations and possibly fertiliser or irrigation treatments, the resulting cross-validation-based prediction accuracy measures do not take into account the extra level of uncertainty that is caused by  $G \times E$  effects (Welham et al. 2004). This implies that the observed correlation between the predicted marginal genotypic values and those estimated conditional on a specific level of the environmental factors (i.e. in an additional field trial in a specific year and geographical region) might differ substantially from the cross-validation-based prediction accuracy. To quantify this expected discrepancy, we performed a validation field trial using 49 hybrids which were created by crossing seven Iodent lines with seven Iowa stiff stalk synthetic (ISSS) lines. The phenotypic performance of these hybrids was measured in a multi-environment trial at three locations in the South of France. Prediction accuracy is determined by correlating the resulting estimates for total genotypic value and SCA to the predictions of  $\varepsilon$ -SVR and BLP models, constructed from the unbalanced training data.

To summarise, we recapitulate the three main objectives of our research: (1) to identify the best method for distilling a single phenotypic score for each hybrid in an unbalanced data set, (2) to compare the prediction accuracy of  $\varepsilon$ -SVR and BLP when the sample size and information content of the molecular marker fingerprint are reduced and (3) to compare the prediction accuracy measures obtained through various cross-validation schemes with those obtained by means of a validation field trial.

## Materials and methods

To achieve the three objectives, this study investigates the impact of changing the levels of the factors influencing them which are summarised in Table 1 and discussed below.

**Table 1** Overview of the different traits, training data preparation methods, molecular marker-based predictors, prediction methods, sampling schemes and methods for prediction accuracy measurement that are combined in this study

Factor	Levels
Trait	Grain yield Grain moisture content Days until flowering
Training data preparation	( $\hat{y}_T^{rp}$ ) random phenotypes ( $\hat{y}_T^{rg}$ ) random genotypes ( $\hat{y}_T^{fg}$ ) fixed genotypes ( $\hat{d}_T^{rs}$ ) random SCA
Predictor	AFLP SSR
Prediction method	$\epsilon$ -SVR BLP
Sampling scheme	Random sampling Test-cross sampling New-cross sampling Random marker reduction
Prediction accuracy measurement	Cross-validation Validation field trial: ( $\hat{y}_V^{rg}$ ) random genotypes ( $\hat{d}_V^{rs}$ ) random SCA

## Training data

### Data description

The data used in this study are a subset of the genotypic and phenotypic information generated by the grain maize breeding programme of the private company RAGT R2n, and is described in detail in Maenhout et al. (2007, 2008, 2009). It contains 40,432 phenotypic measurements on 2,354 hybrids originating from unbalanced crosses between 92 Iodent and 105 ISSS lines. We study the traits grain yield, grain moisture content and days until flowering, which were measured in 1,280 multi-environment trials representing 110 locations spread over Europe from 1989 to 2005. The 197 parental inbred lines are genotyped with 101 SSR markers, which are evenly distributed over the maize genome according to the proprietary linkage map of RAGT R2n. Due to problems identifying some SSR alleles (null alleles), only 75 markers, which have a complete profile over all inbred lines, are used. AFLP fingerprints are generated using 11 *Pst*I–*Mse*I and 4 *Eco*RI–*Mse*I primer combinations producing 569 polymorphic bands in total.

### Data analysis

The construction of an  $\epsilon$ -SVR or BLP prediction model for a specific quantitative trait requires a single response value

for each training example representing the genetic potential of each genotype at each location and year. We consider three methods of constructing such a response value based on linear mixed modelling of the trial data. We also predict SCA values from a mixed model analysis.

**Random phenotypes** In the first approach, we consider the environmental effects (e.g. year, location, block, etc.) as fixed effects, while we consider GCA, SCA and all  $G \times E$  interactions as random effects. A detailed description of this linear mixed model for the three traits under study can be found in Maenhout et al. (2009). The variance structures of GCA and SCA effects are modelled according to Stuber and Cockerham (1966) where we use the AFLP fingerprints to obtain estimators for the pairwise coefficient of coancestry between inbred lines  $i$  and  $j$  belonging to the same heterotic group as (Bernardo 1993)

$$f_{ij} = \frac{f_{ij}^{JAC} - \frac{1}{2}(\bar{f}_i^{JAC} + \bar{f}_j^{JAC})}{1 - \frac{1}{2}(\bar{f}_i^{JAC} + \bar{f}_j^{JAC})}, \quad (1)$$

where  $f_{ij}^{JAC}$  is the Jaccard similarity coefficient between the AFLP fingerprints of lines  $i$  and  $j$ .  $\bar{f}_i^{JAC}$  is the average Jaccard similarity coefficient between inbred line  $i$  and all lines belonging to the opposite heterotic group. This estimator for the coefficient of coancestry resulted in the highest restricted log-likelihood, when compared with several other estimators that use pedigree, AFLP or SSR marker information (Maenhout et al. 2009). The genotypic estimate is obtained by averaging over all measurements of a single hybrid in the response vector  $\mathbf{y}$  after correction for the estimated fixed environmental effects as

$$\hat{\mathbf{y}}_T^{rp} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where  $\mathbf{Z}$  is a design matrix linking the phenotypic measurements in vector  $\mathbf{y}$  to each hybrid in vector  $\hat{\mathbf{y}}_T^{rp}$ . Vector  $\hat{\boldsymbol{\beta}}$  contains the estimated effects for the levels of each nuisance factor and these are linked to the response vector  $\mathbf{y}$  by means of the design matrix  $\mathbf{X}$ . Bernardo (1994, 1995, 1996a, b) calls the entries in vector  $\hat{\mathbf{y}}_T^{rp}$  phenotypes, as these are not corrected for  $G \times E$  interaction effects or residual error. The superscript rp is shorthand for random phenotypes, while the subscript T indicates that this vector was obtained from the training data.

**Random genotypes** The second method is to sum the appropriate GCA and SCA BLUPs obtained from the aforementioned linear mixed model analysis as

$$\hat{\mathbf{y}}_T^{rg} = \mathbf{Z}_s\hat{\mathbf{a}}_s + \mathbf{Z}_o\hat{\mathbf{a}}_o + \hat{\mathbf{d}}_T^{rs}, \quad (2)$$

where  $\hat{\mathbf{a}}_s$  and  $\hat{\mathbf{a}}_o$  are vectors containing BLUPs of the GCA values of lines belonging to the ISSS and Iodent heterotic groups, respectively. The design matrices  $\mathbf{Z}_s$  and  $\mathbf{Z}_o$  link

each hybrid to the appropriate parental inbred lines. Vector  $\hat{\mathbf{d}}_T^{\text{rs}}$  contains a BLUP of the SCA value for each hybrid. As we treat the GCA and SCA effects as random model factors, we use the superscript rg to indicate this random nature of the genotypic values in vector  $\hat{\mathbf{y}}_T^{\text{rg}}$ . This approach implicitly produces genotypic scores that are marginal to all environmental factors in the model such as growing seasons and locations. These marginal scores have larger standard errors compared with estimators that are conditional on one or more environmental factors (Welham et al. 2004), but we prefer them here as they do not require knowledge of the future environmental conditions in which the predicted hybrids will be grown.

