1 **Efficiency of genomic prediction of non-assessed single crosses**

2 José Marcelo Soriano Viana,[*1] Helcio Duarte Pereira,[1] Gabriel Borges Mundim,[†] Hans-Peter

3 Piepho,[‡] and Fabyano Fonseca e Silva[§]

4 [*]Federal University of Viçosa, Department of General Biology, 36570-900, Viçosa, MG, Brazil.

5 [†]Dow AgroSciences Seeds and Biotechnology Brazil Ltda, 38490-000, Indianópolis, MG, Brazil.

6 [‡]University of Hohenheim, Institute of Crop Science, Biostatistics Unit, 70599, Stuttgart, Germany.

7 [§]Federal University of Viçosa, Department of Animal Science, 36570-900, Viçosa, MG, Brazil.

8 Reference number for data available in public repository:

9 https://doi.org/10.6084/m9.figshare.5035130.v3

10 *REALbreeding* private link: https://figshare.com/s/618bee7accd410464232.

11   Running title: Genomic prediction of single crosses.

12   **KEYWORDS** genomic selection; linkage disequilibrium; general combining ability; specific

13   combining ability; doubled haploids.

14   [1]Corresponding author: José Marcelo Soriano Viana. Federal University of Viçosa, Department of

15   General Biology, 36570-900, Viçosa, MG, Brazil. E-mail: jmsviana@ufv.br. Telephone:

16   +55(31)3899-2514.

17   **ABSTRACT** An important application of genomic selection in plant breeding is the prediction of

18   untested single crosses (SCs). Most investigations on the prediction efficiency were based on tested

19   SCs, using cross-validation. The main objective was to assess the prediction efficiency by

20   correlating the predicted and true genotypic values of untested SCs (accuracy) and measuring the

21   efficacy of identification of the best 300 untested SCs (coincidence), using simulated data. We

22   assumed 10,000 SNPs, 400 QTLs, two groups of 70 selected DH lines, and 4,900 SCs. The

23   heritabilities for the assessed SCs were 30, 60 and 100%. The scenarios included three sampling

24   processes of DH lines, two sampling processes of SCs for testing, two SNP densities, DH lines from

25   distinct and same populations, DH lines from populations with lower LD, two genetic models, three

26   statistical models, and three statistical approaches. We derived a model for genomic prediction

27   based on SNP average effects of substitution and dominance deviations. The prediction accuracy is

28   not affected by the linkage phase. The prediction of untested SCs is very efficient. The accuracies

29   and coincidences ranged from approximately 0.8 and 0.5, respectively, under low heritability, to 0.9

30   and 0.7, assuming high heritability. Additionally, we highlighted the relevance of the overall LD

31   and evidenced that efficient prediction of untested SCs can be achieved for crops that show no

32   heterotic pattern, for reduced training set size (10%), for SNP density of 1 cM, and for distinct

33   sampling processes of DH lines, based on random choice of the SCs for testing.

34   **INTRODUCTION**

2

35   Genomic selection is very commonly used in animal breeding programs, especially for dairy

36   cattle (Van Eenennaam et al. 2014). The same cannot yet be said to the same degree concerning

37   crop breeding. The main reasons for the effective application of genomic selection in livestock

38   breeding are: it is efficient, that is, the process has high prediction accuracy, the cost of phenotyping

39   (mainly progeny test) is higher than the cost of genotyping, and the process significantly shortens

40   the selection cycle (Meuwissen et al. 2013). In spite of the many field- and simulation-based studies

41   with genomic selection in plant breeding, in general the cost of phenotyping is often still much

42   lower than the cost of genotyping, restricting its application in breeding programs. Jonas and de

43   Koning (2013) consider that genomic selection has the potential to improve existing plant breeding

44   schemes. However, based also on the high diversity and complexity of plant breeding methods, they

45   stated that there are great obstacles to overcome.

46   An important application of genomic selection in plant breeding is the prediction of untested

47   single crosses (genotypic value prediction) and testcrosses (general combining ability effect

48   prediction) in hybrid breeding (Zhao et al. 2015). Genomic prediction of two- and three-way crosses

49   has been investigated (Philipp et al. 2016). The prediction of untested single crosses was pioneered

50   by Bernardo (1994), based on best linear unbiased prediction (BLUP). Many significant studies on

51   prediction of untested single cross and testcross performance have been published in the last 23

52   years, focused on the assessment of the prediction accuracy. Most investigations were based on

53   empirical data and estimated the prediction accuracy using a cross-validation procedure. Very few

54   were based on simulated data (Li et al. 2017; Technow et al. 2012). With no exception, the

55   inference was that prediction of untested single crosses and testcrosses can be an efficient,

56   depending on heritability, training set size, and number of tested inbreds in hybrid combination

57   (both, one, and none parents tested). Remarkably, this conclusion was drawn from studies differing

58   in the type of molecular marker, density of markers, number of inbreds, level of relatedness,

59   diversity, and linkage disequilibrium (LD) between inbreds, heterotic pattern, training set size,

3

60   genetic model, and statistical approach (Zhao et al. 2015). Efficient prediction of barley two- and

61   three-way crosses has been achieved when training and validation sets include the same class of

62   hybrids (Philipp et al. 2016).

63   Most studies on genomic prediction of maize single cross performance published since 2011

64   have employed single nucleotide polymorphisms (SNP), with the number SNPs filtered ranging

65   from 425 (Zhao et al. 2013a) to 39,627 (Technow et al. 2012). Based on the physical length of the

66   maize genome (approximately 2,106 megabase pairs (Mb) according to Maize genetics and

67   genomics database), the SNP density ranged from approximately 5 to 0.05 Mb, respectively. For

68   grain yield, the relative prediction accuracies (computed as accuracy/root square of the heritability)

69   in the two previously cited papers ranged from 0.27 to 0.62 and from 0.65 to 0.95, respectively. The

70   number of inbreds in each heterotic group was highly variable too, ranging from six and nine

71   (Bernardo 1994) to 75 and 75 (Technow et al. 2012). The relative accuracy observed by Bernardo

72   (1994) ranged between 0.72 and 0.89. The number of testcrosses ranged between 255 (Windhausen

73   et al. 2012) and 1,894 (Albrecht et al. 2014). The relative accuracies ranged from 0.46 to 0.52 and

74   from 0.33 to 0.65, respectively. The level of relatedness ranged from non-related inbreds in each

75   group (Technow et al. 2012) to a maximum average value of 0.58 (Bernardo 1995). The relative

76   accuracy obtained by Bernardo (1995) ranged from 0.41 to 0.80. The common heterotic groups

77   were Stiff Stalk and non-Stiff Stalk (Kadam et al. 1916) or Dent and Flint (Technow et al. 2014).

78   The study of Bernardo (1996a) involved nine heterotic groups and the (statistically significant from

79   zero) relative accuracies ranged from 0.43 to 0.88. No study provided clearly greater prediction

80   accuracy of the additive-dominance model relative to the additive model. Finally, only with

81   testcrosses the genomic BLUP (GBLUP) approach outperformed pedigree-based BLUP (Albrecht

82   et al. 2014; Albrecht et al. 2011) concerning prediction accuracy.

83   Genomic prediction of single crosses has been made based on tested single crosses, using

84   cross-validation. Thus, the estimated prediction accuracies are not for untested single crosses.

4

85   Consequently, none of the previous studies on efficiency of genomic prediction of single cross

86   performance measured the efficacy of identification of the best untested single crosses. Our main

87   objective was to assess the efficiency of prediction of untested single crosses by correlating the

88   predicted and true genotypic values of untested single crosses (prediction accuracy) and measuring

89   the efficacy of identification of the best 300 untested single crosses (coincidence index), using a

90   large simulated data set. The secondary objectives were to highlight that the prediction accuracy

91   depends primarily on the overall LD in the groups of selected doubled haploid (DH) lines, that the

92   prediction efficiency when there is no heterotic pattern can be as high as the prediction efficiency

93   when there are heterotic groups, and that the choice of single crosses for testing should be random,

94   instead of selecting DH lines for a diallel, to maximize the prediction efficiency. Further, we

95   derived a model for genomic prediction of untested single crosses based on the SNP average effects

96   of substitution and dominance deviations.

MATERIALS AND METHODS

98   **Theory**

99   Generally, most papers on genomic selection presents only statistical aspects and the genetic

100 models are deduced from gene to SNP effects. Importantly, when there is some quantitative

101 genetics theory, the LD between QTLs and SNPs is usually completely ignored. The quantitative

102 genetics theory developed in this paper provides a genetic model for genomic prediction of untested

103 single crosses that accounts for the LD between QTLs and SNPs. The model developed offers the

104 genetic background to the models fitted in important previously papers on prediction of untested

105 single crosses and testcrosses (Massman et al. 2013; Technow et al. 2012; Albrecht et al. 2011).

106 ***LD in a group of selected DH or inbred lines***

107 Consider a group of DH or inbred lines selected from a population or heterotic group. Assume

108 also a QTL (alleles B/b) and a SNP (alleles C/c) where B and b are the alleles that increase and

109 decrease the trait expression, respectively. Define the joint genotype probabilities as

5

110     $P(BBCC) = f_{22}$, $P(BBcc) = f_{20}$, $P(bbCC) = f_{02}$,   and $P(bbcc) = f_{00}$, where the subscript

111   indicates the number of copies of the major allele (B and C). The measure of LD between the QTL

112   and the SNP is $\Delta_{bc} = f_{22}f_{00} - f_{20}f_{02}$ (Kempthorne 1954) and the haplotype frequencies are

113     $P(BC) = f_{22} = p_b p_c + \Delta_{bc}$, $P(Bc) = f_{20} = p_b q_c - \Delta_{bc}$, $P(bC) = f_{02} = q_b p_c - \Delta_{bc}$,   and

114     $P(bc) = f_{00} = q_b q_c + \Delta_{bc}$, where p is the frequency of the major allele (B or C) and $q = 1 - p$ is

115   the frequency of the minor allele (b or c). Notice that $p_b = f_{22} + f_{20}$ and $p_c = f_{22} + f_{02}$. It is

116   important to highlight the fact that we are not assuming that the QTL and the SNP are linked and in

117   LD in the population or heterotic group, because this is not a necessary condition for genomic

118   prediction. But we are assuming that they are in LD in the group of DH or inbred lines.

119   Furthermore, because of selection, genetic drift, and inbreeding (only for inbreds and linked QTLs

120   and SNPs), the gene and genotypic frequencies and the LD values concerning the selected DH or

121   inbred lines cannot be traced to the values in the population or heterotic group.

122   ***SNP genotypic values of DH or inbred lines***

123   The average genotypic value for a group of selected DH or inbred lines is

124     $M_{IL} = m_b + \left(p_b - q_b\right)a_b$, where m_b is the mean of the genotypic values of the homozygotes and

125   a_b is the deviation between the genotypic value of the homozygote of higher expression and m_b.

126   Thus, the average SNP genotypic values for the DH or inbred lines CC and cc are

127     $G_{CC} = \dfrac{1}{f_{.2}}\left[f_{22}\left(m_b + a_b\right) + f_{02}\left(m_b - a_b\right)\right] = M_{IL} + 2q_c\alpha_{SNP} = M_{IL} + A_{CC}$

128     $G_{cc} = \dfrac{1}{f_{.0}}\left[f_{20}\left(m_b + a_b\right) + f_{00}\left(m_b - a_b\right)\right] = M_{IL} - 2p_c\alpha_{SNP} = M_{IL} + A_{cc}$

6

129 where $\alpha_{SNP} = \left[\dfrac{\Delta_{bc}}{p_c q_c}\right] a_b = \kappa_{bc} a_b$ is the average effect of a SNP substitution in the group of DH

130 or inbred lines and A is the SNP additive value for a DH or inbred line. Notice that E(A) = 0.

131   Assuming two QTLs (alleles B and b, and E and e) in LD with the SNP, the average effect of

132 a SNP substitution in the selected DH or inbred lines is $\alpha_{SNP} = \kappa_{bc} a_b + \kappa_{ce} a_e$, where

133 $\kappa_{ce} = \left[\dfrac{\Delta_{ce}}{p_c q_c}\right]$. Thus, in general, the average effect of a SNP substitution (and the SNP additive

134 value) is proportional to the LD measure and to the a deviation for each QTL that is in LD with the

135 marker.

SNP genotypic values of single crosses

137   Aiming to maximize the heterosis, maize breeders commonly assess single crosses originating

138 from selected DH or inbred lines from distinct heterotic groups. Consider n_1 DH or inbred lines

139 from a population or heterotic group and n_2 DH or inbred lines from a distinct population or

140 heterotic group. The average genotypic value for the single crosses derived by crossing the DH or

141 inbred lines from group 1 with the DH or inbred lines from group 2 is

142 $M_H = m_b + \left(p_{b1}p_{b2} - q_{b1}q_{b2}\right) a_b + \left(p_{b1}q_{b2} + q_{b1}p_{b2}\right) d_b$

143 where d_b is the dominance deviation (the deviation between the genotypic value of the

144 heterozygote and m_b).

145   The average genotypic values for the single crosses derived from DH or inbred lines CC and

146 cc of the group 1 are

147 $M_{CC1} = M_H + q_{c1}\kappa_{bc1}\left[a_b + \left(q_{b2} - p_{b2}\right)d_b\right] = M_H + q_{c1}\kappa_{bc1}\alpha_{b2} = M_H + q_{c1}\alpha_{SNP1}$

$= M_H + GCA_{CC1}$

7

148 $\quad M_{cc1} = M_H - p_{c1}\kappa_{bc1}\alpha_{b2} = M_H - p_{c1}\alpha_{SNP1} = M_H + GCA_{cc1}$

149 $\quad$ where α_{b2} is the average effect of allelic substitution in the population derived by random crosses

150 $\quad$ between the DH or inbred lines of group 2, α_{SNP1} is the SNP effect of allelic substitution in the

151 $\quad$ hybrid population relative to a SNP derived from group 1, and GCA stands for the general

152 $\quad$ combining ability effect for a SNP locus. Notice that α_{SNP1} depends on the LD in group 1

153 $\quad$ ($\kappa_{bc1} = \Delta_{bc1}/p_{c1}q_{c1}$) and the average effect of allelic substitution in the population derived by

154 $\quad$ random crosses between the DH or inbred lines of group 2. Further,

155 $\quad$ $E(GCA) = p_{c1}GCA_{CC1} + q_{c1}GCA_{cc1} = 0$. Concerning the single crosses derived from DH or

156 $\quad$ inbred lines CC and cc of the group 2 we have

157
$$M_{CC2} = M_H + q_{c2}\kappa_{bc2}\left[a_b + \left(q_{b1} - p_{b1}\right)d_b\right] = M_H + q_{c2}\kappa_{bc2}\alpha_{b1} = M_H + q_{c2}\alpha_{SNP2}$$
$$= M_H + GCA_{CC2}$$

158 $\quad$ $M_{cc2} = M_H - p_{c2}\kappa_{bc2}\alpha_{b1} = M_H - p_{c2}\alpha_{SNP2} = M_H + GCA_{cc2}$

159 $\quad$ Notice that E(GCA) = 0 also. The average genotypic values for the single crosses concerning

160 $\quad$ the SNP locus are

161
$$M_{CC1xCC2} = M_H + q_{c1}\alpha_{SNP1} + q_{c2}\alpha_{SNP2} - 2q_{c1}q_{c2}\kappa_{bc1}\kappa_{bc2}d_b$$
$$= M_H + GCA_{CC1} + GCA_{CC2} + SCA_{CC1xCC2}$$

162
$$M_{cc1xcc2} = M_H - p_{c1}\alpha_{SNP1} - p_{c2}\alpha_{SNP2} - 2p_{c1}p_{c2}\kappa_{bc1}\kappa_{bc2}d_b$$
$$= M_H + GCA_{cc1} + GCA_{cc2} + SCA_{cc1xcc2}$$

163
$$M_{CC1xcc2} = M_H + q_{c1}\alpha_{SNP1} - p_{c2}\alpha_{SNP2} + 2q_{c1}p_{c2}\kappa_{bc1}\kappa_{bc2}d_b$$
$$= M_H + GCA_{CC1} + GCA_{cc2} + SCA_{CC1xcc2}$$

164
$$M_{cc1xCC2} = M_H - p_{c1}\alpha_{SNP1} + q_{c2}\alpha_{SNP2} + 2p_{c1}q_{c2}\kappa_{bc1}\kappa_{bc2}d_b$$
$$= M_H + GCA_{cc1} + GCA_{CC2} + SCA_{cc1xCC2}$$

165 where $\kappa_{bc1}\kappa_{bc2}d_b = d_{SNP}$ is the SNP dominance deviation in the hybrid population and SCA

166 stands for the specific combining ability effect for a SNP locus. Notice that $E(SCA) =$

167 $p_{c1}p_{c2}SCA_{CC1xCC2} + p_{c1}q_{c2}SCA_{CC1xcc2} + q_{c1}p_{c2}SCA_{cc1xCC2} + q_{c1}q_{c2}SCA_{cc1xcc2} = 0$ and

168 , for each group, $E(SCA|CC) = E(SCA|cc) = 0$. That is, the expectation of the SNP SCA effects

169 given a SNP genotype for the common DH or inbred line is also zero. Notice also that the four

170 genotypic values depends on four unknown parameters (M_H, α_{SNP1}, α_{SNP2}, and d_{SNP}).

171 Assuming two QTLs (alleles B and b, and E and e) in LD with the SNP, the SNP dominance

172 deviation is $d_{SNP} = \kappa_{bc1}\kappa_{bc2}d_b + \kappa_{ce1}\kappa_{ce2}d_e$. Thus, generally, the SNP dominance deviation

173 (and the SNP SCA effect) is proportional to the product of the LD values in both groups of DH or

174 inbred lines and to the dominance deviation for each QTL that is in LD with the marker.

175 The previous model expressed as a function of the SNP GCA and SCA effects was proposed

176 by Massman et al. (2013), but these authors assumed $GCA_{CC} + GCA_{cc} = 0$ (for each heterotic

177 group and for each SNP) and $SCA_{CC1xCC2} = SCA_{cc1xcc2} = -SCA_{CC1xcc2} = -SCA_{cc1xCC2}$.

178 Technow et al. (2012) have used a standard extension from QTL to SNP, defining the single cross

179 genotypic value for a SNP as a function of the SNP a and d deviations. That is,

180 $M = M_H + u_1a_1 + u_2a_2 + u_3d$, where u_1 and u_2 equal to 1/2 or −1/2 if the corresponding DH or

181 inbred line is homozygous for distinct SNP alleles (CC or cc), and u_3 equal to 0 if the single cross

182 is homozygous or 1 if heterozygous.

183 ***SNP genotypic values of single crosses from DH or inbred lines derived from the same***

184 ***population or heterotic group***

185   Well defined heterotic groups are known for maize, but not for special maize such as popcorn

186   and sweet corn and for other crops such as wheat (Zhao et al. 2013b), rice (Xu et al. 2014), and

187   barley (Philipp et al. 2016). Thus, for many breeders, it is interesting to know about the efficiency

188   of genomic prediction of singles crosses when there are no heterotic groups. Assuming n DH or

189   inbred lines derived from the same population or heterotic group, the average genotypic values for

190   the single crosses concerning the SNP locus are

191      $M_{CCxCC} = M + 2q_c\alpha_{SNP} - 2q_c^2\kappa_{bc}^2 d_b = M + 2GCA_{CC} + SCA_{CCxCC}$

192      $M_{ccxcc} = M - 2p_c\alpha_{SNP} - 2p_c^2\kappa_{bc}^2 d_b = M + 2GCA_{cc} + SCA_{ccxcc}$

193      $M_{CCxcc} = M + 2(q_c - p_c)\alpha_{SNP} + 2p_cq_c\kappa_{bc}^2 d_b = M + GCA_{CC} + GCA_{cc} + SCA_{CCxcc}$

194   where $M = m_b + (p_c - q_c)a_b + 2p_cq_cd_b$ is   the   hybrid   population   mean,

195      $\alpha_{SNP} = \kappa_{bc}[a_b + (q_b - p_b)d_b] = \kappa_{bc}\alpha_b$ is the average effect of a SNP substitution in the hybrid

196   population, and $d_{SNP} = \kappa_{bc}^2 d_b$ is the SNP dominance deviation. Notice that the SNP GCA effects

197   are equal to half the SNP additive value for the single crosses (A), the SNP SCA effects are the SNP

198   dominance deviations for the single crosses (D), and that the three genotypic values depends on

199   three unknown parameters (M, α_{SNP}, and d_{SNP}). Notice also that E(GCA) = E(A) = E(SCA) =

200   E(SCA|CC) = E(SCA|cc) = E(D) = 0.

201   ***Accuracy of single cross genomic prediction***

202   Assuming a QTL and a SNP in LD in the two groups of DH or inbred lines, the predictor of

203   the single cross QTL genotypic value is the single cross SNP genotypic value (because they are

204   proportional). Thus, the covariance between the predictor and the genotypic value is

$$\text{Cov}(\widetilde{G}, G) = f_{22}^1 f_{22}^2 \left[M_H + GCA_{CC1} + GCA_{CC2} + SCA_{CC1xCC2} \right]\left[M_H + GCA_{BB1} + GCA_{BB2} + SCA_{BB1xBB2} \right] +$$

$$+ f_{22}^1 f_{20}^2 \left[M_H + GCA_{CC1} + GCA_{cc2} + SCA_{CC1xcc2} \right]\left[M_H + GCA_{BB1} + GCA_{BB2} + SCA_{BB1xBB2} \right] +$$

$$\dots$$

$$+ f_{00}^1 f_{00}^2 \left[M_H + GCA_{cc1} + GCA_{cc2} + SCA_{cc1xcc2} \right]\left[M_H + GCA_{bb1} + GCA_{bb2} + SCA_{bb1xbb2} \right] - (M_H)^2$$

$$= p_{c1}q_{c1}\left(\kappa_{bc1}\alpha_{b2} \right)^2 + p_{c2}q_{c2}\left(\kappa_{bc2}\alpha_{b1} \right)^2 + 4p_{c1}q_{c1}p_{c2}q_{c2}\left(\kappa_{bc1}\kappa_{bc2}d_b \right)^2$$

$$= p_{c1}q_{c1}\left(\alpha_{SNP1} \right)^2 + p_{c2}q_{c2}\left(\alpha_{SNP2} \right)^2 + 4p_{c1}q_{c1}p_{c2}q_{c2}(d_{SNP})^2$$

$$= \sigma^{2(1)}_{GCA_{SNP}} + \sigma^{2(2)}_{GCA_{SNP}} + \sigma^2_{SCA_{SNP}} = \sigma^2_{G(SNP)}$$

where the GCA and SCA effects for the QTL are $GCA_{BB1} = q_{b1}\alpha_{b2}$, $GCA_{bb1} = -p_{b1}\alpha_{b2}$,

$GCA_{BB2} = q_{b2}\alpha_{b1}$, $\qquad GCA_{bb2} = -p_{b2}\alpha_{b1}$, $\qquad SCA_{BB1xBB2} = -2q_{b1}q_{b2}d_b$,

$SCA_{BB1xbb2} = 2q_{b1}p_{b2}d_b$, $\quad SCA_{bb1xBB2} = 2p_{b1}q_{b2}d_b$, 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 $\sigma^{2(1)}_{GCA} = p_{b1}q_{b1}\left(\alpha_{b2} \right)^2$,

$\sigma^{2(2)}_{GCA} = p_{b2}q_{b2}\left(\alpha_{b1} \right)^2$, 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

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

Assuming s SNPs,

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

11

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

218   genotypic variance. Further,

219    $\alpha_{SNP(r)1} = \sum\limits_{i=1}^{k'} \left[\dfrac{\Delta_{ri1}}{p_{r1}q_{r1}} \right] \alpha_{i2} = \sum\limits_{i=1}^{k'} \kappa_{ri1}\alpha_{i2}$, where k' is the number of QTLs in LD with the SNP

220   r in group 1, and

221    $d_{SNP(r)} = \sum\limits_{i=1}^{k''} \left[\dfrac{\Delta_{ri1}}{p_{r1}q_{r1}} \right]\left[\dfrac{\Delta_{ri2}}{p_{r2}q_{r2}} \right] d_i = \sum\limits_{i=1}^{k''} \kappa_{ri1}\kappa_{ri2}d_i$ where k" is the number of QTLs in LD with

222   the SNP r in both groups

223   Notice that because the accuracy of genomic prediction of single crosses depends on the

224   squares of the average effects of SNP substitution and the SNP dominance deviations, it is not

225   affected by the linkage phase (coupling or repulsion), as it does not depend on linkage. But it

226   depends on the magnitude of the LD in each group of DH or inbred lines.

227   Assuming single crosses derived from DH or inbred lines of a single population or heterotic

228   group we have $\sigma^2_{G(SNP)} = 2p_cq_c\left(\alpha_{SNP}\right)^2 + \left(2p_cq_cd_{SNP}\right)^2$ and

229    $\sigma^2_G = 2p_bq_b\left(\alpha_b\right)^2 + \left(2p_bq_bd_b\right)^2$.

230   **The statistical model for single cross genomic prediction**

231   Assume n_1 and n_2 (several tens) DH or inbred lines from two populations or heterotic groups

232   genotyped for s (thousands) SNPs and the experimental assessment of h (few hundred) single-

233   crosses (h much lower than $n_1.n_2$) in e (several) environments (a combination of growing seasons,

234   years, and locals). Defining y as the adjusted single cross phenotypic mean, the statistical model

235   for prediction of the average effects of SNP substitution and the SNP dominance deviations is

236    $y = M_H + \sum\limits_{r=1}^{s} \left(z_{1_r}\alpha_{SNP1_r} + z_{2_r}\alpha_{SNP2_r} + z_{3_r}d_{SNP_r} \right) + error$

12

237   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

238   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

239   genotypes for the DH or inbred lines are cc (group 1) and cc (group 2), $z_{1_r} = q_{r1}$, $z_{2_r} = -p_{r2}$, and

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

241   and $z_{1_r} = -p_{r1}$, $z_{2_r} = q_{r2}$, and $z_{3_r} = p_{r1}q_{r2}$ if the SNP genotypes for the DH or inbred lines are

242   cc (group 1) and CC (group 2).

243   Regarding the single crosses obtained from DH or inbred lines of the same population or

244   heterotic group we have

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

246   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

247   CC and CC, $z_{1_r} = -2p_r$ and $z_{2_r} = -2p_r^2$ if the SNP genotypes for the two DH or inbred lines are

248   cc and cc, and $z_{1_r} = 2(q_r - p_r)$ and $z_{2_r} = 2p_r q_r$ if the SNP genotypes for the two DH or inbred

249   lines are CC and cc.

250   The statistical problem of genomic prediction when there are a very large number of

251   molecular markers and relatively few observations have been addressed thorough several

252   regularized whole-genome regression and prediction methods (Daetwyler et al. 2013; de Los

253   Campos et al. 2013). Based on one of these approaches, the SNP average effects of substitution and

254   SNP dominance deviations are predicted and used to provide genomic prediction of non-assessed

255   single crosses. The predicted genotypic value for a non-assessed single cross of DH or inbred lines

256   from two groups is

13

257 $$\widetilde{G} = \hat{M}_H + \sum_{r=1}^{s} \left(z_{1_r} \widetilde{\alpha}_{SNP1_r} + z_{2_r} \widetilde{\alpha}_{SNP2_r} + z_{3_r} \widetilde{d}_{SNP_r} \right)$$

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

259   genotypic value is

260 $$\widetilde{G} = \hat{M} + \sum_{r=1}^{s} \left(z_{1_r} \widetilde{\alpha}_{SNP_r} + z_{2_r} \widetilde{d}_{SNP_r} \right)$$

261 **Simulation**

262   The SNP and QTL genotypic data for DH lines, the QTL genotypic data of single crosses, and

263   the phenotypic data for DH lines and single crosses were simulated using the software

264   *REALbreeding*. The program has been developed by the first author using the software *REALbasic*

265   *2009* (Viana et al. 2017a; Viana et al. 2017b; Viana et al. 2016; Azevedo et al. 2015; Viana et al.

266   2013). Based on our input, the software distributed 10,000 SNPs and 400 QTLs in ten

267   chromosomes (1,000 SNPs and 40 QTLs by chromosome). The average SNP density was 0.1 cM.

268   The QTLs were distributed in the regions covered by the SNPs (approximately 100

269   cM/chromosome). Initially, *REALbreeding* sampled 700 DH lines from two non-inbred populations

270   (heterotic groups) in LD (350 from each population). The populations were composites of two

271   populations in linkage equilibrium. In a composite, there is LD only for linked SNPs and QTLs

272   (Viana et al. 2016). The number of DH lines from each S_0 plant was one (scenario 1) or ranged

273   from 1 to 5 (scenario 2). We also sampled 350 DH lines from each population after three

274   generations of selfing (using a single seed descent process). The number of DH lines from each S_3

275   plant ranged from 1 to 5 (scenario 3). For each scenario, the software then crossed 70 selected DH

276   lines from each population, using a diallel design. The heritability for the DH lines was 30%.

277   The genotypic values of the DH lines and of the single crosses were generated assuming a

278   single set of 400 QTLs and two degrees of dominance. To simulate grain yield and expansion

279   volume, a measure of popcorn quality, we defined positive dominance ($0 < (d/a)_i \leq 1.2$, $i = 1, ...,$

280   400) and bidirectional dominance ($-1.2 \leq (d/a)_i \leq 1.2$), respectively, where d/a is the degree of

281   dominance. To compute the genotypic values, *REALbreeding* used our input relative to the

282   maximum and minimum genotypic values for homozygotes. For grain yield and expansion volume,

283   we defined 140 and 30 g/plant and 55 and 15 mL/g, respectively. The phenotypic values were

284   obtained from the sum of the population mean, genotypic value, and experimental error. The error

285   variance was computed from the broad sense heritability. To avoid outliers, we defined the

286   maximum and minimum phenotypic values as 160 and 10 g/plant and 65 and 5 mL/g.

287   The heritabilities for the assessed single crosses were 30, 60, and 100%. Thus, the genotypic

288   value prediction accuracies of the assessed single crosses were 0.55, 0.77, and 1.00, respectively.

289   For each scenario were processed 50 resamplings of 30 and 10% of the single crosses (1,470 and

290   490 assessed single crosses). That is, we predicted 70 and 90% of the single crosses (3,430 and

291   4,410 non-assessed single crosses). Additionally, to assess the relevance of the number of DH lines

292   sampled, we fixed the number of DH lines to achieve the same number of assessed single crosses,

293   using a diallel. That is, we sampled 50 times 38 and 22 DH lines in each group for a diallel

294   (scenario 4), generating 1,444 and 484 single crosses for assessment, respectively. We denote these

295   processes as sampling of single crosses (scenarios 1 to 3) and sampling of DH lines (scenario 4).

296   Other additional scenarios were: genomic prediction of single crosses from selected DH lines from

297   same heterotic group (interestingly for wheat, rice, and barley breeders, for example) (scenario 5)

298   and from selected DH lines from populations with lower LD (scenario 6), to emphasize that the

299   prediction accuracy depends on the LD in the groups of DH or inbred lines. A last scenario

300   (seventh) was genomic prediction of single crosses under an average density of one SNP each cM.

301   This lower density was obtained by random sampling of 100 SNPs per chromosome using a

302   *REALbreeding* tool (*sampler*). To investigate the single cross prediction efficiency based on our

303   model and on the models proposed by Massman et al. (2013) and Technow et al. (2012), we used

304   another *REALbreeding* tool (*Incidence matrix*) to generate the incidence matrices for the three

15

305   models and for the two DH lines sampling processes. To assess the relevance of the SCA effects

306   prediction on genomic prediction of single cross performance, we also fitted the additive model

307   (including only the GCA effects). For comparison purpose, we also processed single cross

308   prediction based on GBLUP (with the observed additive and dominance relationship matrices) and

309   pedigree-based BLUP (with the expected additive and dominance relationship matrices).

310   **Statistical analysis**

311   The methods used for prediction were ridge regression BLUP (RR-BLUP), GBLUP and

312   BLUP. For the analyses we used the *rrBLUP* package (Endelman 2011). The accuracies of single

313   cross genotypic value prediction were obtained by the correlation between the true values of the

314   non-assessed single crosses computed by *REALbreeding* and the values predicted by RR-BLUP,

315   GBLUP, and BLUP. We also computed the efficiency of identification of the 300 non-assessed

316   single crosses of higher genotypic value (coincidence index). The coincidence index was computed

317   from the 300 higher predicted untested single crosses as the number of predicted untested single

318   crosses among the 300 untested single crosses of greater true genotypic value/300. For each DH

319   lines derivation process and heritability, the parametric average coincidence index was computed

320   from the average phenotypic values of the 4,900 single crosses as the number of single crosses

321   among the 300 single crosses of greater true genotypic value/300. Regarding grain yield, for

322   heritability of 30% the coincidence index was 0.2533, 0.2833, and 0.2433 assuming one DH line

323   per S_0 plant, one to five DH lines per S_0 plant, and one to five DH lines per S_3 plant, respectively.

324   The corresponding values for heritability of 60% were, respectively, 0.4800, 0.4900, and 0.4567.

325   Concerning expansion volume, the corresponding values for heritabilities of 30 and 60% were,

326   respectively, 0.2600, 0.2833, and 0.2700, and 0.4733, 0.5100, and 0.4533. The assumed average

327   parametric coefficient index was 0.26 and 0.48 for heritabilities of 30 and 60%, respectively, for

328   both traits. For the population structure analysis we employed *Structure* (Falush et al. 2003) and

329   fitted the no admixture model with independent allelic frequencies. The number of SNPs, sample

16

330   size, burn-in period, and number of MCMC (Markov chain Monte Carlo) replications were 1,000

331   (sampled at random), 140 (70 DH lines from each population), 10,000, and 40,000, respectively.

332   The number of populations assumed (K) ranged from 1 to 4, and the most probable K value was

333   determined based on the inferred plateau method (Viana et al. 2013). The LD analyses were

334   performed with *Haploview* (Barrett et al. 2005).

335   **Data availability**

336   *REALbreeding* is available upon request. The data set is available at

337   https://doi.org/10.6084/m9.figshare.5035130.v3. Data citation:

338   Viana, José Marcelo Soriano; Pereira, Helcio Duarte; Mundim, Gabriel Borges; Piepho, Hans-Peter;

339   Fonseca e Silva, Fabyano (2017): Efficiency of genomic prediction of non-assessed single crosses.

340   figshare. https://doi.org/10.6084/m9.figshare.5035130.v3

341   **RESULTS**

342   The parametric mean and genotypic variance in the populations 1 and 2 were 108.5 and 87.3

343   (g/plant) and 4.7680 and 6.2580 (g/plant)^2, respectivelly. The DH lines derivation processes (one

344   and one to five per S_0 plant and one to five per S_3 plant) provided, for each population, selected DH

345   lines with similar mean (approximately 97 and 76 g/plant for populations 1 and 2), inbreeding

346   depression (approximately −10 and −13% for populations 1 and 2), and genotypic variance

347   (approximately 6 and 7 (g/plant)^2 for populations 1 and 2) and groups of single crosses also similar

348   for mean (approximately 103 g/plant), heterosis (approximately 19%), and genotypic variance

349   (approximately 4 (g/plant)^2). Because we derived one to few DH lines from unrelated S_0 and S_3

350   plants, the average level of relatedness between the selected DH lines was very low (zero and zero,

351   0.0041 and 0.0041, and 0.0054 and 0.0074 assuming one DH line per S_0, one to five DH lines per

352   S_0, and one to five DH lines per S_3, for populations 1 and 2, respectively). Concerning SNP data,

353   the frequency distribution of the minor allele frequency (MAF) and the absolute value of the

354   difference between a SNP allele frequency were also similar for both groups of selected DH lines,

17

355 regardless of the DH line derivation process (Figure 1a, b, c). The average MAF was 0.33,

356 regardless of the population and DH line derivation process. However, the evidence obtained by the

357 population structure analysis was that the DH lines belong to two distinct subpopulations (suggested

358 K equal to 2.4 by the inferred plateau method). The percentages of non-polymorphic SNPs were

359 very low (0.1 to 0.4%). No differences between allelic frequencies were observed for only 1.7 to

360 2.1% of the SNPs. For approximately 70% of the SNPs, the absolute difference between allelic

361 frequencies ranged from 0.1 to 0.6. Regarding LD, for the groups of selected DH lines the evidence

362 based on the analysis of chromosome 1 (no difference between chromosomes is expected) is that

363 LD extents for up to 35 cM, regardless of the DH lines derivation process (Figure 1c, d). Ignoring

364 the non-significant LD values (LOD score lower than 3), for 17 to 20% of the SNP pairs the r^2

365 values ranged from 0.2 to 0.5 (average of 0.16, regardless of the DH lines group and derivation

366 process).

367   Assuming our model, average SNP density of 0.1 cM, training set size of 30%, positive

368 dominance (grain yield), additive-dominance model, and sampling of single crosses, the prediction

369 accuracies of the non-assessed single crosses were greater than the accuracies of the assessed single

370 crosses for low (up to 46% higher) and intermediate (up to 16% higher) heritabilities (Table 1;

371 Figure 2a). As the prediction accuracy of assessed single crosses approaches 1.0, the accuracy of the

372 non-assessed single crosses approaches approximately 0.9 (up to 11% lower). Sampling one to five

373 DH lines per S_3 plant was only slightly superior to the other DH lines derivation processes,

374 regardless of the prediction accuracy of the assessed single crosses (up to 5% higher). Fitting the

375 additive model provided essentially the same prediction accuracies since the maximum decrease

376 was approximately 1%. No significant differences between the prediction accuracies of non-

377 assessed single crosses were also observed assuming bidirectional dominance (expansion volume).

378 The differences compared to positive dominance ranged from approximately −5 to 2%. However, a

379 striking difference was observed between the sampling processes of single crosses for testing.

18

380   Random sampling of single crosses provided higher prediction accuracies of non-assessed single

381   crosses, compared to sampling DH lines for a diallel. The increases in the accuracies by sampling

382   single crosses ranged from approximately 38 to 77%, proportional to the heritability. Decreasing the

383   average SNP density to 1 cM led to a slight decrease in the prediction accuracy of non-assessed

384   single crosses of approximately −4%). Decreasing the training set size to 10% decreased the

385   prediction accuracy of non-assessed single crosses in approximately −5 to −15%, inversely

386   proportional to the heritability. To establish that the prediction accuracy of non-assessed single

387   crosses depends on the level of (overall) LD in the groups of selected DH or inbred lines, we

388   derived DH lines from the same base populations after 10 generations of random crosses (to

389   decrease the LD). The accuracies were also high, ranging from 0.83 to 0.95, proportional to the

390   heritability. The prediction accuracies of non-assessed single crosses from DH lines of the same

391   population were equivalent to the accuracies for single crosses derived from DH lines belonging to

392   distinct heterotic groups, ranging from 0.83 to 0.91, also proportional to the heritability. Comparing

393   our statistical model with the models proposed by Massman et al. (2013) and Technow et al. (2012),

394   we observed no differences for the prediction accuracies of non-assessed single crosses (maximum

395   difference of 1%). Interestingly, the Massman et al. (2013) and Technow et al. (2012) models

396   provide identical accuracies. Finally, no significant differences between the prediction accuracies

397   for RR-BLUP, GBLUP, and BLUP occurred (maximum of 2%), excepting for one to five DH lines

398   per S_3 plant, where BLUP was 9 to 10% inferior, regardless of the heritability.

399   Concerning the coincidence index, in general the inferences are the same established from the

400   prediction accuracy analysis (Table 2; Figure 2b). There were no differences between the

401   coincidence indexes regarding our model and the models proposed by Massman et al. (2013) and

402   Technow et al. (2012) (maximum difference of 3%), and between the RR-BLUP, GBLUP, and

403   BLUP approaches, except for one to five DH lines per S_3 plant, where BLUP was −19 to −27%

404   inferior, proportional to the heritability. The coincidence indexes were also high for single crosses

405 derived from selected DH lines obtained from the base populations with lower LD (ranging from

406 0.55 to 0.76, proportional to the heritability) and from selected DH lines of the same population

407 (ranging from 0.61 to 0.76, also proportional to the heritability). Sampling single crosses for

408 assessment also provided higher coincidence index compared to sampling DH lines for a diallel (39

409 to 98% higher, proportional to the heritability). Decreasing the SNP density and the training set size

410 decreased the coincidence index from 5 to 10% (proportional to the heritability) and from 17 to

411 26% (inversely proportional to the heritability), respectively. The maximum difference in the

412 coincidence index by fitting the additive-dominant and the additive models was −3%. Only for one

413 DH line per S_0 plant the coincidence indexes assuming bidirectional dominance were slightly

414 greater than the values assuming positive dominance (9 to 14% greater). This sampling process of

415 DH lines provided the higher values of coincidence index, compared to the other sampling

416 processes (7 to 26% higher, inversely proportional to the heritability). Finally, the coincidence

417 index of the non-assessed single crosses are greater than the parametric values for all assessed

418 single crosses assuming low (up to 117% higher) and intermediate (up to 39% higher) heritabilities

419 (Table 1). However, as the parametric coincidence of assessed single crosses approaches 1.0, the

420 coincidence values of the non-assessed single crosses approach approximately 0.60 to 0.74 (up to

421 26 to 40% lower), depending on the DH line sampling process.

DISCUSSION

423   Twenty-three years ago, Bernardo (1994) first suggested to use BLUP for predicting untested

424 maize single cross performance. Based on the prediction accuracies obtained by Bernardo (1994,

425 1995, 1996a, 1996b, 1996c), for grain yield and other traits (distinct genetic controls), a breeder

426 should realize that the performance of untested single crosses can be effectively predicted using

427 relationship information from molecular or pedigree data, unbalanced and large data set, and

428 diverse heterotic patterns. The significance of genomic prediction has been confirmed with maize

429 (Zhao et al. 2015) and other important crops, as rice (Xu et al. 2014), wheat (Zhao et al. 2013b) and

430 barley (Philipp et al. 2016), along the last 10 years. Why, then, is there no published evidence that

431 prediction of untested single crosses is of general use by breeders of worldwide seed companies?

432 What should be additionally proved to make prediction of untested single crosses as successful as

433 the Jenkins' (1934) method for predicting double crosses performance was? We believe that this

434 paper offers a significant contribution.

435 Our assessment on efficiency of prediction of untested single cross performance keeps some

436 similarities with few earlier studies but sharp differences for most previous investigations. This

437 study is based on simulated data set, as the study of Technow et al. (2012), assuming 400 QTLs

438 distributed along ten chromosomes. Thus, the prediction accuracies and coincidence indexes (a

439 measure of untested single crosses selection efficiency) are available for non-assessed single crosses

440 since the values were computed based on the true genotypic values of the non-assessed single

441 crosses and not on a cross-validation procedure involving assessed single crosses. This does not

442 mean that we consider simulated data better than field data or have any criticism on the cross-

443 validation procedure. We know that simulated data, because the assumptions, cannot integrally

444 describe the complexity of populations and genetic determination of traits (Daetwyler et al. 2013).

445 To highlight the relevance of (overall) LD, our study is based on scenarios not favorable to

446 prediction of untested single cross performance: very low level of relationship between the DH

447 lines, low and intermediate heritabilities for the assessed single crosses, and not higher heterotic

448 pattern. In the studies of Massman et al. (2013) and Bernardo (1994, 1995, 1996a) the relationship

449 among inbreds from the same heterotic group ranged from 0.11 to 0.58. Riedelsheimer et al. (2012)

450 observed high relationship only between the non-Stiff Stalk inbreds. Technow et al. (2012) assumed

451 non-related inbreds. For most of the investigations on prediction of untested single crosses and

452 testcrosses, the grain yield heritability ranged from 0.72 to 0.88. The common heterotic patterns in

453 these previous studies are Stiff Stalk and non-Stiff Stalk, and Dent and Flint. The MAF in the

454   groups of Dent and Flint inbreds were approximately 0.10 and 0.20, respectively, and

455   approximately 20% of the SNPs showed a difference of allelic frequency of at least 0.6.

456   Concerning the prediction accuracy and the efficiency of identification of the superior 300

457   non-assessed single crosses, our results prove that prediction of untested single crosses is a very

458   efficient procedure (note that we are not saying genomic prediction), especially for low and

459   intermediate heritabilities of the assessed single crosses. The prediction accuracy of the non-

460   assessed single crosses under low (0.55 to 0.71) and intermediate (0.74 to 0.87) accuracies of

461   assessed single crosses achieved 0.85 and 0.89, respectively. It is important to highlight that these

462   are not relative accuracies. Most important, the coincidence of the non-assessed single crosses

463   under low (0.26 to 0.39) and intermediate (0.44 to 0.66) parametric coincidences of assessed single

464   crosses achieved 0.59 and 0.64, respectively. For high heritability (80 to 95%; accuracies from 0.89

465   to 0.97), as observed in most of the studies on prediction of untested single cross performance, we

466   can state (based on values predicted by fitting a quadratic regression model) that the prediction

467   accuracy of non-assessed single crosses is up to only 10% lower (0.87 to 0.92) and, most

468   impressive, the coincidence index can range from 0.61 to 0.71 (parametric coincidences between

469   0.72 to 0.93). Under maximum accuracy of assessed single crosses (1.0), the prediction accuracy

470   and coincidence of non-assessed single crosses achieved 0.93 and 0.76. Thus, assuming high

471   heritability, high density, and training set size of 30%, the accuracy can achieve 0.92 and the

472   efficiency of identification of the best 9% of the non-assessed single crosses can achieve 0.71. It is

473   important to highlight that this efficacy can be higher by using more related DH or inbred lines,

474   under high LD. Thus, we strong recommend that maize breeders, as well as rice, wheat, and barley

475   breeders, make widespread use of prediction of non-assessed single crosses, at least for preliminary

476   screening or prior to field testing.

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

478   breeding programs. However, because breeders are unlikely to rely solely on genomic predictions

479   when selecting superior untested hybrids, Technow et al. (2014) believe that genomic prediction

480   will be combined with field testing of the most promising experimental hybrids. For grain yield, the

481   prediction accuracies observed by Bernardo (1994, 1995, 1996a) ranged from 0.14 to 0.80,

482   proportional to the heritability (in the range 35-74%) and training set size. The non-relative

483   accuracies (relative accuracy x root square of heritability) observed in the studies of Kadam et al.

484   (2016), Technow et al. (2014), Massman et al. (2013), Technow et al. (2012), and Riedelsheimer et

485   al. (2012) ranged between 0.20 and 0.86, also proportional to the heritability (in the range 53-98%)

486   and training set size.

487   We hope that readers of this paper have realized the importance of (overall) LD for effective

488   prediction of non-assessed single crosses, as well as genetic variability (see the parametric accuracy

489   of genomic prediction). Breeders have no control over LD and relatedness between the DH or

490   inbred lines. However, selection should always provide high level of overall LD in the groups of

491   selected DH or inbred lines. Comparison of our LD assessment with the LD analyses from other

492   studies is inadequate because we have distances in cM and not in base-pairs. But in general the level

493   of LD was high (r^2 of approximately 0.3) only for SNPs separated by up to 0.5 Mb (Technow et al.

494   2014; Massman et al. 2013; Technow et al. 2012; Riedelsheimer et al. 2012). To maximize the

495   prediction accuracy and the efficiency of identification of the best non-assessed single crosses it is

496   necessary to adopt the random sampling of single crosses for testing instead of the random sampling

497   of DH or inbred lines for a diallel. This is because sampling 30 or even 10% of the single crosses

498   leads to single crosses for testing derived from all DH or inbred lines from each group. In our case,

499   in every resampling assuming training set size of 30 and 10% we always get groups of assessed

500   single crosses (1,470 and 490 single crosses, respectively) derived from the 70 DH lines of each

501   group. However, sampling DH lines for a diallel provided 1,440 and 484 single crosses for testing

502   derived from 38 and 22 DH lines, respectively. Thus, the sampling of single crosses provides best

503   prediction of the SNP average effects of substitution. Riedelsheimer et al. (2012) emphasized the

23

504   need for large genetic variability to obtain high prediction accuracies. Further, their results indicated

505   that pairs of closely related lines and population structuring only weakly contributed to the high

506   prediction accuracies. Regarding dominance, because it can be a relevant genetic effect, breeders

507   should always fit the additive-dominance model to maximize the prediction accuracy and the

508   efficiency of identification of the best non-assessed single crosses. Interestingly, in most of the

509   studies on prediction of non-assessed single crosses the prediction accuracy did not significantly

510   increase when modeling SCA in addition to GCA effects (Zhao et al. 2015).

511   Concerning SNP density and training set size, factors related with the costs of genotyping and

512   phenotyping, breeders should find a balance between efficiency and expenses, since maximizing

513   SNP density and training set size maximizes the efficiency of untested single cross prediction.

514   Based on our results, because the decreases in the prediction accuracy (approximately 4%) and

515   coincidence index (5 to 10%) by decreasing the average SNP density from 0.1 to 1 cM are of

516   reduced magnitude, we consider sufficient to employ custom genotyping to provide an average SNP

517   density of 1 cM. Decreasing the training set size from 30 to 10% of the single crosses does not

518   significantly affect the prediction accuracy under intermediate to high heritability (decrease of up to

519   9%), but the coincidence index can be reduced in up to 21%. However, considering that the

520   coincidence index will be kept in the range 0.48 to 0.61, proportional to the heritability, and that the

521   maximum values are in the range 0.48 to 0.61, we also consider sufficient to assess at least 10% of

522   the possible single crosses. As highlighted by Zhao et al. (2015), marker density only marginally

523   affects the prediction accuracy of untested single crosses and, for biparental populations, a plateau

524   for the accuracy is reached with a few hundred markers. Technow et al. (2014) did not find an

525   improvement of prediction accuracies by using higher SNP density. Additionally, the increase in the

526   training set size led to a relative small increase in the prediction accuracy. However, the prediction

527   accuracies obtained by Riedelsheimer et al. (2012) under high density (38,019 SNPs) were

528   substantially greater than those reached with a low-density marker panel (1,152 SNPs). In the study

529 of Technow et al. (2012), the prediction accuracies increased with SNP density and number of

530 parents tested in hybrid combination.

531   The DH lines sampling process, the heterotic pattern, and the statistical approach should not

532 be worries for breeders. However, under high heritability notice that sampling more than one DH

533 line per S_0 or S_3 plant provided the higher coincidence values and high prediction accuracy in our

534 study. For rice, wheat, and barley breeders our message is: high prediction accuracy and high

535 efficiency of identification of superior non-assessed single crosses does not depend on heterotic

536 groups but on the (overall) LD in the group or in each group of DH or inbred lines. In other words,

537 the efficiency of prediction of non-assessed single crosses derived from DH or inbred lines from the

538 same population can be as high as the efficiency of prediction of untested single crosses derived

539 from DH or inbred lines from distinct heterotic groups. This is not confirmed comparing the relative

540 prediction accuracies for grain yield of maize untested single crosses (from approximately 0.50 to

541 0.95, for most studies) with those obtained with rice, wheat, and barley untested hybrids (0.50 to

542 0.60, approximately) (Philipp et al. 2016; Xu et al. 2014; Zhao et al. 2013b). However, the lower

543 relative prediction accuracies for untested rice, wheat, and barley hybrids should be due to

544 prediction of two- and three-way crosses. Regarding the statistical approach, our model did not

545 provide an increase in the efficiency of non-assessed single cross prediction, compared to the

546 models proposed by Massman et al. (2013) and Technow et al. (2012). It is important to highlight

547 that our results showed that these two models are really identical (data no shown). Thus, because

548 the simplified definition of the incidence matrices for these two previous models, it is quite safe to

549 use any of them. Finally, the choice between the statistical approaches RR-BLUP (prediction of

550 genotypic values of non-assessed single crosses based on prediction of SNP average effects of

551 substitution), GBLUP (prediction of genotypic values of non-assessed single crosses based on

552 additive and dominance genomic matrices), and BLUP (prediction of genotypic values of non-

553 assessed single crosses based on additive and dominance matrices from pedigree records) is not a

554   serious worry for breeders too. Our evidence is that there is no significant difference between RR-

555   BLUP and GBLUP regarding prediction accuracy and efficiency of identification of the best

556   untested single crosses. Further, even when the level of relatedness between the DH or inbred lines

557   in each group is low, in general pedigree-based BLUP is as efficient as genomic prediction,

558   excepting when the DH lines are derived from inbred population. Thus, DNA polymorphism is not

559   essential for an efficient prediction of non-assessed single cross performance. In a review on

560   genomic selection in hybrid breeding, Zhao et al. (2015) state that the choice of the biometrical

561   model has no substantial impact on the prediction accuracy of untested single crosses. Technow et

562   al. (2014) observed that prediction methods GBLUP and BayesB resulted in very similar prediction

563   accuracies. According to Massman et al. (2013), pedigree-based BLUP and RR-BLUP models did

564   not lead to prediction accuracies that differed significantly. Comparing GBLUP and BayesB,

565   Technow et al. (2012) concluded that the latter method produced significantly higher accuracies for

566   the additive-dominance model.

567   Our main contributions on the assessment of prediction efficiency of untested single cross

568   performance are: 1) the prediction accuracy of untested single crosses ranged from approximately

569   0.80 to 0.90 as the heritability of tested single crosses ranged from low (30%) to high (100%);

570   however, the efficacy of identification of the best 9% of the untested single crosses ranged from

571   approximately 0.50 to 0.70, depending on the DH lines sampling process; 2) the prediction accuracy

572   for crops showing no defined heterotic pattern can be as efficient as with maize, for which there are

573   well defined heterotic groups; this is because the most important factor affecting the prediction

574   efficiency is the overall LD; 3) to maximize prediction accuracy and coincidence the choice of

575   single crosses for testing should be based on a random process; this procedure maximizes the

576   number of DH lines in hybrid combinations and provides better predictions of the SNP average

577   effects of substitution and dominance deviations, compared to sampling DH lines for a diallel; 4)

578   because non significant decreases in the prediction accuracy and coincidence, the prediction of

579   untested single crosses can be efficient assuming reduced training set size (10%) and SNP density

580   of 1 cM; 5) RR-BLUP and GBLUP provide equivalent prediction efficiencies of untested single

581   crosses; 6) excepting for DH lines derived from inbred populations, pedigree-based BLUP is as

582   efficient as genomic prediction of untested single crosses; and 7) the theoretical accuracy shows that

583   the prediction accuracy is not affected by the linkage phase.

588   **LITERATURE CITED**

589   Albrecht, T., H.-J. Auinger, V. Wimmer, J.O. Ogutu, C. Knaak *et al.*, 2014 Genome-based

590   prediction of maize hybrid performance across genetic groups, testers, locations, and years.

591   *Theoretical and Applied Genetics* 127 (6):1375-1386.

592   Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction

593   of testcross values in maize. *Theoretical and Applied Genetics* 123 (2):339-350.

594   Azevedo, C.F., M.D. Vilela de Resende, F. Fonseca e Silva, J.M. Soriano Viana, M.S. Ferreira

595   Valente *et al.*, 2015 Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC*

596   *Genet* 16.

597   Barrett, J.C., B. Fry, J. Maller, and M.J. Daly, 2005 Haploview: analysis and visualization of LD

598   and haplotype maps. *Bioinformatics* 21 (2):263-265.

599   Bernardo, R., 1996a Best linear unbiased prediction of maize single-cross performance. *Crop*

600   *Science* 36: 50-56.

601   Bernardo, R., 1996b Best linear unbiased prediction of maize single-cross performance given

602   erroneous inbred relationships. *Crop Science* 36: 862-866.

603   Bernardo, R., 1996c Best linear unbiased prediction of the performance of crosses between untested

604   maize inbreds. *Crop Science* 36: 872-876.

605   Bernardo, R., 1995 Genetic models for predicting maize single-cross performance in unbalanced

606   yield trial data. *Crop Science* 35: 141-147.

607   Bernardo, R., 1994 Prediction of maize single-cross performance using RFLPs and information

608   form related hybrids. *Crop Science* 34: 20-25.

609   Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey, 2013 Genomic

610   Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and

611   Benchmarking. *Genetics* 193 (2):347-+.

612   de Los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P. Calus, 2013 Whole-

613   genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193

614   (2):327-345.

615   Endelman, J.B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package

616   rrBLUP. *Plant Genome* 4 (3):250-255.

617   Falush, D., M. Stephens, and J.K. Pritchard, 2003 Inference of population structure using multilocus

618   genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567-1587.

619   Jonas, E., and D.J. de Koning, 2013 Does genomic selection have a future in plant breeding? *Trends*

620   *in Biotechnology* 31 (9):497-504.

621   Jenkins, M.T., 1934 Methods of estimating the performance of double crosses in corn. *Journal of*

622   *the American Society of Agronomy* 26:199-204.

623   Kadam, D.C., S.M. Potts, M.O. Bohn, A.E. Lipka, and A.J. Lorenz, 2016 Genomic Prediction of

624   Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. *G3-Genes Genomes*

625   *Genetics* 6 (11):3443-3453.

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

28

627 Li, Z., N. Philipp, M. Spiller, G. Stiewe, J.C. Reif *et al.*, 2017 Genome-Wide Prediction of the
628   Performance of Three-Way Hybrids in Barley. *Plant Genome* 10 (1).

629 Massman, J.M., A. Gordillo, R.E. Lorenzana, and R. Bernardo, 2013 Genomewide predictions from
630   maize single-cross data. *Theor Appl Genet* 126 (1):13-22.

631 Meuwissen, T., B. Hayes, and M. Goddard, 2013 Accelerating Improvement of Livestock with
632   Genomic Selection. *Annual Review of Animal Biosciences, Vol 1* 1:221-237.

633 Philipp, N., G.Z. Liu, Y.S. Zhao, S. He, M. Spiller *et al.*, 2016 Genomic Prediction of Barley
634   Hybrid Performance. *Plant Genome* 9 (2).

635 Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow *et al.*, 2012 Genomic
636   and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* 44
637   (2):217-220.

638 Technow, F., C. Riedelsheimer, T.A. Schrag, and A.E. Melchinger, 2012 Genomic prediction of
639   hybrid performance in maize with models incorporating dominance and population specific
640   marker effects. *Theoretical and Applied Genetics* 125 (6):1181-1194.

641 Technow, F., T.A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome Properties
642   and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize.
643   *Genetics* 197 (4):1343-U1469.

644 Van Eenennaam, A.L., K.A. