Heredity (2018) 120:283–295
https://doi.org/10.1038/s41437-017-0027-0

the geneticssociety

ARTICLE

Efficiency of genomic prediction of non-assessed single crosses

José Marcelo Soriano Viana[1] · Helcio Duarte Pereira[1] · Gabriel Borges Mundim[2] · Hans-Peter Piepho[3] · Fabyano Fonseca e Silva[4]

Received: 28 July 2017 / Revised: 26 October 2017 / Accepted: 30 October 2017 / Published online: 28 November 2017

Abstract

An important application of genomic selection in plant breeding is predicting untested single crosses (SCs). Most investigations on the prediction efficiency were based on tested SCs using cross-validation. The main objective was to assess the prediction efficiency by correlating the predicted and true genotypic values of untested SCs (accuracy) and measuring the efficacy of identification of the best 300 untested SCs (coincidence) using simulated data. We assumed 10,000 SNPs, 400 QTLs, two groups of 70 selected DH lines, and 4900 SCs. The heritabilities for the assessed SCs were 30, 60, and 100%. The scenarios included three sampling processes of DH lines, two sampling processes of SCs for testing, two SNP densities, DH lines from distinct and the same populations, DH lines from populations with lower LD, two genetic models, three statistical models, and three statistical approaches. We derived a model for genomic prediction based on SNP average effects of substitution and dominance deviations. The prediction accuracy is not affected by the linkage phase. The prediction of untested SCs is very efficient. The accuracies and coincidences ranged from ~0.8 and 0.5 at low heritability to 0.9 and 0.7 at high heritability, respectively. We also highlight the relevance of the overall LD and demonstrate that efficient prediction of untested SCs can be achieved for crops that show no heterotic pattern, for reduced training set size (10%), for SNP density of 1 cM, and for distinct sampling processes of DH lines based on random choice of the SCs for testing.

Introduction

Genomic selection is commonly used in animal breeding programs, especially for dairy cattle (Van Eenennaam et al. 2014). The main reasons for the effective application of genomic selection in livestock breeding are that it is efficient, that is, the process has high prediction accuracy, the cost of phenotyping (mainly progeny tests) is higher than the cost of genotyping, and the process significantly shortens the selection cycle (Meuwissen et al. 2013). An important application of genomic selection in plant breeding is the prediction of untested single crosses (genotypic value prediction) and testcrosses (prediction of general combining ability effect) in hybrid breeding (Zhao et al. 2015). Genomic prediction of two-way and three-way crosses has been investigated (Philipp et al. 2016). Bernardo (1994) pioneered the prediction of untested single crosses based on best linear unbiased prediction (BLUP). Many significant studies on the prediction of untested single cross and testcross performance have been published in the last 23 years with a focus on the assessment of prediction accuracy. Most investigations were based on empirical data and estimated the prediction accuracy using a cross-validation procedure. Very few were based on simulated data (Li et al. 2017; Technow et al. 2012). With no exception, the inference was that prediction of untested single crosses and testcrosses can be efficient, depending on the heritability, training set size, and number of tested inbreds in hybrid combination (both, one, and none parents tested). Remarkably, this conclusion was drawn from studies differing in molecular marker, marker density, number of inbreds, level of relatedness, diversity, and linkage disequilibrium (LD) between inbreds, heterotic pattern, training set size, genetic model, and statistical approach (Zhao et al. 2015). Efficient prediction of two-way and three-way crosses of barley has been achieved

✉ José Marcelo Soriano Viana
jmsviana@ufv.br

[1] Department of General Biology, Federal University of Viçosa, 36570-900 Viçosa, MG, Brazil

[2] Dow AgroSciences Seeds and Biotechnology Brazil Ltda, 38490-000 Indianópolis, MG, Brazil

[3] Institute of Crop Science, Biostatistics Unit, University of Hohenheim, 70599 Stuttgart, Germany

[4] Department of Animal Science, Federal University of Viçosa, 36570-900 Viçosa, MG, Brazil

using training and validation sets that include the same class of hybrids (Philipp et al. 2016).

Most studies on genomic prediction of maize single cross performance published since 2011 have used single nucleotide polymorphisms (SNPs), with the number of filtered SNPs ranging from 425 (Zhao et al. 2013a) to 39,627 (Technow et al. 2012). For grain yield, the relative prediction accuracies (computed as the accuracy divided by the root square of the heritability) in these studies ranged from 0.27 to 0.62 and from 0.65 to 0.95, respectively. The number of inbreds in each heterotic group was highly variable as well, ranging from 6 and 9 (Bernardo 1994) to 75 and 75 (Technow et al. 2012), respectively. The relative accuracy for grain yield observed by Bernardo (1994) ranged between 0.72 and 0.89. The level of relatedness between inbreds ranged from non-related inbreds in each group (Technow et al. 2012) to a maximum average value of 0.58 (RFLP-based coancestry coefficient) (Bernardo 1995). The relative accuracy obtained by Bernardo (1995) ranged from 0.41 to 0.80 for grain yield. The common heterotic groups were Stiff Stalk and non-Stiff Stalk (Kadam et al. 2016) or Dent and Flint (Technow et al. 2014). The relative accuracies for grain yield ranged from 0.28 to 0.77 and from 0.75 and 0.92, respectively. The study of Bernardo (1996a) involved nine heterotic groups and the relative accuracies for grain yield ranged from 0.43 to 0.88. These results evidence that prediction accuracy is proportional to the molecular marker density and that high accuracy can be achieved regardless of number of inbreds, level of relatedness, and number of heterotic groups. No study provided distinctly higher prediction accuracy of the additive-dominance model relative to the additive model. Finally, with only testcrosses the genomic BLUP (GBLUP) approach outperformed pedigree-based BLUP (Albrecht et al. 2014; Albrecht et al. 2011) in regard to prediction accuracy.

Genomic prediction of single crosses has been carried out based on tested single crosses using cross-validation. Thus, the estimated prediction accuracies are not for untested single crosses. Consequently, none of the previous studies on the efficiency of genomic prediction of single cross performance measured the efficacy of identification of the best untested single crosses. Our main objective was to assess the prediction efficiency of untested single crosses by correlating the predicted and true genotypic values of untested single crosses (prediction accuracy) and measuring the efficacy of identification of the best 300 untested single crosses (coincidence index) using a large simulated data set. The secondary objectives were to highlight that the prediction accuracy primarily depends on the overall LD in the groups of selected doubled haploid (DH) lines, that the prediction efficiency with no heterotic pattern can be as high as the prediction efficiency involving heterotic groups, and that the choice of single crosses for testing should be random instead of selecting DH lines for a diallel to maximize the prediction efficiency. Further, we derived a model for genomic prediction of untested single crosses based on the SNP average effects of substitution and dominance deviations.

Materials and methods

Theory

Most important papers on genomic selection offer deep statistical aspects on the whole-regression models, extending to SNP effects a previously derived gene model. Some important papers include only basic quantitative genetics theory based on linkage equilibrium. The quantitative genetics theory developed in this paper provides a genetic model for genomic prediction of untested single crosses that accounts for the LD between QTLs and SNPs. The model offers the genetic background to the models fitted in previous papers on the prediction of untested single crosses and testcrosses (Albrecht et al. 2011; Massman et al. 2013; Technow et al. 2012). The theory is comprehensive, i.e., it is adequate for DH and inbred lines, for predicting untested single crosses and testcrosses, and for crops with and without defined heterotic groups, and it is easily extended to genomic prediction of two-way and three-way crosses (relevant for rice, wheat, and barley breeders), based on Jenkins (1934). The theoretical accuracy can be used in future investigations on the efficiency of genomic prediction of untested single crosses based on a deterministic approach, as in the study of Grattapaglia and Resende (2010).

LD in a group of selected DH or inbred lines

Consider a group of DH or inbred lines selected from a population or heterotic group. Assume also a QTL (alleles B/b) and a SNP (alleles C/c) where B and b are the alleles that increase and decrease the trait expression, respectively. Define the joint genotype probabilities as $P(BBCC) = f_{22}$, $P(BBcc) = f_{20}$, $P(bbCC) = f_{02}$, and $P(bbcc) = f_{00}$, where the subscripts indicate the numbers of copies of the major allele (B and C). The measure of LD between the QTL and the SNP is $\Delta_{bc} = f_{22}f_{00} - f_{20}f_{02}$ (Kempthorne 1957) and the haplotype frequencies are $P(BC) = f_{22} = p_b p_c + \Delta_{bc}$, $P(Bc) = f_{20} = p_b q_c - \Delta_{bc}$, $P(bC) = f_{02} = q_b p_c - \Delta_{bc}$, and $P(bc) = f_{00} = q_b q_c + \Delta_{bc}$, where p is the frequency of the major allele (B or C) and $q = 1 - p$ is the frequency of the minor allele (b or c). Note that $p_b = f_{22} + f_{20}$ and $p_c = f_{22} + f_{02}$. It is important to highlight that we are not assuming that the QTL and the SNP are linked and in LD in

the population or heterotic group, because this is not a necessary condition for genomic prediction. But we are assuming that they are in LD in the group of DH or inbred lines. Furthermore, because of selection, genetic drift, and inbreeding (only for inbreds and linked QTLs and SNPs), the gene and genotypic frequencies and the LD values concerning the selected DH or inbred lines cannot be traced to the values in the population or heterotic group.

SNP genotypic values of DH or inbred lines

The average genotypic value for a group of selected DH or inbred lines is $M_{IL} = m_b + (p_b - q_b)\,a_b$, where m_b is the mean of the genotypic values of the homozygotes and a_b is the deviation between the genotypic value of the homozygote of higher expression and m_b. Thus, the average SNP genotypic values for the DH or inbred lines CC and cc are

$$\begin{aligned} G_{CC} &= \frac{1}{f_{.2}}[f_{22}(m_b + a_b) + f_{02}(m_b - a_b)] \\ &= M_{IL} + 2q_c\alpha_{SNP} = M_{IL} + A_{CC} \end{aligned}$$

$$\begin{aligned} G_{cc} &= \frac{1}{f_{.0}}[f_{20}(m_b + a_b) + f_{00}(m_b - a_b)] \\ &= M_{IL} - 2p_c\alpha_{SNP} = M_{IL} + A_{cc} \end{aligned}$$

where $\alpha_{SNP} = \left[\frac{\Delta_{bc}}{p_c q_c}\right] a_b = \kappa_{bc} a_b$ is the average effect of a SNP substitution in the group of DH or inbred lines, and A is the SNP additive value for a DH or inbred line. Note that $E(A) = 0$.

Assuming two QTLs (alleles B and b and E and e) in LD with the SNP, the average effect of a SNP substitution in the selected DH or inbred lines is $\alpha_{SNP} = \kappa_{bc} a_b + \kappa_{ce} a_e$, where $\kappa_{ce} = \left[\frac{\Delta_{ce}}{p_c q_c}\right]$. Thus, the average effect of a SNP substitution (and the SNP additive value) is proportional to the LD measure and to the deviation a for each QTL that is in LD with the marker.

SNP genotypic values of single crosses

To maximize the heterosis, maize breeders commonly assess single crosses originating from selected DH or inbred lines from distinct heterotic groups. Consider n_1 DH or inbred lines from a population or heterotic group and n_2 DH or inbred lines from a distinct population or heterotic group. The average genotypic value for the single crosses derived by crossing the DH or inbred lines from group 1 with the DH or inbred lines from group 2 is

$$M_H = m_b + (p_{b1}p_{b2} - q_{b1}q_{b2})\,a_b + (p_{b1}q_{b2} + q_{b1}p_{b2})\,d_b$$

where d_b is the dominance deviation (the deviation between the genotypic value of the heterozygote and m_b).

The average genotypic values for the single crosses derived from DH or inbred lines CC and cc of group 1 are

$$\begin{aligned} G_{CC1} &= M_H + q_{c1}\kappa_{bc1}[a_b + (q_{b2} - p_{b2})d_b] \\ &= M_H + q_{c1}\kappa_{bc1}\alpha_{b2} = M_H + q_{c1}\alpha_{SNP1} = M_H + GCA_{CC1} \end{aligned}$$

$$\begin{aligned} G_{cc1} &= M_H - p_{c1}\kappa_{bc1}\alpha_{b2} = M_H - p_{c1}\alpha_{SNP1} \\ &= M_H + GCA_{cc1} \end{aligned}$$

where α_{b2} is the average effect of allelic substitution in the population derived by random crosses between the DH or inbred lines from group 2, α_{SNP1} is the SNP effect of allelic substitution in the hybrid population relative to a SNP derived from group 1, and GCA is the general combining ability effect for a SNP locus. Note that α_{SNP1} depends on the LD in group 1 ($\kappa_{bc1} = \Delta_{bc1}/p_{c1}q_{c1}$) and the average effect of allelic substitution in the population derived by random crosses between the DH or inbred lines from group 2. Furthermore, $E(GCA) = p_{c1}GCA_{CC1} + q_{c1}GCA_{cc1} = 0$. Concerning the single crosses derived from DH or inbred lines CC and cc of group 2, we have

$$\begin{aligned} G_{CC2} &= M_H + q_{c2}\kappa_{bc2}[a_b + (q_{b1} - p_{b1})d_b] \\ &= M_H + q_{c2}\kappa_{bc2}\alpha_{b1} = M_H + q_{c2}\alpha_{SNP2} = M_H + GCA_{CC2} \end{aligned}$$

$$G_{cc2} = M_H - p_{c2}\kappa_{bc2}\alpha_{b1} = M_H - p_{c2}\alpha_{SNP2} = M_H + GCA_{cc2}$$

Note that $E(GCA) = 0$. The average genotypic values for the single crosses concerning the SNP locus are

$$\begin{aligned} G_{CC1xCC2} &= M_H + q_{c1}\alpha_{SNP1} + q_{c2}\alpha_{SNP2} - 2q_{c1}q_{c2}\kappa_{bc1}\kappa_{bc2}d_b \\ &= M_H + GCA_{CC1} + GCA_{CC2} + SCA_{CC1xCC2} \end{aligned}$$

$$\begin{aligned} G_{cc1xcc2} &= M_H - p_{c1}\alpha_{SNP1} - p_{c2}\alpha_{SNP2} - 2p_{c1}p_{c2}\kappa_{bc1}\kappa_{bc2}d_b \\ &= M_H + GCA_{cc1} + GCA_{cc2} + SCA_{cc1xcc2} \end{aligned}$$

$$\begin{aligned} G_{CC1xcc2} &= M_H + q_{c1}\alpha_{SNP1} - p_{c2}\alpha_{SNP2} + 2q_{c1}p_{c2}\kappa_{bc1}\kappa_{bc2}d_b \\ &= M_H + GCA_{CC1} + GCA_{cc2} + SCA_{CC1xcc2} \end{aligned}$$

$$\begin{aligned} G_{cc1xCC2} &= M_H - p_{c1}\alpha_{SNP1} + q_{c2}\alpha_{SNP2} + 2p_{c1}q_{c2}\kappa_{bc1}\kappa_{bc2}d_b \\ &= M_H + GCA_{cc1} + GCA_{CC2} + SCA_{cc1xCC2} \end{aligned}$$

where $\kappa_{bc1}\kappa_{bc2}d_b = d_{SNP}$ is the SNP dominance deviation in the hybrid population, and SCA stands for the specific combining ability effect for a SNP locus. Note that $E(SCA) = p_{c1}p_{c2}SCA_{CC1xCC2} + p_{c1}q_{c2}SCA_{CC1xcc2} + q_{c1}p_{c2}SCA_{cc1xCC2} + q_{c1}q_{c2}SCA_{cc1xcc2} = 0$ and $E(SCA|CC) = E(SCA|cc) = 0$ for each group. That is, the expectation of the SNP SCA effects given a SNP genotype for the common DH or inbred line is also zero. Also note that the four genotypic values depend on four unknown parameters (M_H, α_{SNP1}, α_{SNP2}, and d_{SNP}).

Assuming two QTLs (alleles B and b and E and e) in LD with the SNP, the SNP dominance deviation is $d_{SNP} = \kappa_{bc1}\kappa_{bc2}d_b + \kappa_{ce1}\kappa_{ce2}d_e$. Thus, the SNP dominance deviation (and the SNP SCA effect) is proportional to the

product of the LD values in both groups of DH or inbred lines and to the dominance deviation for each QTL that is in LD with the marker.

The previous model expressed as a function of the SNP GCA and SCA effects was proposed by Massman et al. (2013), but the authors assumed $GCA_{CC} + GCA_{cc} = 0$ (for each heterotic group and for each SNP) and $SCA_{CC1xCC2} = SCA_{cc1xcc2} = -SCA_{CC1xcc2} = -SCA_{cc1xCC2}$. Technow et al. (2012) used a standard extension from QTL to SNP and defined the single cross genotypic value for a SNP as a function of the SNP deviations a and d. That is, $G = M_H + u_1 a + u_2 a + u_3 d$, where u_1 and u_2 are equal 1/2 or −1/2 if the corresponding DH or inbred line is homozygous for distinct SNP alleles (CC or cc), and u_3 equals 0 if the single cross is homozygous or 1 if heterozygous.

SNP genotypic values of single crosses from DH or inbred lines derived from the same population or heterotic group

Well-defined heterotic groups are known for maize but not for special maize such as popcorn and sweet corn, and for other crops such as wheat (Zhao et al. 2013b), rice (Xu et al. 2014), and barley (Philipp et al. 2016). Thus, for many breeders, it is interesting to know about the efficiency of genomic prediction of singles crosses when there are no heterotic groups. Assuming n DH or inbred lines derived from the same population or heterotic group, the average genotypic values for the single crosses concerning the SNP locus are

$$\begin{aligned} G_{CCxCC} &= M + 2q_c\alpha_{SNP} - 2q_c^2\kappa_{bc}^2 d_b \\ &= M + 2GCA_{CC} + SCA_{CCxCC} \end{aligned}$$

$$\begin{aligned} G_{ccxcc} &= M - 2p_c\alpha_{SNP} - 2p_c^2\kappa_{bc}^2 d_b \\ &= M + 2GCA_{cc} + SCA_{ccxcc} \end{aligned}$$

$$\begin{aligned} G_{CCxcc} &= M + 2(q_c - p_c)\,\alpha_{SNP} + 2p_c q_c \kappa_{bc}^2 d_b \\ &= M + GCA_{CC} + GCA_{cc} + SCA_{CCxcc}, \end{aligned}$$

where $M = m_b + (p_c - q_c)\,a_b + 2p_c q_c d_b$ is the hybrid population mean, $\alpha_{SNP} = \kappa_{bc}[a_b + (q_b - p_b)\,d_b] = \kappa_{bc}\alpha_b$ is the average effect of a SNP substitution in the hybrid population, and $d_{SNP} = \kappa_{bc}^2 d_b$ is the SNP dominance deviation. Note that the SNP GCA effects are equal to half the SNP additive value for the single crosses (A), the SNP SCA effects are the SNP dominance deviations for the single crosses (D), and the three genotypic values depend on three unknown parameters (M, α_{SNP}, and d_{SNP}). Also note that E(GCA) = E(A) = E(SCA) = E(SCA|CC) = E(SCA|cc) = E(D) = 0.

Accuracy of single cross genomic prediction

Assuming a QTL and a SNP in LD in the two groups of DH or inbred lines, the predictor of the single cross QTL genotypic value is the single cross SNP genotypic value (because they are proportional). Thus, the covariance between the predictor and the genotypic value is

$$\begin{aligned} &\mathrm{Cov}\left(\tilde{G}, G\right) = f_{22}^1 f_{22}^2 \\ &[M_H + GCA_{CC1} + GCA_{CC2} + SCA_{CC1xCC2}] \\ &[M_H + GCA_{BB1} + GCA_{BB2} + SCA_{BB1xBB2}] + \\ &+ f_{22}^1 f_{20}^2 [M_H + GCA_{CC1} + GCA_{cc2} + SCA_{CC1xcc2}] \\ &[M_H + GCA_{BB1} + GCA_{BB2} + SCA_{BB1xBB2}] + \\ &\cdots \\ &+ f_{00}^1 f_{00}^2 [M_H + GCA_{cc1} + GCA_{cc2} + SCA_{cc1xcc2}] \\ &[M_H + GCA_{bb1} + GCA_{bb2} + SCA_{bb1xbb2}] - (M_H)^2 \\ &= p_{c1}q_{c1}(\kappa_{bc1}\alpha_{b2})^2 + p_{c2}q_{c2}(\kappa_{bc2}\alpha_{b1})^2 + 4p_{c1}q_{c1}p_{c2}q_{c2}(\kappa_{bc1}\kappa_{bc2}d_b)^2 \\ &= p_{c1}q_{c1}(\alpha_{SNP1})^2 + p_{c2}q_{c2}(\alpha_{SNP2})^2 + 4p_{c1}q_{c1}p_{c2}q_{c2}(d_{SNP})^2 \\ &= \sigma^{2(1)}_{GCA_{SNP}} + \sigma^{2(2)}_{GCA_{SNP}} + \sigma^2_{SCA_{SNP}} = \sigma^2_{G(SNP)} \end{aligned}$$

where the GCA and SCA effects for the QTL are $GCA_{BB1} = q_{b1}\alpha_{b2}$, $GCA_{bb1} = -p_{b1}\alpha_{b2}$, $GCA_{BB2} = q_{b2}\alpha_{b1}$, $GCA_{bb2} = -p_{b2}\alpha_{b1}$, $SCA_{BB1xBB2} = -2q_{b1}q_{b2}d_b$, $SCA_{BB1xbb2} = 2q_{b1}p_{b2}d_b$, $SCA_{bb1xBB2} = 2p_{b1}q_{b2}d_b$, and $SCA_{bb1xbb2} = -2p_{b1}p_{b2}d_b$, σ^2_{GCA} and σ^2_{SCA} are the GCA and SCA variances for the SNP locus, and σ^2_G is the SNP genotypic variance. The GCA and SCA variances for the QTL are $\sigma^{2(1)}_{GCA} = p_{b1}q_{b1}(\alpha_{b2})^2$, $\sigma^{2(2)}_{GCA} = p_{b2}q_{b2}(\alpha_{b1})^2$, and $\sigma^2_{SCA} = 4p_{b1}q_{b1}p_{b2}q_{b2}(d_b)^2$. The QTL genotypic variance is $\sigma^2_G = \sigma^{2(1)}_{GCA} + \sigma^{2(2)}_{GCA} + \sigma^2_{SCA}$. Thus, the single cross prediction accuracy is

$$\rho_{\tilde{G},G} = \sqrt{\frac{\sigma^2_{G(SNP)}}{\sigma^2_G}}$$

Assuming s SNPs,

$$\rho_{\tilde{G},G} = \sum_{r=1}^{s} \sigma^2_{G(SNP(r))} \Big/ \sqrt{\sigma^2_{\tilde{G}}\sigma^2_G}$$

where $\sigma^2_{\tilde{G}}$ is the variance of the predicted single cross genotypic values, and σ^2_G is the single cross genotypic variance. Furthermore,

$$\alpha_{SNP(r)1} = \sum_{i=1}^{k'} \left[\frac{\Delta_{ri1}}{p_{r1}q_{r1}}\right]\alpha_{i2} = \sum_{i=1}^{k'} \kappa_{ri1}\alpha_{i2},$$

where k' is the number of QTLs in LD with the SNP r in group 1, and

$$d_{SNP(r)} = \sum_{i=1}^{k''} \left[\frac{\Delta_{ri1}}{p_{r1}q_{r1}}\right]\left[\frac{\Delta_{ri2}}{p_{r2}q_{r2}}\right] d_i = \sum_{i=1}^{k''} \kappa_{ri1}\kappa_{ri2}d_i$$

where k'' is the number of QTLs in LD with the SNP r in both groups.

Because the accuracy of genomic prediction of single crosses depends on the squares of the average effects of

SNP substitution and the SNP dominance deviations, it is not affected by the linkage phase (coupling or repulsion), as it does not depend on linkage. But it does depend on the magnitude of the LD in each group of DH or inbred lines.

Assuming single crosses derived from DH or inbred lines of a single population or heterotic group we have $\sigma^2_{G(SNP)} = 2p_c q_c(\alpha_{SNP})^2 + (2p_c q_c d_{SNP})^2$ and $\sigma^2_G = 2p_b q_b(\alpha_b)^2 + (2p_b q_b d_b)^2$.

The statistical model for single cross genomic prediction

Consider n_1 and n_2 (several tens) DH or inbred lines from two populations or heterotic groups genotyped for s (thousands) SNPs and the experimental assessment of h (a few hundred) single-crosses (h much lower than $n_1.n_2$) in e (several) environments (a combination of growing seasons, years, and locations). Defining y as the adjusted single cross phenotypic mean, the statistical model for prediction of the average effects of SNP substitution and the SNP dominance deviations is

$$y = M_H + \sum_{r=1}^{s} (z_{1_r}\alpha_{SNP1_r} + z_{2_r}\alpha_{SNP2_r} + z_{3_r}d_{SNP_r}) + \text{error}$$

where $z_{1_r} = q_{r1}$, $z_{2_r} = q_{r2}$, and $z_{3_r} = -2q_{r1}q_{r2}$ if the SNP genotypes for the DH or inbred lines are CC (group 1) and CC (group 2), $z_{1_r} = -p_{r1}$, $z_{2_r} = -p_{r2}$, and $z_{3_r} = -2p_{r1}p_{r2}$ if the SNP genotypes are cc (group 1) and cc (group 2), $z_{1_r} = q_{r1}$, $z_{2_r} = -p_{r2}$, and $z_{3_r} = 2q_{r1}p_{r2}$ if the SNP genotypes are CC (group 1) and cc (group 2), and $z_{1_r} = -p_{r1}$, $z_{2_r} = q_{r2}$, and $z_{3_r} = p_{r1}q_{r2}$ if the SNP genotypes are cc (group 1) and CC (group 2).

Regarding the single crosses obtained from DH or inbred lines of the same population or heterotic group, we have

$$y = M + \sum_{r=1}^{s} (z_{1_r}\alpha_{SNP_r} + z_{2_r}d_{SNP_r}) + \text{error}$$

where $z_{1_r} = 2q_r$ and $z_{2_r} = -2q_r^2$ if the SNP genotypes for the two crossed DH or inbred lines are CC and CC, $z_{1_r} = -2p_r$ and $z_{2_r} = -2p_r^2$ if the SNP genotypes are cc and cc, and $z_{1_r} = 2(q_r - p_r)$ and $z_{2_r} = 2p_r q_r$ if the SNP genotypes are CC and cc.

The statistical problem of genomic prediction when there is a very large number of molecular markers and relatively few observations has been addressed through several regularized whole-genome regression and prediction methods (Daetwyler et al. 2013; de Los Campos et al. 2013). Based on one of these approaches, the SNP average effects of substitution and SNP dominance deviations are predicted and used to provide genomic prediction of non-assessed single crosses. The predicted genotypic value for a non-assessed single cross of DH or inbred lines from two groups is

$$\tilde{G} = \hat{M}_H + \sum_{r=1}^{s} (z_{1_r}\tilde{\alpha}_{SNP1_r} + z_{2_r}\tilde{\alpha}_{SNP2_r} + z_{3_r}\tilde{d}_{SNP_r})$$

For a non-assessed single cross of DH or inbred lines from the same group, the predicted genotypic value is

$$\tilde{G} = \hat{M} + \sum_{r=1}^{s} (z_{1_r}\tilde{\alpha}_{SNP_r} + z_{2_r}\tilde{d}_{SNP_r})$$

Simulation

The SNP and QTL genotypes for DH lines, the QTL genotypes for single crosses, and the phenotypes for DH lines and single crosses were simulated using the software *REALbreeding* (available by request). The software does not assume a distribution for the LD values and gene effects, but computes the true LD values and gene effects based on quantitative genetics theory (Viana 2004). SNP and QTL allele frequencies follow a beta distribution. The parameters m, a, and d for each QTL are computed from the maximum and minimum genotypic values for homozygotes informed by the user. Based on our input, the software distributed 10,000 SNPs and 400 QTLs on ten chromosomes (1000 SNPs and 40 QTLs by chromosome). The average SNP density was 0.1 cM. The QTLs were distributed in the regions covered by the SNPs (~100 cM/chromosome).

The genotypic values of the DH lines and single crosses were generated assuming a single set of 400 QTLs and two degrees of dominance. To simulate grain yield and expansion volume (a measure of popcorn quality), we defined positive dominance ($0 <$ degree of dominance ≤ 1.2) and bidirectional dominance ($-1.2 \leq$ degree of dominance ≤ 1.2), respectively. For grain yield and expansion volume, the maximum and minimum genotypic values for homozygotes were 140 and 30 g/plant and 55 and 15 mL/g, respectively. The phenotypic values were obtained from the sum of the population mean, genotypic value, and experimental error. The error variance was computed from the broad sense heritability.

Initially, the software simulated 350 S_0 plants of the first heterotic group. The population was a second generation composite. In a composite, there is LD only for linked SNPs and QTLs (Viana et al. 2016). Then the software sampled one (scenario 1) or one to five (scenario 2) gametes from the S_0 plants, generating 350 DH lines. To generate 350 DH lines from S_3 plants, the software selfed S_0 plants for three generations using the single seed descent process. The number of DH lines per S_3 plant ranged from one to five (scenario 3). For each DH line sampling process, the software selected 70 DH lines, assuming a trait heritability of

30%. The same computational procedures provided the three groups of 70 selected DH lines from the second heterotic group (a second composite). For each DH line sampling process, the software crossed 70×70 DH lines to generate 4900 single crosses.

To investigate the efficiency of genomic prediction of untested single crosses when there is no heterotic group (relevant for rice, wheat, and barley breeders), the software also crossed 70 selected DH lines from the same heterotic group for generating 2415 single crosses (scenario 4). To highlight that the efficiency of genomic prediction of untested single crosses does not depend on the LD in the reference population, but on the LD in the groups of selected DH lines, the same computational procedures were used to derive 70 selected DH lines from the first and second heterotic groups after 10 generations of random crosses (to decrease the LD) (scenario 5).

Data files

The data for processing was obtained from 50 random samplings of 1470 (30%) and 490 (10%) of the single crosses to be assessed, assuming a trait heritability of 30, 60, and 100%. Thus, the genotypic value prediction accuracies of the assessed single crosses were 0.55, 0.77, and 1.00, respectively. With no exception, all DH lines from both heterotic groups were represented in the tested single crosses. Additionally, to assess the relevance of the number of DH lines sampled, we fixed the number of DH lines in each heterotic group to achieve approximately the same number of assessed single crosses using a diallel. That is, we sampled 38 and 22 DH lines in each heterotic group 50 times for a diallel (scenario 6), generating 1444 (30%) and 484 (10%) single crosses for assessment, respectively. In this case, only 54 and 31% of the DH lines are represented in the tested single crosses. We denoted these processes as sampling of single crosses and sampling of DH lines.

Assuming no heterotic groups, we proceeded to 50 random samplings of 724 (30%) and 241 (10%) of the single crosses from the same heterotic group for testing, also assuming a trait heritability of 30, 60, and 100%. With few exceptions when sampling 10% of the single crosses for testing, all DH lines from the heterotic group were represented in the assessed single crosses. The last scenario was genomic prediction of untested single crosses under an average density of one SNP for each cM. This lower density was obtained by random sampling of 100 SNPs per chromosome using a *REALbreeding* tool (*sampler*).

Statistical analysis

The methods used for prediction of the non-assessed single crosses (70 and 90% of the single crosses) were ridge regression BLUP (RR-BLUP), GBLUP, and pedigree-based BLUP. We used the *rrBLUP* package (Endelman 2011) for the analyses. To investigate the single cross prediction efficiency based on our model and on the models proposed by Massman et al. (2013) and Technow et al. (2012), we used another *REALbreeding* tool (*Incidence matrix*) to generate the incidence matrices for the three models and for the two DH line sampling processes. We also fitted the additive model (including only the GCA effects) to assess the relevance of the SCA effects on genomic prediction of single cross performance. The accuracies of single cross genotypic value prediction were obtained from the correlation between the true genotypic values of the non-assessed single crosses computed by *REALbreeding* and the values predicted by RR-BLUP, GBLUP, and BLUP. We also computed the efficiency of identification of the 300 non-assessed single crosses of higher genotypic value (coincidence index). The coincidence index was computed as the number of the best 300 predicted untested single crosses among the 300 untested single crosses of greater true genotypic value divided by 300. For each DH lines derivation process and heritability, the parametric average coincidence index was computed from the average phenotypic values of the 4,900 single crosses as the number of the 300 single crosses of greater average phenotypic value among the 300 single crosses of greater true genotypic value divided by 300. Regarding grain yield, for heritability of 30% the coincidence index was 0.2533, 0.2833, and 0.2433 assuming one DH line per S_0 plant, one to five DH lines per S_0 plant, and one to five DH lines per S_3 plant, respectively. The corresponding values for heritability of 60% were, respectively, 0.4800, 0.4900, and 0.4567. Concerning expansion volume, the corresponding values for heritabilities of 30 and 60% were, respectively, 0.2600, 0.2833, and 0.2700, and 0.4733, 0.5100, and 0.4533. The assumed average parametric coincidence index was 0.26 and 0.48 for heritabilities of 30 and 60%, respectively, for both traits.

Results

Using our model, average SNP density of 0.1 cM, training set size of 30%, positive dominance (grain yield), additive-dominance model, and sampling of single crosses, the prediction accuracies of the non-assessed single crosses were greater than the accuracies of the assessed single crosses for low (up to 46% higher) and intermediate (up to 16% higher) heritabilities (Table 1; Fig. 1a). As the prediction accuracy of assessed single crosses approaches 1.0, the accuracy of the non-assessed single crosses approaches ~0.9 (up to 11% lower). Sampling one to five DH lines per S_3 plant was only slightly superior to the other DH lines derivation processes, regardless of the prediction accuracy

Table 1 Average prediction accuracies of non-assessed single crosses and its standard deviation, assuming single crosses from selected DH lines, 30 and 10% of assessed single crosses, two traits, two sampling processes of single crosses, four statistical models, three DH line sampling processes, two genetic models, and three accuracies of assessed single crosses

Trait	Samp. proc.	Statistical model	DH lines	Gen. mod.	Accuracy of assessed single crosses		
					0.55	0.77	1.00
GY	SCs	Viana et al.	$1/S_0$	AD	0.7790 ± 0.0124	0.8447 ± 0.0066	0.8859 ± 0.0018
				A	0.7688 ± 0.0132	0.8380 ± 0.0067	0.8821 ± 0.0019
			$1–5/S_0$	AD	0.7947 ± 0.0125	0.8525 ± 0.0072	0.8896 ± 0.0025
				A	0.7895 ± 0.0126	0.8465 ± 0.0077	0.8858 ± 0.0027
			$1–5/S_3$	AD	0.8010 ± 0.0145	0.8678 ± 0.0054	0.9276 ± 0.0025
				A	0.7954 ± 0.0145	0.8627 ± 0.0056	0.9238 ± 0.0026
			$1–5/S_3$	AD[a]	0.7718 ± 0.0161	0.8371 ± 0.0079	0.8888 ± 0.0043
			$1–5/S_3$	AD[b]	0.6836 ± 0.0277	0.7885 ± 0.0139	0.8817 ± 0.0049
			$1/S_0$	AD[c]	0.8293 ± 0.0131	0.8944 ± 0.0049	0.9479 ± 0.0017
			$1–5/S_3$	AD[d]	0.8267 ± 0.0082	0.8928 ± 0.0043	0.9083 ± 0.0023
		Massman et al.[e]	$1/S_0$	AD	0.7874 ± 0.0118	0.8519 ± 0.0053	0.8924 ± 0.0026
			$1–5/S_0$	AD	0.7982 ± 0.0140	0.8622 ± 0.0055	0.8973 ± 0.0025
			$1–5/S_3$	AD	0.8074 ± 0.0112	0.8753 ± 0.0056	0.9314 ± 0.0026
		GBLUP	$1/S_0$	AD	0.7841 ± 0.0122	0.8477 ± 0.0064	0.8906 ± 0.0019
			$1–5/S_0$	AD	0.7973 ± 0.0124	0.8574 ± 0.0070	0.8978 ± 0.0019
			$1–5/S_3$	AD	0.7911 ± 0.0146	0.8639 ± 0.0056	0.9319 ± 0.0023
		BLUP	$1/S_0$	AD	0.7855 ± 0.0129	0.8541 ± 0.0059	0.8899 ± 0.0019
			$1–5/S_0$	AD	0.7803 ± 0.0143	0.8435 ± 0.0074	0.8830 ± 0.0024
			$1–5/S_3$	AD	0.7227 ± 0.0203	0.7915 ± 0.0077	0.8373 ± 0.0048
	DHs	Viana et al.	$1/S_0$	AD	0.5012 ± 0.0416	0.5117 ± 0.0467	0.5343 ± 0.0467
			$1–5/S_0$	AD	0.4827 ± 0.0423	0.5000 ± 0.0420	0.5036 ± 0.0465
			$1–5/S_3$	AD	0.5799 ± 0.0437	0.6106 ± 0.0413	0.6357 ± 0.0429
EV	SCs	Viana et al.	$1/S_0$	AD	0.7779 ± 0.0157	0.8458 ± 0.0069	0.8820 ± 0.0024
			$1–5/S_0$	AD	0.8019 ± 0.0155	0.8656 ± 0.0050	0.9055 ± 0.0020
			$1–5/S_3$	AD	0.7589 ± 0.0143	0.8424 ± 0.0058	0.9165 ± 0.0027

GY grain yield g/plant, *EV* expansion volume mL/g

[a]Density of 1 cM;

[b]Training set of 490 single crosses (10%);

[c]After 10 generations of random crosses;

[d]Single crosses from DH lines of the same population;

[e]and Technow et al

of the assessed single crosses (up to 5% higher). Fitting the additive model provided essentially the same prediction accuracies since the maximum decrease was ~1%. No significant differences between the prediction accuracies of non-assessed single crosses were observed when assuming bidirectional dominance (expansion volume). The differences compared to positive dominance ranged from approximately −5 to 2%. However, a striking difference was observed between the sampling processes of single crosses for testing. Random sampling of single crosses provided higher prediction accuracies of non-assessed single crosses compared to sampling DH lines for a diallel. The increases in the accuracies by sampling single crosses ranged from ~38 to 77%, which was proportional to the heritability. Decreasing the average SNP density to 1 cM led to a slight decrease in the prediction accuracy of non-assessed single crosses of approximately −4%. Decreasing the training set size to 10% decreased the prediction accuracy of non-assessed single crosses in approximately −5 to −15%, inversely proportional to the heritability. To establish that the prediction accuracy of non-assessed single crosses depends on the level of (overall) LD in the groups of selected DH or inbred lines, we derived DH lines from the same base populations after 10 generations of random crosses (to decrease the LD). The accuracies were also high, ranging from 0.83 to 0.95, proportional to the heritability. The prediction accuracies of non-assessed single crosses from DH lines of the same population were equivalent to the accuracies for single crosses derived from DH lines belonging to distinct heterotic groups, ranging from 0.83 to

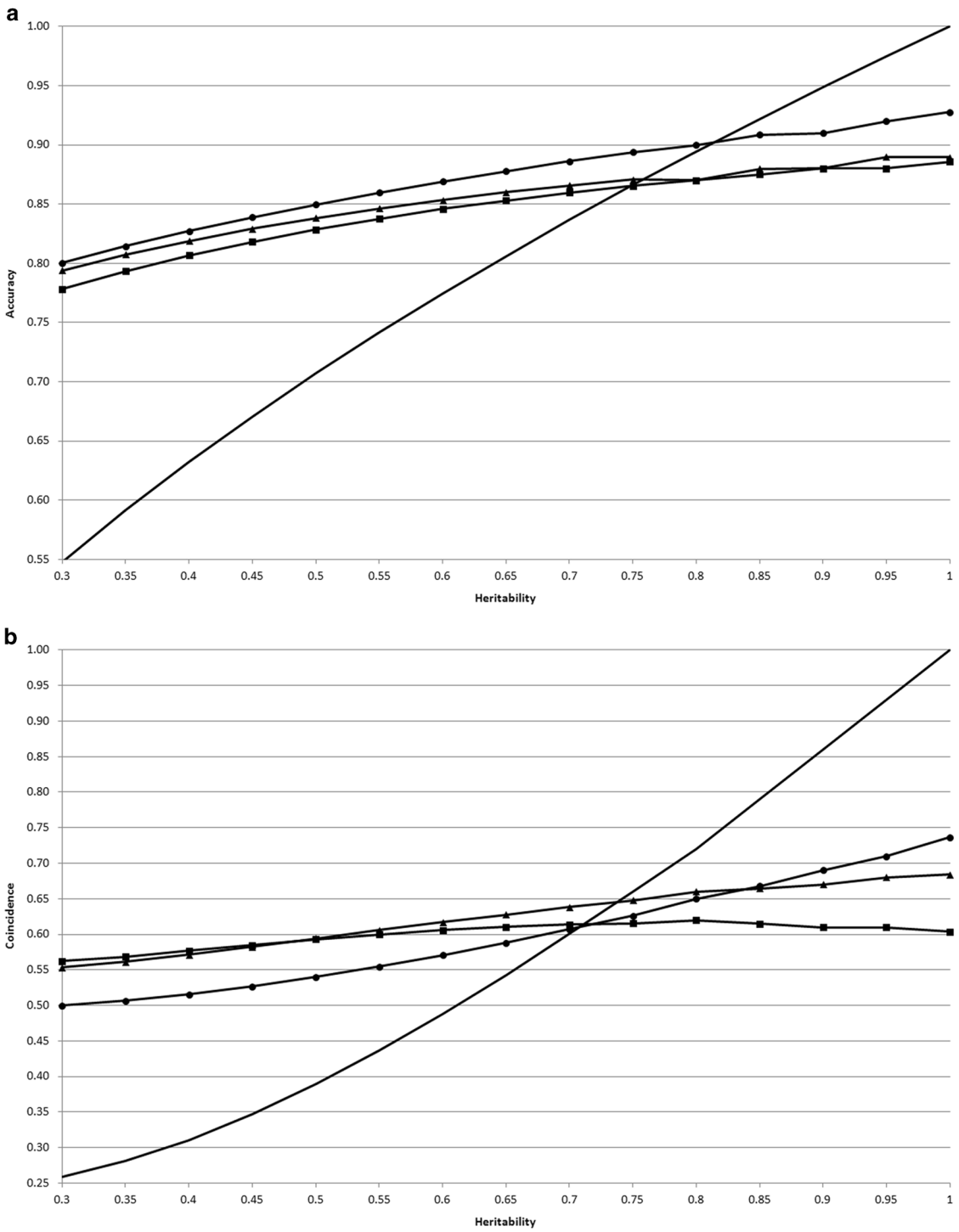

Fig. 1 Predicted accuracies (**a**) and coincidence indexes (**b**) for untested single crosses (square: 1 DH line/S_0; triangle: 1–5 DH lines/S_0; circle: 1–5 DH lines/S_3), and parametric accuracies (**a**) and coincidence indexes (**b**) for tested single crosses (continuous line), assuming our model, average SNP density of 0.1 cM, training set size of 30%, positive dominance (grain yield), additive-dominance model, and sampling of single crosses

0.91, also proportional to the heritability. When comparing our statistical model with those proposed by Massman et al. (2013) and Technow et al. (2012), we observed no differences in the prediction accuracies of non-assessed single crosses (maximum difference of 1%). Interestingly, the Massman et al. (2013) and Technow et al. (2012) models provide identical accuracies. Finally, no significant differences between the prediction accuracies for RR-BLUP,

GBLUP, and BLUP occurred (maximum of 2%), except for one to five DH lines per S_3 plant, where BLUP was 9 to 10% lower, regardless of the heritability.

Concerning the coincidence index, in general the inferences are the same as those established from the prediction accuracy analysis (Table 2; Fig. 1b). There were no differences between the coincidence indexes regarding our model and the models proposed by Massman et al. (2013) and Technow et al. (2012) (maximum difference of 3%) and between the RR-BLUP, GBLUP, and BLUP approaches, except for one to five DH lines per S_3 plant, where the coincidence for BLUP was −19 to −27% lower, proportional to the heritability. The coincidence indexes were also high for single crosses derived from selected DH lines obtained from the base populations with lower LD (ranging from 0.55 to 0.76, proportional to the heritability) and from selected DH lines of the same population (ranging from 0.61 to 0.76, also proportional to the heritability). Sampling single crosses for assessment also provided a higher coincidence index compared to sampling DH lines for a diallel (39 to 98% higher, proportional to the heritability). Decreasing the SNP density and the training set size decreased the coincidence index from 5 to 10% (proportional to the heritability) and from 17 to 26% (inversely proportional to the heritability), respectively. The maximum difference in the coincidence index by fitting the additive-dominance and the additive models was −3%. Only for one DH line per S_0 plant and assuming bidirectional dominance, the coincidence indexes were slightly greater than the values obtained assuming positive dominance (9–14%

Table 2 Average coincidence of the best 300 predicted single crosses and its standard deviation, assuming single crosses from selected DH lines, 30 and 10% of assessed single crosses, two traits, two sampling processes of single crosses, four statistical models, three DH line sampling processes, two genetic models, and three parametric coincidence of assessed single crosses

Trait	Samp.	Statistical	DH	Gen.	Coincidence of assessed single crosses		
	Proc.	Model	Lines	Mod.	0.26	0.48	1.00
GY	SCs	Viana et al.	$1/S_0$	AD	0.4523 ± 0.0334	0.5525 ± 0.0190	0.6037 ± 0.0170
				A	0.4396 ± 0.0346	0.5449 ± 0.0176	0.5976 ± 0.0172
			$1–5/S_0$	AD	0.5686 ± 0.0273	0.6369 ± 0.0221	0.6842 ± 0.0140
				A	0.5640 ± 0.0283	0.6299 ± 0.0221	0.6816 ± 0.0152
			$1–5/S_3$	AD	0.5129 ± 0.0235	0.6044 ± 0.0200	0.7363 ± 0.0183
				A	0.5063 ± 0.0225	0.5993 ± 0.0193	0.7305 ± 0.0190
			$1–5/S_3$	AD[a]	0.4881 ± 0.0278	0.5691 ± 0.0229	0.6620 ± 0.0215
			$1–5/S_3$	AD[b]	0.3805 ± 0.0511	0.4797 ± 0.0354	0.6087 ± 0.0233
			$1/S_0$	AD[c]	0.5528 ± 0.0298	0.6489 ± 0.0203	0.7571 ± 0.0162
			$1–5/S_3$	AD[d]	0.6116 ± 0.0214	0.7156 ± 0.0150	0.7581 ± 0.0166
		Massman et al.[e]	$1/S_0$	AD	0.4670 ± 0.0346	0.5663 ± 0.0174	0.6157 ± 0.0157
			$1–5/S_0$	AD	0.5651 ± 0.0310	0.6431 ± 0.0164	0.6955 ± 0.0144
			$1–5/S_3$	AD	0.5279 ± 0.0291	0.6139 ± 0.0204	0.7423 ± 0.0172
		GBLUP	$1/S_0$	AD	0.4622 ± 0.0308	0.5660 ± 0.0190	0.6092 ± 0.0163
			$1–5/S_0$	AD	0.5650 ± 0.0280	0.6384 ± 0.0204	0.6849 ± 0.0137
			$1–5/S_3$	AD	0.5010 ± 0.0245	0.5937 ± 0.0216	0.7294 ± 0.0168
		BLUP	$1/S_0$	AD	0.4641 ± 0.0331	0.5709 ± 0.0176	0.6081 ± 0.0127
			$1–5/S_0$	AD	0.5531 ± 0.0323	0.6272 ± 0.0194	0.6699 ± 0.0130
			$1–5/S_3$	AD	0.4172 ± 0.0258	0.4731 ± 0.0211	0.5377 ± 0.0196
	DHs	Viana et al.	$1/S_0$	AD	0.2753 ± 0.0374	0.3056 ± 0.0445	0.3169 ± 0.0401
			$1–5/S_0$	AD	0.3268 ± 0.0642	0.3400 ± 0.0691	0.3461 ± 0.0728
			$1–5/S_3$	AD	0.3699 ± 0.0583	0.3931 ± 0.0579	0.4300 ± 0.0633
EV	SCs	Viana et al.	$1/S_0$	AD	0.5156 ± 0.0331	0.6081 ± 0.0159	0.6599 ± 0.0146
			$1–5/S_0$	AD	0.5506 ± 0.0285	0.6337 ± 0.0203	0.6944 ± 0.0141
			$1–5/S_3$	AD	0.4746 ± 0.0294	0.5843 ± 0.0174	0.7141 ± 0.0171

GY grain yield g/plant, *EV* expansion volume mL/g

[a]density of 1 cM;

[b]training set of 490 single crosses (10%);

[c]after 10 generations of random crosses;

[d]single crosses from DH lines of the same population;

[e]and Technow et al

greater). This sampling process of DH lines provided the higher values of the coincidence index compared to the other sampling processes (7–26% higher, inversely proportional to the heritability). Finally, the coincidence index value of the non-assessed single crosses were greater than the parametric values for all assessed single crosses when assuming low (up to 117% higher) and intermediate (up to 39% higher) heritabilities (Table 1). However, as the parametric coincidence of assessed single crosses approached 1.0, the coincidence values of the non-assessed single crosses approached 0.60–0.74 (up to 26–40% lower), depending on the DH line sampling process.

Discussion

Bernardo (1994) first suggested using BLUP for predicting untested maize single cross performance. Based on the prediction accuracies obtained by Bernardo (1994, 1995, 1996a, 1996b, 1996c) for grain yield and other traits (distinct genetic controls), a breeder should realize that the performance of untested single crosses can be effectively predicted using relationship information from molecular or pedigree data, unbalanced and large data set, and diverse heterotic patterns. The significance of genomic prediction has been confirmed with maize (Zhao et al. 2015) and other important crops, such as rice (Xu et al. 2014), wheat (Zhao et al. 2013b), and barley (Philipp et al. 2016). However, there has been no published evidence that the prediction of untested single crosses is of general use by breeders of worldwide seed companies. Additional proof may be needed to make the prediction of untested single crosses as successful as Jenkins' (1934) method for predicting double-cross performance. This paper offers a significant contribution in this direction.

Our assessment on efficiency of prediction of untested single cross performance maintains some similarities with a few earlier studies, but there are sharp differences compared to most investigations. This study is based on a simulated data set, an approach also used by Technow et al. (2012), assuming 400 QTLs distributed along ten chromosomes. Thus, the prediction accuracies and coincidence indexes (a measure of untested single crosses selection efficiency) are available for non-assessed single crosses since the values were computed based on the true genotypic values of the non-assessed single crosses and not on a cross-validation procedure involving assessed single crosses. This does not mean that we consider simulated data to be better than field data or have any criticism of the cross-validation procedure. Because of the assumptions, we know that simulated data cannot integrally describe the complexity of populations and genetic determination of traits (Daetwyler et al. 2013). To highlight the relevance of (overall) LD, our study is based on conditions that are not favorable to the prediction of untested single cross performance: a very low level of relatedness between the DH lines, low and intermediate heritabilities for the assessed single crosses, and not a higher heterotic pattern. In studies by Massman et al. (2013) and Bernardo (1994, 1995, 1996a), the coancestry coefficient between inbreds from the same heterotic group ranged from 0.11 to 0.58. Riedelsheimer et al. (2012) observed high relatedness only between the non-Stiff Stalk inbreds. Technow et al. (2012) assumed non-related inbreds. For most of the investigations on prediction of untested single crosses and testcrosses, the grain yield heritability ranged from 0.72 to 0.88. The common heterotic patterns in these studies are Stiff Stalk and non-Stiff Stalk and Dent and Flint. The minor allele frequency in the groups of Dent and Flint inbreds were ~0.10 and 0.20, respectively, and ~20% of the SNPs showed a difference of allelic frequency of at least 0.60.

Concerning the prediction accuracy and the efficiency of identification of the best 300 non-assessed single crosses, our results prove that the prediction of untested single crosses is a very efficient procedure (note that we are not saying genomic prediction), especially for low and intermediate heritabilities of the assessed single crosses. The prediction accuracies of the non-assessed single crosses under low (0.55–0.71) and intermediate (0.74–0.87) accuracies of assessed single crosses achieved 0.85 and 0.89, respectively. It is important to highlight that these are not relative accuracies. Most importantly, the coincidence of the non-assessed single crosses under low (0.26–0.39) and intermediate (0.44–0.66) parametric coincidences of assessed single crosses achieved 0.59 and 0.64, respectively. For high heritability (80–95%; accuracies from 0.89 to 0.97), as observed in most studies on prediction of untested single cross performance, we can state (based on values predicted by fitting a quadratic regression model) that the prediction accuracy of non-assessed single crosses is up to only 10% lower (0.87–0.92). Most impressively, the coincidence index can range from 0.61 to 0.71 (parametric coincidences between 0.72–0.93). Under maximum accuracy of assessed single crosses (1.00), the prediction accuracy and coincidence of non-assessed single crosses achieved 0.93 and 0.76. Thus, assuming high heritability, high SNP density, and a training set size of 30%, the accuracy can achieve 0.92 and the efficiency of identification of the best 9% of the non-assessed single crosses can achieve 0.71. It is important to highlight that this efficacy can be increased by using more related DH or inbred lines, under high LD. Thus, we strongly recommend that maize breeders, as well as rice, wheat, and barley breeders, make widespread use of prediction of non-assessed single crosses, at least for preliminary screening or prior to field testing.

To take advantage of genomic prediction, Kadam et al. (2016) recommend redesigning hybrid breeding programs.

However, because breeders are unlikely to rely solely on genomic predictions when selecting superior untested hybrids, Technow et al. (2014) believe that genomic prediction will be combined with field testing of the most promising experimental hybrids. For grain yield, the prediction accuracies observed by Bernardo (1994, 1995, 1996a) ranged from 0.14 to 0.80, proportional to the heritability (in the range 35–74%) and training set size. The non-relative accuracies (relative accuracy × root square of heritability) observed in the studies of Kadam et al. (2016), Technow et al. (2014), Massman et al. (2013), Technow et al. (2012), and Riedelsheimer et al. (2012) ranged between 0.20 and 0.86, also proportional to the heritability (in the range 53–98%) and training set size.

We hope that readers have realized the importance of (overall) LD for effective prediction of non-assessed single crosses, as well as genetic variability. Breeders have no control over LD and relatedness between the DH or inbred lines. However, selection should always provide a high level of overall LD in the groups of selected DH or inbred lines. Comparison of our LD assessment with the LD analyses from other studies is inadequate because our distances are in cM and not in base-pairs. But in general, the level of LD was high (r^2 of ~0.3) for only SNPs separated by up to 0.5 Mb (Massman et al. 2013; Riedelsheimer et al. 2012; Technow et al. 2012, 2014). To maximize the prediction accuracy and the efficiency of identification of the best non-assessed single crosses it is necessary to adopt random sampling of single crosses for testing instead of the random sampling of DH or inbred lines for a diallel. This is because sampling 30 or even 10% of the single crosses leads to single crosses for testing derived from all DH or inbred lines from each group. In our case, in every resampling assuming training set size of 30 and 10% we always get groups of assessed single crosses (1470 and 490 single crosses, respectively) derived from the 70 DH lines of each group. However, sampling DH lines for a diallel provided 1440 and 484 single crosses for testing derived from 38 and 22 DH lines, respectively. Thus, the sampling of single crosses provides the best prediction of the SNP average effects of substitution and dominance deviations. Riedelsheimer et al. (2012) emphasized the need for large genetic variability to obtain high prediction accuracies. Furthermore, their results indicated that pairs of closely related lines and population structuring only weakly contributed to the high prediction accuracies. Because dominance can be a relevant genetic effect, breeders should always fit the additive-dominance model to maximize the prediction accuracy and the efficiency of identification of the best non-assessed single crosses. Interestingly, in most of the studies on prediction of non-assessed single crosses the prediction accuracy did not increase significantly when modeling SCA in addition to GCA effects (Zhao et al. 2015).

Concerning SNP density and training set size, factors related to the costs of genotyping and phenotyping, breeders should find a balance between efficiency and expenses, since maximizing SNP density and training set size maximizes the efficiency of untested single cross prediction. Based on our results, because the decreases in the prediction accuracy (~4%) and coincidence index (5–10%) by decreasing the average SNP density from 0.1 to 1 cM are of reduced magnitude, we consider sufficient to employ custom genotyping to provide an average SNP density of 1 cM. Decreasing the training set size from 30 to 10% of the single crosses does not significantly affect the prediction accuracy under intermediate to high heritability (decrease of up to 9%), but the coincidence index can be reduced by up to 21%. However, considering that the coincidence index will be kept in the range 0.48–0.61, proportional to the heritability, and that the maximum values are in the range 0.48 to 0.61, we also consider sufficient to assess at least 10% of the possible single crosses. As highlighted by Zhao et al. (2015), marker density only marginally affects the prediction accuracy of untested single crosses and for biparental populations a plateau for the accuracy is reached with a few hundred markers. Technow et al. (2014) did not find an improvement in prediction accuracies when using higher SNP density. Additionally, increasing the training set size led to a relatively small increase in the prediction accuracy. However, the prediction accuracies obtained by Riedelsheimer et al. (2012) under high density (38,019 SNPs) were substantially higher than those reached with a low-density marker panel (1152 SNPs). In the study of Technow et al. (2012), the prediction accuracies increased with SNP density and number of parents tested in hybrid combination.

The DH line sampling process, heterotic pattern, and statistical approach should not be worries for breeders. However, under high heritability, sampling more than one DH line per S_0 or S_3 plant provided higher coincidence values and high prediction accuracy in our study. For rice, wheat, and barley breeders, our message is that high prediction accuracy and high efficiency of identification of the best non-assessed single crosses does not depend on heterotic groups but on the (overall) LD in the group or in each group of DH or inbred lines. In other words, the efficiency of prediction of non-assessed single crosses derived from DH or inbred lines from the same population can be as high as the prediction efficiency of untested single crosses derived from DH or inbred lines from distinct heterotic groups. This was not confirmed comparing the relative prediction accuracies for the grain yield of maize untested single crosses (from ~0.50 to 0.95, for most studies) with those obtained with rice, wheat, and barley untested hybrids (0.50–0.60, approximately) (Philipp et al. 2016; Xu et al. 2014; Zhao et al. 2013b). However, the lower relative prediction accuracies for untested rice, wheat, and barley

hybrids should be due to prediction of two-way and three-way crosses. Regarding the statistical approach, our model did not provide an increase in the efficiency of non-assessed single cross prediction compared to the models proposed by Massman et al. (2013) and Technow et al. (2012). Importantly, our results showed that these two models are really identical (data not shown). Thus, because of the simplified definition of the incidence matrices for these two previous models, it is quite safe to use either of them. Finally, the choice between the statistical approaches RR-BLUP (based on prediction of SNP average effects of substitution and dominance deviations), GBLUP (based on additive and dominance genomic matrices), and pedigree-based BLUP (prediction of genotypic values of non-assessed single crosses based on additive and dominance matrices from pedigree records) should not be a serious worry for breeders as well. Our evidence is that there is no significant difference between RR-BLUP and GBLUP regarding the prediction accuracy and efficiency of identification of the best untested single crosses. Furthermore, even when the level of relatedness between the DH or inbred lines in each group is low, pedigree-based BLUP is generally as efficient as genomic prediction, except when the DH lines are derived from an inbred population. Thus, DNA polymorphism is not essential for efficient prediction of non-assessed single cross performance. In a review on genomic selection in hybrid breeding, Zhao et al. (2015) state that the choice of the biometrical model has no substantial impact on the prediction accuracy of untested single crosses. Technow et al. (2014) observed that the GBLUP and BayesB prediction methods resulted in very similar prediction accuracies. According to Massman et al. (2013), the pedigree-based BLUP and RR-BLUP models did not lead to significantly different prediction accuracies. Technow et al. (2012) concluded that BayesB produced significantly higher accuracies for the additive-dominance model than GBLUP.

Our main contributions to the assessment of prediction efficiency of untested single cross performance are the following: (1) the prediction accuracy of untested single crosses ranged from ~0.80 to 0.90 as the heritability of tested single crosses ranged from low (30%) to high (100%); however, the efficacy of identification of the best 9% of the untested single crosses ranged from ~0.50 to 0.70, depending on the DH line sampling process; (2) the prediction accuracy for crops showing no defined heterotic pattern can be as efficient as with maize, for which there are well-defined heterotic groups; this is because the most important factor affecting the prediction efficiency is the overall LD; (3) to maximize prediction accuracy and coincidence the choice of single crosses for testing should be based on a random process; this procedure maximizes the number of DH lines in hybrid combinations and provides better predictions of the SNP average effects of substitution and dominance deviations compared to sampling DH lines for a diallel; (4) because of the non-significant decreases in the prediction accuracy and coincidence, the prediction of untested single crosses can be efficient when assuming a reduced training set size (10%) and SNP density of 1 cM; (5) RR-BLUP and GBLUP provide equivalent prediction efficiencies of untested single crosses; (6) except for DH lines derived from inbred populations, pedigree-based BLUP is as efficient as genomic prediction of untested single crosses; and (7) the theoretical accuracy shows that the prediction accuracy is not affected by the linkage phase.

Data archiving

The data set is available at https://doi.org/10.6084/m9.figshare.5035130.v3.

Acknowledgements We thank the National Council for Scientific and Technological Development (CNPq), the Brazilian Federal Agency for Support and Evaluation of Graduate Education (Capes), and the Foundation for Research Support of Minas Gerais State (Fapemig) for financial support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Albrecht T, Auinger H-J, Wimmer V, Ogutu JO, Knaak C, Ouzunova M et al. (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127(6):1375–1386

Albrecht T, Wimmer V, Auinger H-J, Erbe M, Knaak C, Ouzunova M et al. (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123(2):339–350

Bernardo R (1996a) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50–56

Bernardo R (1996b) Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. Crop Sci 36:862–866

Bernardo R (1996c) Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci 36:872–876

Bernardo R (1995) Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci 35:141–147

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25

Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193 (2):347–365

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2):327–345

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3):250–255

Grattapaglia D, Resende MDV (2010) Genomic selection in forest tree breeding. Tree Genet & Genomes 7(2):241–255

Jenkins MT (1934) Methods of estimating the performance of double crosses in corn. J Am Soc Agron 26:199–204

Kadam DC, Potts SM, Bohn MO, Lipka AE, Lorenz AJ (2016) Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. G3-Genes Genomes Genet 6 (11):3443–3453

Kempthorne O (1957) An Introduction to Genetic Statistics. John Wiley and Sons Inc, New York

Li Z, Philipp N, Spiller M, Stiewe G, Reif JC, Zhao YS (2017) Genome-wide prediction of the performance of three-way hybrids in barley. Plant Genome 10:1

Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2013) Genomewide predictions from maize single-cross data. Theor Appl Genet 126(1):13–22

Meuwissen T, Hayes B, Goddard M (2013) Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci, Vol 1 1:221–237

Philipp N, Liu GZ, Zhao YS, He S, Spiller M, Stiewe G et al. (2016). Genomic prediction of barley hybrid performance. Plant Genome **9**(2).

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R et al. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44 (2):217–220

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125(6):1181–1194

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197(4):1343–U1469

Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM (2014) Applied animal genomics: results from the field. Annu Rev Anim Biosci 2(2):105–139

Viana JMS (2004) Quantitative genetics theory for non-inbred populations in linkage disequilibrium. Genet Mol Biol 27(4):594–601

Viana JMS, Piepho H-P, Silva FF (2016) Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. Sci Agric 73(3):243–251

Xu S, Zhu D, Zhang Q (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci USA 111(34):12456–12461

Zhao Y, Gowda M, Liu W, Wuerschum T, Maurer HP, Longin FH et al. (2013a) Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. Plant Breed 132(1):99–106

Zhao Y, Mette MF, Reif JC (2015) Genomic selection in hybrid breeding. Plant Breed 134(1):1–10

Zhao Y, Zeng J, Fernando R, Reif JC (2013b) Genomic prediction of hybrid wheat performance. Crop Sci 53(3):802