**Fixed genotypes** The third method of forming a genotypic response is from a linear mixed model with genotypes fixed and non-genetic effects fitted as random. This approach allows to obtain a vector of estimated genotypic fixed effects  $\hat{\mathbf{y}}_T^{\text{fg}}$  without making prior assumptions on the covariance structure of the GCA and SCA components.

**Random SCA** Besides training on genotypic or phenotypic scores, we also construct prediction models for the values in vector  $\hat{\mathbf{d}}_T^{\text{rs}}$  of Eq. 2.

#### Validation data

#### Data description

Seven ISSS and Seven Iodent lines were selected from the initial set of 197 inbred lines and pairwise intermated to produce 49 cross-heterotic hybrids. For these hybrids and an additional six check varieties, the traits grain yield, grain moisture content and days until flowering were measured in a balanced field trial at three locations in the South of France during the growing season of 2008. The initial selection of 14 parental inbred lines was based on the  $\varepsilon$ -SVR and BLP predictions of all 9,660 possible hybrids between the 105 ISSS and 92 Iodent lines. A greedy search heuristic was used to approach the optimal selection of 14 parental inbred lines such that the  $\varepsilon$ -SVR and BLP predictions of the 49 hybrids show the largest variance in grain yield. However, several lines in this initial selection were replaced by other lines so that all hybrids had a comparable maturity index. Only 11 of the 49 hybrids were in fact new combinations, while the other 38 already had phenotypic records in the training data. Regardless of potential seed availability, each of the 49 crosses were (re)created under the exact same circumstances, as to avoid non-genetic seed quality differences. At each location of the trial, the 55 hybrids are laid out as a two-replicate resolvable row-column design with 22 rows and 5 columns.

#### Data analysis

A linear mixed model analysis is performed assuming location effects as fixed and all genetic components and  $G \times E$  interactions as random. The description of the statistical model follows the notation of Smith et al. (2001) where the vector of phenotypic measurements  $\mathbf{y}$  is decomposed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_0\mathbf{u} + \mathbf{e}, \quad (3)$$

and  $\boldsymbol{\tau}$  is a vector of fixed effects containing main location effects and location-specific effects correcting for extraneous field variation.  $\mathbf{g} = (\mathbf{g}'_1, \mathbf{g}'_2, \mathbf{g}'_3)'$  is a vector containing the random effects of the 55 hybrids in each of the three locations with an associated design matrix  $\mathbf{Z}_g$ .  $\mathbf{u}$  is also a vector of random effects modelling for location-specific blocking factors. The vector of residuals  $\mathbf{e} = (\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)'$  is partitioned in three subvectors corresponding to the three locations. For the trait grain moisture contents, the values in vector  $\mathbf{y}$  were logit transformed.

The vector of genetic effects  $\mathbf{g}$  is partitioned as

$$\mathbf{Z}_g\mathbf{g} = \mathbf{Z}_c\mathbf{c} + \mathbf{Z}_s\mathbf{a}_s + \mathbf{Z}_0\mathbf{a}_0 + \mathbf{Z}_d\mathbf{d}, \quad (4)$$

where  $\mathbf{c} = (\mathbf{c}'_1, \mathbf{c}'_2, \mathbf{c}'_3)'$  represents a vector containing the genetic effects of the six check varieties at each of the three locations, vectors  $\mathbf{a}_s$  and  $\mathbf{a}_0$  contain the GCA effects of the parental inbred lines belonging to the ISSS and Iodent heterotic groups, respectively, and vector  $\mathbf{d}$  contains the SCA effects of the 49 hybrids at each location. The design matrices  $\mathbf{Z}_c$ ,  $\mathbf{Z}_s$ ,  $\mathbf{Z}_0$  and  $\mathbf{Z}_d$  separate check and non-check entries and matrices  $\mathbf{Z}_s$  and  $\mathbf{Z}_0$  have the additional function of linking the appropriate parental inbred lines to each non-check hybrid in vector  $\mathbf{g}$ . Additional details on the fitted variance structures for vectors  $\mathbf{g}$  and  $\mathbf{e}$  can be found in the [Appendix](#).

The linear mixed model analysis of the validation trial data provides BLUPs for the GCA and SCA components which are summed according to Eq. 2 to obtain an estimate of the genotypic value for each of the 49 hybrids. These estimates are grouped in the vector  $\hat{\mathbf{y}}_V^{\text{rg}}$  where the subscript  $V$  indicates their validation trial origin. The vector  $\hat{\mathbf{d}}_V^{\text{rs}}$  contains the BLUPs of the 49 SCA values.

#### Prediction methods

##### $\varepsilon$ -Insensitive support vector machines regression

Support vector machines (SVM) are a class of machine learning methods developed by Vapnik (1995) for classification and regression purposes. A good tutorial on SVM classification can be found in Burges (1998), while Smola and Schölkopf (2004) present the underlying ideas of  $\varepsilon$ -insensitive support vector machines regression. Maenhout

et al. (2007) show how  $\varepsilon$ -SVR can be used to predict the phenotypic performance of new hybrids using unbalanced phenotypic training data and AFLP or SSR marker fingerprints as predictors. Cross-validation results indicated that solving the linear regression problem in an infinite-dimensional space by means of a Gaussian kernel function results in a higher prediction accuracy compared with a linear solution in the original input space. Using the Gaussian kernel function requires a value for the kernel parameter  $\gamma$  and the optimisation function that is minimised during the construction of an  $\varepsilon$ -SVR prediction model requires two additional parameters  $C$  and  $\varepsilon$ . In Maenhout et al. (2007, 2008) optimal values for  $C$ ,  $\varepsilon$  and  $\gamma$  are found by an expensive grid-search over this three-dimensional space with a  $\nu$ -fold cross-validation prediction accuracy as optimisation criterion. To reduce the computational effort, the  $\varepsilon$ -SVR parameter searches in the present study were guided by the efficient global optimisation or EGO algorithm reported by Jones et al. (1998). The criterion to be optimised was the squared Pearson correlation coefficient obtained by a  $\nu$ -fold cross-validation where  $\nu = 20$ .

#### Best linear prediction

Bernardo (1994, 1995, 1996a, b) makes predictions for a set of single crosses as

$$\hat{y}_P = C_{PT} V_T^{-1} \hat{y}_T, \quad (5)$$

where  $C_{PT}$  is the genetic covariance matrix between the hybrids in the training set  $\hat{y}_T$  and the hybrids to be predicted and  $V_T = \text{Var}(\hat{y}_T)$  is the variance matrix of the hybrids in the training set. The genetic covariances in the matrices  $C_{PT}$  and  $V_T$  are obtained from a simplification of the covariance model described in Stuber and Cockerham (1966)

$$\text{Cov}(h_{ij}, h_{i'j'}) = \theta_{ii'} \sigma_s^2 + \theta_{jj'} \sigma_o^2 + \theta_{ii'} \theta_{jj'} \sigma_d^2,$$

where  $h_{ij}$  and  $h_{i'j'}$  are two hybrids for which the parental inbred lines  $i$  and  $i'$  belong to the ISSS heterotic group and the lines  $j$  and  $j'$  belong to the Iodent group.  $\theta_{ii'}$  and  $\theta_{jj'}$  are the coefficients of coancestry estimated from SSR (Bernardo 1993) or AFLP marker information, the latter based on Eq. 1. The additive variance parameters  $\sigma_s^2$  and  $\sigma_o^2$  and the dominance variance  $\sigma_d^2$  are obtained from the REML analysis of the training data.

We obtain  $\hat{y}_P$  from Eq. 5 by solving the system of linear equations

$$V_T x_T = \hat{y}_T \quad (6)$$

for  $x_T$  through a Cholesky decomposition of  $V_T$ . The vector  $x_T$  then allows to calculate  $\hat{y}_P$  as

$$\hat{y}_P = C_{PT} x_T.$$

#### Reduction of the training data

Previous reports on  $\varepsilon$ -SVR and BLP hybrid prediction have assumed the availability of phenotypic measurements on a large number of hybrids. For both prediction methods, a reduction in prediction accuracy is to be expected if the size of the training set is decreased. A large sample size does, however, not necessarily imply a high prediction accuracy as the relevance of the training examples with respect to the future cross predictions, is of equal importance. Also the size and information content of the molecular fingerprints has an impact on the reliability of the prediction model as a smaller marker resolution implies a reduced chance of detecting marker-trait associations and less precise estimates of the genetic covariance between relatives.

#### Training sample size

In an attempt to assess the impact of the size of the training sample on the prediction accuracy of both  $\varepsilon$ -SVR and BLP, we employ three sampling schemes to obtain subsets of the original RAGT data set. For each sampling scheme, the prediction accuracy is determined in two ways: (1) by means of cross-validation on the training vectors  $\hat{y}_T^{\text{rg}}$  and  $\hat{d}_T^{\text{rs}}$  for predictions on total genotypic value and SCA, respectively, (2) by correlating against the validation vectors  $\hat{y}_V^{\text{rg}}$  and  $\hat{d}_V^{\text{rs}}$ . The 38 hybrids that are common to training and validation data, are removed from the vectors  $\hat{y}_T^{\text{rg}}$  and  $\hat{d}_T^{\text{rs}}$  when the second prediction accuracy measure is used.

**Random sampling** For the random sampling scheme, the hybrids in the full training set are successively split at random to form smaller data sets from which  $\varepsilon$ -SVR and BLP prediction models are constructed. Initially, the prediction accuracy of both methods using all but one training examples is determined by means of a leave-one-out cross-validation (1). Predictions on the 49 hybrids that were tested in the validation field trial are obtained from  $\varepsilon$ -SVR and BLP models that were constructed from the 2,316 non-validated hybrids (2). In the next step, the number of training examples made available to  $\varepsilon$ -SVR and BLP is cut in half and the cross-validation-based prediction accuracy is determined by making predictions on the other half of the training examples (1). The set of 2,316 non-validated hybrids is also randomly split in half and used to make  $\varepsilon$ -SVR and BLP-based predictions on the 49 validation hybrids (2). In subsequent steps, the number of training examples made available to  $\varepsilon$ -SVR and BLP is reduced further by randomly splitting the training data in  $2^p$  pieces for  $p = 1, \dots, 6$ . The whole process is repeated 100 times resulting in  $100 \sum_{p=1}^6 2^p = 12,600$  distinct  $\varepsilon$ -SVR and BLP prediction models.



**Test-cross sampling** The test-cross sampling scheme simulates the prediction of a hybrid formed by crossing a newly created inbred line with a well-known tester line. For each of the 197 inbred lines in the original data set, a separate  $\varepsilon$ -SVR and BLP prediction model is constructed using only information from hybrids that are not a child of that particular inbred. The resulting prediction models are used to predict the performance of the left-out hybrids and those hybrids in the validation data set that also have that particular inbred line as a parent. This sampling scheme, therefore, results in two predictions for each hybrid as both parental inbred lines function once as tester and once as newly developed line. In a balanced mating design (i.e. all 9,960 distinct crosses between the Iodent and the ISSS lines are made), this sampling scheme would allow to assess the obtained prediction accuracy for Type 1 hybrids as defined by Schrag et al. (2009a, b).

**New-cross sampling** The third sampling scheme simulates the prediction of a hybrid formed by crossing two newly developed inbred lines. Although this situation is rather uncommon in hybrid breeding programmes, it allows to compare  $\varepsilon$ -SVR and BLP in a worst-case scenario. For each hybrid in the dataset, a specific  $\varepsilon$ -SVR and BLP prediction model is constructed by removing all hybrids from the training set that have a parental inbred line in common with the selected hybrid. This sampling scheme relates to the Type 0 hybrids of Schrag et al. (2009a, b).

#### Molecular marker information content

The impact of the information content of the molecular fingerprints is examined by taking random subsets of the available SSR or AFLP markers and subsequent construction of the  $\varepsilon$ -SVR and BLP prediction models. Again, prediction accuracy is determined by means of (1) cross-validation and (2) correlating against the estimates obtained from the validation trial. The size of the set of predictor markers is reduced in steps of 10% of the original fingerprint size and at each step, 100 iterations of the sampling routine are performed. Reducing the set of molecular markers often results in a singular coancestry matrix which prevents its inversion during the construction of a BLP prediction model. This situation occurs if the marker-based estimate of the variance matrix of the training hybrids is rank deficient and, therefore, does not allow for a unique solution of the system of linear equations in Eq. 6. Any estimated variance matrix should be at least positive semi-definite as explained in Maenhout et al. (2009) but in the present case, the marker-based estimate of the genetic covariance matrix  $V_T$  should be strictly positive definite as its Cholesky decomposition is used to make predictions on new hybrids. If the estimated covariance

matrix, obtained from the reduced set of molecular markers, is singular, we obtain the minimum norm, least squares solution to Eq. 6. Other solutions might result in higher correlations but without relying on the validation data, there is no biological justification for preferring these solutions over the least squares solution.

## Results

### Unbalanced data handling

Three quarters of the hybrids in the validation field trial have measurements in the unbalanced training data set. These 38 hybrids, therefore, allow to identify the best way of obtaining a single hybrid score from unbalanced phenotypic data. Table 2 gives an overview of the observed correlations between the different types of hybrid scores and the genotypic estimates obtained from the validation field trial measurements. The latter were collected during one growing season at three locations in a specific region of France and as such, represent only a small part of the  $G \times E$  space spanned by the training data. The correlations presented are, therefore, susceptible to environmental changes but should, however, allow for a relative comparison between the different data handling methods.

### Reduction of the training data

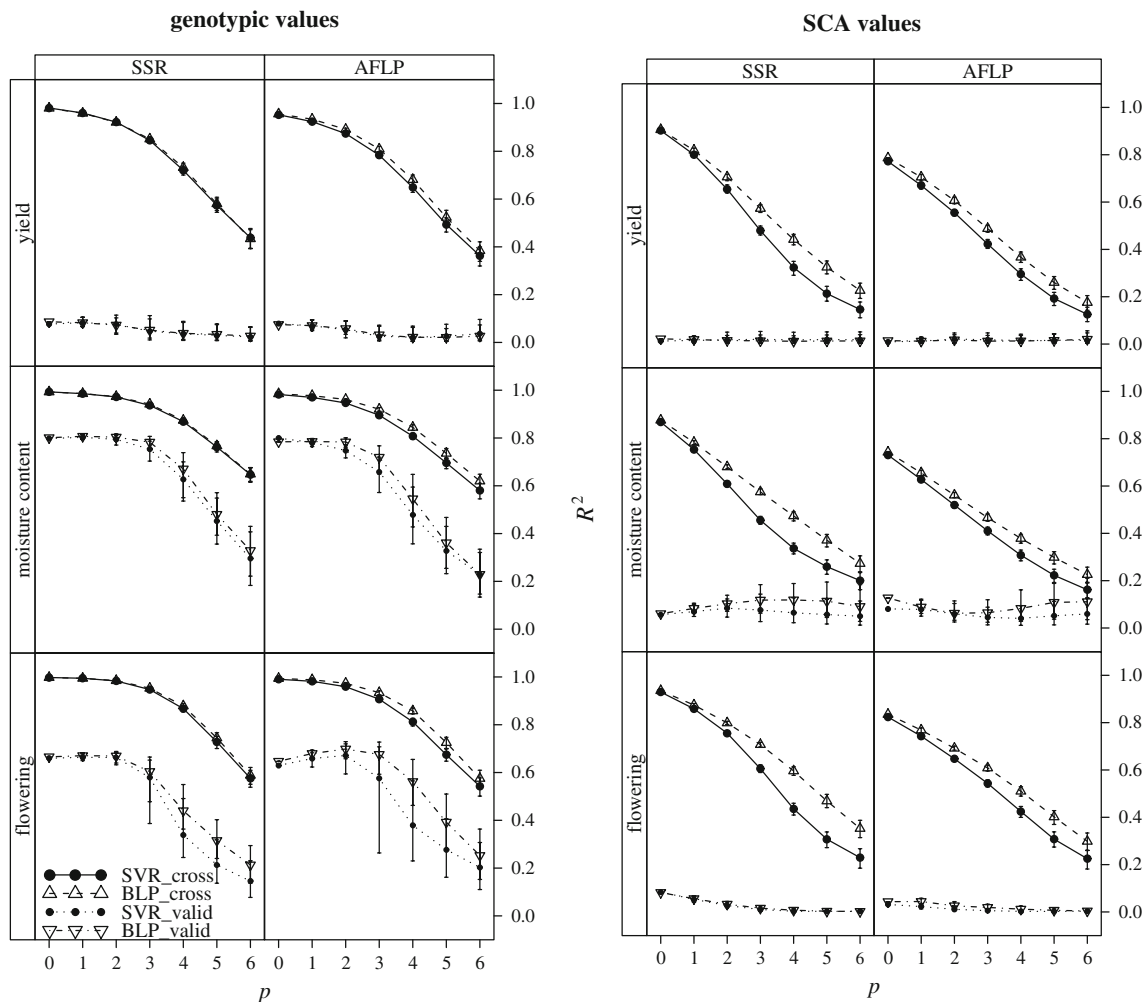
#### Training sample size

**Random sampling** Figure 1 shows the prediction accuracy obtained by  $\varepsilon$ -SVR and BLP prediction models that were constructed by reducing the initial set of the training examples in the vectors  $\hat{y}_T^{rg}$  and  $\hat{d}_T^{rs}$ .  $p = 0$  indicates that a leave-one-out cross-validation is performed and predictions

**Table 2** Squared Pearson correlation coefficients between the different types of training scores ( $\hat{y}_T^{ip}$ ,  $\hat{y}_T^{rg}$ ,  $\hat{y}_T^{fg}$ ) and SCA BLUPs ( $\hat{d}_T^{rs}$ ) obtained from the unbalanced phenotypic data set and the scores ( $\hat{y}_V^{rg}$ ) and SCA estimates ( $\hat{d}_V^{rs}$ ) obtained from measurements taken in the balanced validation field trial for the 38 common hybrids

Score vector	Validation data $\hat{y}_V^{rg} / \hat{d}_V^{rs}$		
	Yield	Moist. cont.	Flowering
Training data			
$\hat{y}_T^{ip}$	0.04	0.61	0.43
$\hat{y}_T^{rg}$	<b>0.19</b>	<b>0.79</b>	<b>0.72</b>
$\hat{y}_T^{fg}$	0.05	0.59	0.43
$\hat{d}_T^{rs}$	0.03	0.15	0.17

For each trait, the combination of scores with the highest correlation is set in bold



**Fig. 1**  $\epsilon$ -SVR and BLP prediction accuracies obtained by training on subsets of the vector of genotypic values  $\hat{y}_T^{rg}$  and the vector of SCA BLUPs  $\hat{d}_T^{rs}$ . At  $p = 0$ , a leave-one-out cross-validation is performed on the training data and predictions on the 49 hybrids are made by training on all 2,316 training hybrids. At  $p = 1, \dots, 6$ , an  $\epsilon$ -SVR and BLP prediction model are constructed from the  $2^p$  subsets of the original vectors and AFLP or SSR predictor information. For each of these models, predictions are made for all training hybrids that are not

in the particular subset and all 49 hybrids of the validation data set. This subset assignment procedure is replicated 100 times. Accuracy is expressed as the median of the squared Pearson correlation coefficient between the predictions for all hybrids and their corresponding entries in the training vectors  $\hat{y}_T^{rg}$ ,  $\hat{d}_T^{rs}$  (suffix cross), and the validation vectors  $\hat{y}_V^{rg}$  and  $\hat{d}_V^{rs}$  (suffix valid). The error bars indicate the 0.25 and 0.75 quantiles of each sampling distribution

for the validation trial hybrids are obtained from  $\epsilon$ -SVR and BLP models that are trained on the full vector  $\hat{y}_T^{rg}$  or  $\hat{d}_T^{rs}$ , minus the entries of the 38 common hybrids. For each of the 100 iterations at  $p = 1, \dots, 6$ , the training hybrids are randomly assigned to one of  $2^p$  subsets and for each of these subsets, an  $\epsilon$ -SVR and BLP prediction model is constructed. These models are subsequently used to make predictions on (1) all hybrids that are not included in the training subset and (2) the 49 hybrids tested in the validation field trial. Despite the promising cross-validation results for SCA values, the observed correlations for the SCA predictions of the 49 validation hybrids, indicate that predicting SCA values by training on this set of unbalanced

phenotypic data, is well beyond the capabilities of both  $\epsilon$ -SVR and BLP.

**Test-cross and new-cross sampling** Table 3 gives an overview of the BLP and  $\epsilon$ -SVR prediction accuracies when the training set is reduced in a non-random fashion to simulate predictions on hybrids for which one or both parental inbred lines are new and, therefore, untested. Squared Pearson correlation coefficients between the entries in vector  $\hat{y}_T^{rg}$  and their SSR or AFLP-based cross-validation predictions are presented for both sampling schemes as well as the squared correlations between the entries of the validation set vectors  $\hat{y}_V^{rg}$  and their predictions.

**Table 3** Prediction accuracies, expressed as squared Pearson correlation coefficients, obtained from two sampling schemes simulating predictions on hybrids where one (test-cross sampling) or both parents (new-cross sampling) are newly developed inbred lines

Predictor	Trait	Predictand	Test-cross sampling		New-cross sampling	
			$\varepsilon$ -SVR	BLP	$\varepsilon$ -SVR	BLP
AFLP markers	Yield	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.72	0.78	0.48	0.58
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.09	0.10	0.09	0.11
	Moist.	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.80	0.85	0.63	0.71
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.53	0.67	0.31	0.58
	Flower	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.80	0.84	0.62	0.69
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.30	0.43	0.04	0.22
SSR markers	Yield	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.62	0.66	0.32	0.39
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.10	0.05	0.07	0.03
	Moist.	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.77	0.72	0.57	0.51
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.41	0.38	0.15	0.14
	Flower	Cross ( $\hat{y}_T^{\text{FG}}$ )	0.67	0.70	0.41	0.45
		Valid ( $\hat{y}_V^{\text{FG}}$ )	0.31	0.41	0.02	0.18

Cross-validation correlations on the vector  $\hat{y}_T^{\text{FG}}$  (cross) as well as correlations for predictions of the validation vector  $\hat{y}_V^{\text{FG}}$  (valid) are presented for the three traits grain yield, grain moisture content and days until flowering

### Molecular marker information content

The sensitivity of both  $\varepsilon$ -SVR and BLP to a reduction in the size of the molecular fingerprint is shown in Fig. 2 by means of box and whisker plots. The set of SSR and AFLP markers is reduced in steps of 10%. For each step, a random subset of markers is selected and used to construct an  $\varepsilon$ -SVR and BLP prediction model by training on all entries of the vector  $\hat{y}_T^{\text{FG}}$  minus the 38 hybrids that are tested in the validation set. Prediction accuracy is expressed as squared Pearson correlation coefficients between the predictions of the 49 validation hybrids and their corresponding entries in the vector  $\hat{y}_V^{\text{FG}}$ .

## Discussion

### Unbalanced data handling

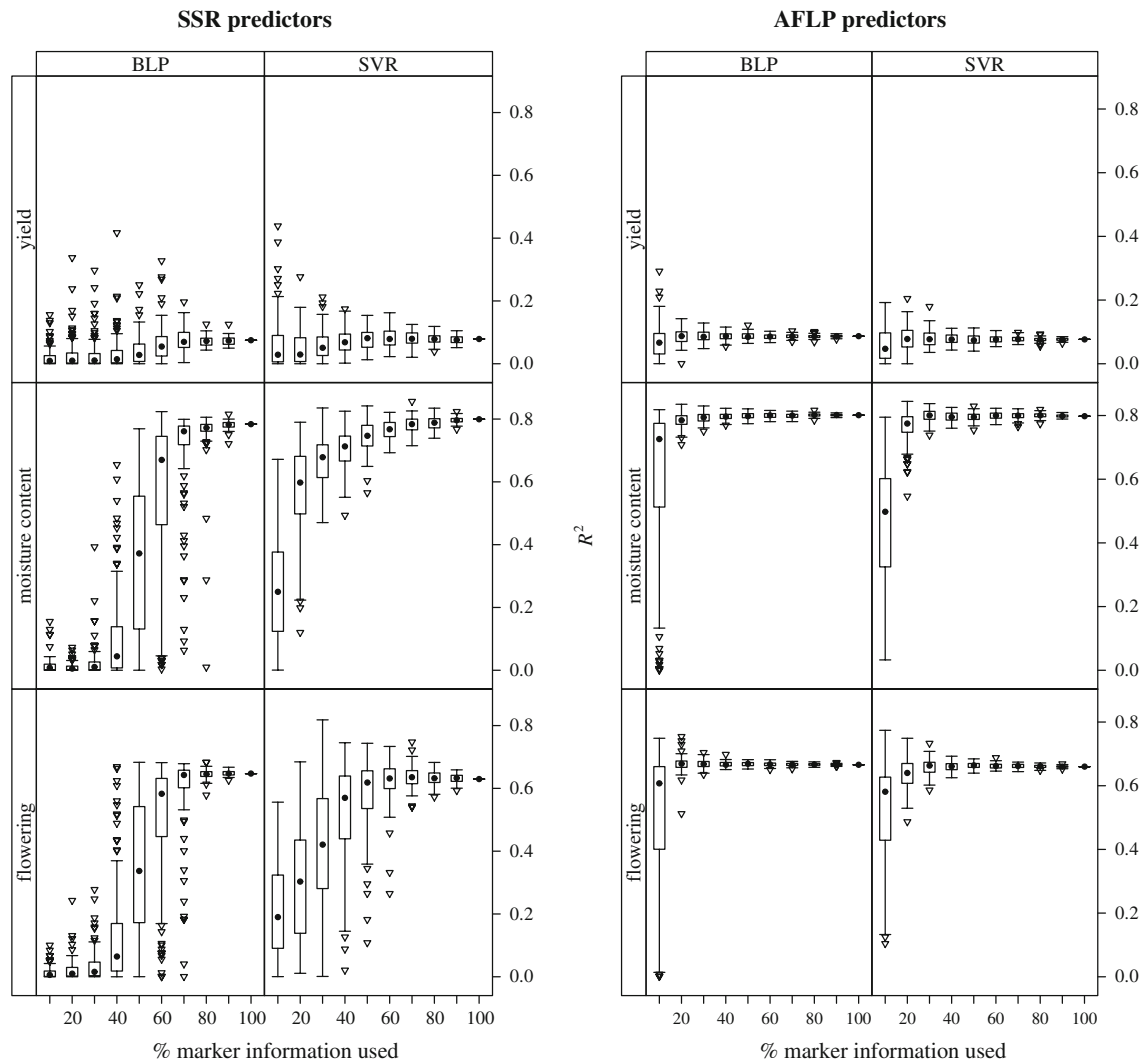
Predicting the phenotypic performance of untested hybrids by means of an  $\varepsilon$ -SVR or BLP model requires a training set of considerable size. Each training example should be represented by a single response value and a set of molecular marker-based predictors. A typical commercial hybrid breeding programme tests hundreds of new inbred combinations in a vast number of multi-location field trials on a yearly basis. The resulting data sets contain phenotypic measurements on numerous hybrids and would, therefore, allow for the construction of an  $\varepsilon$ -SVR or BLP prediction model at a low cost. However, the unbalanced nature of this kind of breeding data makes it hard to distill a single response value that allows to rank all hybrids on the

same scale. We examined three mixed model-based methods to obtain such a score from unbalanced phenotypic data: (1) random phenotypes introduced by Bernardo (1994, 1995, 1996a, b), (2) random genotypes described by Maenhout et al. (2008) and (3) fixed genotypes. Besides these three types of genotypic scores, we also obtain an estimate of the SCA value for each hybrid in the training data.

The random genotypes approach results in the highest correlations for all three traits under study. The fixed genotypes approach seems to result in the lowest correlations and Bernardo's random phenotypes perform only slightly better. The inadequacy of the fixed genotypes is not unexpected because the assumption of fixed genotypic effects is likely to increase the standard error of the estimators of commercially uninteresting hybrids, as these have few records in the data set and no strength can be borrowed from records on related hybrids. The assumption of random nuisance effects on the other hand seems justified for this kind of breeding data as the number of levels of these factors is usually quite high.

Comparing the prediction accuracies of the three traits under study, we see that grain moisture content is the most promising trait for the construction of a reliable prediction model. The large contribution of the main genotypic effects (i.e. GCA and SCA) to the total variance (74%) and the low impact of the GxE components (14.3%) in the linear mixed model analysis with a random genotype assumption, explains these results. For the number of days until flowering, this partition is 44.5 versus 20% which results in the somewhat lowered correlations observed for this trait. The trait grain yield, although of great interest to breeders,





**Fig. 2**  $\epsilon$ -SVR and BLP prediction accuracies obtained by constructing  $\epsilon$ -SVR and BLP prediction models from the 2316 entries in vector  $\hat{y}_T^{rg}$  using subsets of the AFLP or SSR marker information as predictors for each of the three traits under study. Box and whisker

plots show the range of squared Pearson correlation coefficients between the 49 entries in vector  $\hat{y}_V^{rg}$  and their predictions over 100 iterations of the marker sampling routine

looks the least eligible candidate for the construction of a prediction model. This low correspondence between training and validation data estimates can be explained by the fact that the contribution of the  $G \times E$  factors (38%) exceeds the contribution of the main genotypic factors (30.7%) to the total variance. The training examples are constructed marginal to the environmental factors such as growing season and geographical region while the validation data were collected at exactly one specific level of these factors. If a trait is subject to a large  $G \times E$  variance, one can expect a genotypic effect, estimated over a large range of environments, to deviate substantially from an estimate obtained at one particular level of these environmental factors. A similar reasoning can explain the observed lack of correlation for the SCA effects although other aspects like the increased prediction error variance of

the SCA BLUPs, the limited predictive value of a set of random SSR or AFLP markers with respect to a complex phenomenon like heterosis, and possibly reciprocal differences, also have their detrimental influence.

For the two promising traits moisture content and days until flowering the actual prediction accuracies obtained by  $\epsilon$ -SVR and BLP models, when trained on the vectors of random genotypes, are quite close to the theoretical upper bounds presented in Table 2. This can be seen from the SVR\_valid and BLP\_valid lines in Fig. 1 at  $p = 0$ . These specific points are obtained by correlating the  $\epsilon$ -SVR and BLP predictions of the 49 validation hybrids with their random genotypic estimates in vector  $\hat{y}_V^{rg}$ .

Our results indicate that the random genotypes approach is the best way to obtain a single genotypic score for each hybrid in the training data. By contrast, Bernardo (1994,

1995, 1996a, b) makes predictions on new hybrids by fitting the vector of random phenotypes  $\hat{y}_T^{rp}$  in Eq. 5. The entries in the resulting vector  $\hat{y}_p$  are sensu stricto not BLPs as the procedure does not take into account the covariance structure that originated from the measurement adjustments involving estimated fixed effects. This observation seems of minor importance as cross-validation results indicate a superior prediction accuracy compared with several other methods (Charcosset et al. 1998). However, a more straightforward approach would be to simply fit a number of additional parameters for the missing GCA and SCA components of the untested hybrids into the variance structure of the linear mixed model. As there are no phenotypic measurements linked to these effects, the additional columns in the random design matrix can all be set to zero. The estimated values for these additional effects are true best linear unbiased predictions or BLUPs and allow to reconstruct the predicted genotypic value of an untested hybrid by means of Eq. 2. The downside of this approach is that for each new prediction, the full set of mixed model equations needs to be solved. Moreover, an assessment of prediction accuracy by means of cross-validation routines is not only computationally exhausting, but often just not sensible as leaving out the phenotypic measurements on one or more hybrids might divide the training data in two or more disconnected subsets. In this scenario, each of the disconnected subsets contains measurements on a different, non-overlapping set of hybrids which are tested in a different set of environments. Contrasts involving random genotypic effects of hybrids that belong to different, disconnected data subsets are usually estimable but do not conform to the usual interpretation as they rely on the implicit assumption that the genetic levels among the different environmental subsets are equal (Laloë 1993). To avoid these pitfalls, a BLP prediction based on the random genotypic scores of the training hybrids is the next best option.

#### Reduction of the training data

##### *Training sample size*

In the previous section we indicated that using random genotypes to train our  $\varepsilon$ -SVR and BLP prediction models should result in superior prediction accuracies compared with the alternatives examined. For this reason, we continue to work with the random genotypes to evaluate the impact of the training sample size on the prediction accuracy of both  $\varepsilon$ -SVR and BLP.

*Random sampling* In Fig. 1 we see that the behaviour of  $\varepsilon$ -SVR is quite similar to that of BLP when the size of the training set is reduced in a random fashion. For both

methods, it is very clear that the cross-validation-based prediction accuracies consistently overestimate their validation trial counterparts. This is more explicit for the low heritability trait grain yield than for the traits moisture content and days until flowering. The observed disparity can be explained by the specific set of  $G \times E$  effects that affect the validation data while the estimates derived from the training data are marginal to all environmental effects. If  $G \times E$  effects explain a large portion of the observed variance for a trait, the observed heritability will be reduced correspondingly, as is the case for grain yield.

If we focus on the prediction of total genotypic value, the accuracy of  $\varepsilon$ -SVR and BLP shares a similar downward trend when the size training set is reduced, although  $\varepsilon$ -SVR usually performs slightly worse than BLP. The fall in prediction accuracy starts somewhere between  $p = 2$  and  $p = 3$ , which is the equivalent of using 25 and 12.5% of the original training data, respectively. If the training set is further reduced, the sampling variance of the validation trial-based prediction accuracies increases, as indicated by the widening of the interquartile ranges. This increase in sampling error is less pronounced for the cross-validation-based prediction accuracies, giving a false indication of confidence for these favourable estimates. For the three traits under study, there is little difference between the behaviour of prediction models based on SSR markers and those using AFLP markers as predictors when the set of training hybrids is reduced by random selection.

If we focus on the prediction of SCA, we see that neither  $\varepsilon$ -SVR or BLP succeed in raising the median validation prediction accuracy, expressed as a squared Pearson correlation coefficient, above 0.13. Most striking is that the prediction accuracy estimates obtained through cross-validation give the impression that both  $\varepsilon$ -SVR and BLP are quite capable of making SCA predictions with a reasonable accuracy, especially if the full training set is used. The more pronounced impact of  $G \times E$  effects on SCA measurements is again the most likely culprit here.

*Test-cross and new-cross sampling* If a non-random selection of training hybrids is performed, the superiority of the AFLP predictors becomes apparent, as can be seen from Table 3. In all but two scenarios, the prediction models based on AFLP markers have a greater prediction accuracy compared with those based on SSR markers. Table 3 again demonstrates the upward bias of the cross-validation-based prediction accuracy estimates. The  $\varepsilon$ -SVR prediction models are generally inferior to BLP when it comes to predicting the phenotypic performance of hybrids for which at least one of the parental inbred lines has no offspring in the training set. If both parents are unknown, neither  $\varepsilon$ -SVR nor BLP succeeds in making reliable predictions as the highest validation trial prediction accuracy

is 0.58 for a BLP model trained on the trait grain moisture content using AFLP markers as predictors. The combination of a high heritability for grain moisture content and the more informative AFLP markers as predictors should allow this BLP model to be used for screening purposes (Maenhout et al. 2008).

#### *Molecular marker information content*

Reducing the set of predictors, by randomly selecting a subset of markers, has a negative effect on the prediction accuracy of both  $\varepsilon$ -SVR and BLP as can be deduced from Fig. 2. The effect of the number of genotyped markers on the prediction accuracy appears to be subject to the law of diminishing marginal returns and little improvement is to be expected by further saturating the molecular fingerprint with additional AFLP or SSR markers. In this respect, Frisch et al. (2009) even observe a decline in prediction accuracy when the number of genes for which expression data are incorporated in their transcriptome-based prediction models, is increased beyond a certain optimum.

The difference in behaviour between  $\varepsilon$ -SVR and BLP is most apparent for the traits grain moisture content and days until flowering in combination with the less informative SSR markers as predictors. As soon as 30% of the SSR markers are removed from the set of predictors, certain samples generate a substantially lower prediction accuracy of the BLP model while the accuracy of the equivalent  $\varepsilon$ -SVR model is nearly identical to that of the full marker set. Reducing the set of SSR predictors beyond this level, further inflates the sampling error of the BLP prediction accuracies, while at the same time the median of the distribution starts its steep descent.  $\varepsilon$ -SVR handles a reduction of the SSR predictors better than BLP as the sampling error starts to increase at lower values of the fingerprint size, while the median of the prediction accuracy shows a gentle decline as the number of predictors is reduced. The median  $\varepsilon$ -SVR prediction accuracy is for instance always superior to that of BLP as soon as 40% of the markers is removed. This observed superiority of  $\varepsilon$ -SVR over BLP is less pronounced if we use the AFLP markers as predictors. Both methods retain a good and comparable prediction accuracy for the traits grain moisture content and days until flowering, even when the set of AFLP predictors is reduced to 20% of its original size. Beyond this level, the prediction accuracy rapidly declines, while the sample variance increases. Even if 90% of the AFLP markers are removed, which is equivalent to a predictor set size of 57 dominant markers, several samples allow  $\varepsilon$ -SVR and BLP to obtain good prediction accuracies. Moreover, several samples of AFLP and SSR markers result in prediction accuracies that are greater than that of the equivalent model using the full set of markers. These observations indicate that an  $\varepsilon$ -SVR

or BLP prediction model that uses only a specific subset of markers, might possibly improve the presented prediction accuracies but further study is needed to ascertain this point.

#### **Conclusions**

To construct an  $\varepsilon$ -SVR or BLP model for the prediction of phenotypic response based on a hybrid's molecular fingerprint, training data that contains a vector of marker scores and a single response value for every hybrid is needed. The best prediction accuracy is achieved by constructing these hybrid response values by summing the appropriate GCA and SCA BLUPs, obtained from a linear mixed model analysis with a random genotypic effects assumption.

If prediction accuracy is determined by means of a validation trial, both  $\varepsilon$ -SVR and BLP perform close to the theoretical limit for the traits grain moisture content and days until flowering while they both fall short for grain yield, a trait with a low heritability in advanced breeding pools. The accuracy of SCA predictions is similarly insufficient for all three traits. This lack of predictive power is not reflected in the prediction accuracy measures obtained through cross-validation procedures, as these do not take into account the uncertainty introduced by  $G \times E$  effects. Furthermore, if only a limited set of training examples is available but the genotyped markers are either numerous or very informative, BLP is more accurate than  $\varepsilon$ -SVR. If on the other hand the set of molecular markers is either restricted in size or information content,  $\varepsilon$ -SVR is the preferred prediction method.

**Acknowledgments** The authors would like to thank the people from RAGT R2n for their unreserved and open-minded scientific contribution to this research. We also gratefully acknowledge the helpful comments and suggestions of two anonymous referees.

#### **Appendix: Variance structure of the linear mixed models fitted to the phenotypic data of the validation field trial**

The four random vectors  $\mathbf{c}$ ,  $\mathbf{a}_s$ ,  $\mathbf{a}_o$  and  $\mathbf{d}$  of Eq. 4 are assumed to be mutually independent. Furthermore, for each of these vectors  $\mathbf{h} \in \{\mathbf{c}, \mathbf{a}_s, \mathbf{a}_o, \mathbf{d}\}$  we assume that the variance has the separable form

$$\text{Var}(\mathbf{h}) = \mathbf{g}_e \otimes \mathbf{g}_v, \quad (7)$$

where  $\otimes$  denotes the Kronecker product.  $\mathbf{g}_e$  represents a  $3 \times 3$  symmetric matrix containing the covariance between environments while  $\mathbf{g}_v$  represents the covariance between the specified genetic components of the validation

trial entries. We start by fitting a completely unstructured variance matrix for  $\mathbf{g}_e$  while assuming an identity matrix for  $\mathbf{g}_v$ . In subsequent steps, the number of REML estimated variance components is reduced by fitting more parsimonious variance models for  $\mathbf{g}_e$  using restricted maximum likelihood ratio tests in case of comparisons between nested models, or Akaike's information criterion (AIC) otherwise. We attempt to fit a first-order factor analytic variance model such that  $\mathbf{g}_e = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}$  where  $\boldsymbol{\lambda}$  is a vector of factor loadings and the matrix  $\boldsymbol{\Psi}$  is a diagonal matrix containing three location-specific variances (Smith et al. 2001). To obtain a more parsimonious model, the specific variances were sometimes made equal or zero (giving perfect correlation), and/or the loadings made equal (giving a common covariance (Cullis et al. 1998)). In a subsequent reduction, the variances on the diagonal are set equal which results in a compound symmetry model. The simplest model for  $\mathbf{g}_e$  assumed zero covariance and equal variances.

Once the most parsimonious model for  $\mathbf{g}_e$  is determined, we try different formulations for  $\mathbf{g}_v$ . We fit an identity matrix for the variance model of the six check varieties in vector  $\mathbf{c}$  as no molecular marker or pedigree information is available for these varieties. For the vectors  $\mathbf{a}_s$  and  $\mathbf{a}_o$ , containing the GCA effects of the inbred lines, we try to fit the different coefficient of coancestry derived matrices  $\mathbf{a}$  described by Maenhout et al. (2009) or an identity matrix. In a similar way, we compare the different coefficient of fraternity-based matrices  $\mathbf{d}$  for the variance matrix  $\mathbf{g}_v$  pertaining to the vector  $\mathbf{d}$ . Sometimes, the most parsimonious model is obtained by not using the separable form of Eq. 7 but directly fitting a common GCA or SCA effect for all three locations.

**Table 4** Summary of the variance structures fitted on the measurements of the validation data set for the traits grain yield, grain moisture content and days until flowering

Component	Yield	Moisture content	Flowering
Var( $\mathbf{c}$ )	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \mathbf{I}_6$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \mathbf{I}_6$	CS
Var( $\mathbf{a}_s$ )	CS	CS	$(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \otimes \mathbf{I}_7$
Var( $\mathbf{a}_o$ )	$\mathbf{I}_3 \otimes \mathbf{a}_o$	$\mathbf{a}_o$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \mathbf{I}_7$
Var( $\mathbf{d}$ )	$\mathbf{d}$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}) \otimes \mathbf{d}$	$\mathbf{d}$
Var( $\mathbf{e}_1$ )	$\mathbf{I}_5 \otimes \text{AR1}$	$\mathbf{I}_5 \otimes \mathbf{I}_{11}$	$\mathbf{I}_5 \otimes \mathbf{I}_{11}$
Var( $\mathbf{e}_2$ )	$\mathbf{I}_5 \otimes \text{AR1}$	$\text{AR1} \otimes \mathbf{I}_{11}$	$\text{AR1} \otimes \mathbf{I}_{11}$
Var( $\mathbf{e}_3$ )	$\mathbf{I}_5 \otimes \text{AR1}$	$\mathbf{I}_5 \otimes \text{AR1}$	$\text{AR1} \otimes \mathbf{I}_{11}$

$\boldsymbol{\lambda}$ , loadings of a first order factor analytic covariance model;  $\mathbf{I}_l$ , identity matrix of size  $l$ ; CS, compound symmetry, common genetic covariance over locations;  $\boldsymbol{\Psi}$ , site specific variances of a factor analytic covariance model; AR1, first order autoregressive covariance;  $\mathbf{a}_o$ , matrix containing coancestry coefficients of Iodent lines according to Eq. 1. In case  $\mathbf{a}_o$  is not part of a Kronecker product, a common Iodent GCA effect for all three locations was fitted;  $\mathbf{d}$ , matrix containing fraternity coefficients for the 49 hybrids according to Cockerham (1966). In case  $\mathbf{d}$  is not part of a Kronecker product, a common SCA effect for all three locations was fitted

The variance of each vector of residuals  $\mathbf{e}_i$  that make up vector  $\mathbf{e}$  in Eq. 3 is modeled as a separable process in the direction of rows and columns so we can write  $\text{Var}(\mathbf{e}_i) = \boldsymbol{\Sigma}_{ic} \otimes \boldsymbol{\Sigma}_{ir}$  where  $\otimes$  denotes the Kronecker product. The matrices  $\boldsymbol{\Sigma}_{ic}$  and  $\boldsymbol{\Sigma}_{ir}$  are either identity matrices or contain first order autoregressive correlations to account for spatial variation as described in Gilmour et al. (1997), Smith et al. (2001) and Oakey et al. (2007). Table 4 gives an overview of the final model for the variance structure of vectors  $\mathbf{g}$  and  $\mathbf{e}$  for each trait.

## References

- Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. *Theor Appl Genet* 85:1055–1062
- Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci* 34:20–25
- Bernardo R (1995) Genetic models for predicting maize single-cross performance in unbalanced yield trial data. *Crop Sci* 35:141–147
- Bernardo R (1996a) Best linear unbiased prediction of the performance of crosses between untested maize inbreds. *Crop Sci* 36:50–56
- Bernardo R (1996b) Best linear unbiased prediction of maize single-cross performance. *Crop Sci* 36:872–876
- Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci* 48:1649–1664
- Burges C (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2:121–167
- Charcosset A, Bonnissseau B, Touchebeuf O, Burstin J, Dubreuil P, Barrière Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. *Crop Sci* 38:38–44
- Cullis B, Gogel B, Verbyla A, Thompson R (1998) Spatial analysis of multi-environment early generation trials. *Biometrics* 54:1–18
- Frisch M, Thiemann A, Fu J, Schrag TA, Scholten S, Melchinger AE (2009) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theor Appl Genet* (in press)
- Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *J Agric Biol Environ Stat* 2:269–293
- Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *J Global Optim* 13:455–492
- Laloë D, (1993) Precision and information in linear models of genetic evaluation. *Genet Sel Evol* 25:557–576
- Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. *Theor Appl Genet* 115:1003–1013
- Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2008) Marker-based screening of maize inbred lines using support vector machine regression. *Euphytica* 161:123–131
- Maenhout S, De Baets B, Haesaert G (2009) Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theor Appl Genet* 118:1181–1192
- Oakey H, Verbyla AP, Cullis BR, Wei X, Pitchford WS (2007) Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theor Appl Genet* 114:1319–1332

- Schrag TA, Maurer HP, Melchinger AE, Piepho HP, Peleman J, Frisch M (2007) Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theor Appl Genet* 114:1345–1355
- Schrag TA, Möhring J, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sorensen AP, Frisch M (2009) Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theor Appl Genet* 118:741–751
- Schrag TA, Möhring J, Kusterer B, Dhillon BS, Melchinger AE, Piepho HP, Frisch M (2009) Hybrid performance prediction in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theor Appl Genet* (in press)
- Smith A, Cullis B, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147
- Smola A, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14:199–222
- Stuber C, Cockerham C (1966) Gene effects and variances in hybrid populations. *Genetics* 54:1279–1286
- Vapnik V (1995) *The nature of statistical learning theory*. Springer, New York
- Welham SJ, Cullis BR, Gogel BJ, Gilmour AR, Thompson R (2004) Prediction in linear mixed models. *Aust NZ J Stat* 46:325–347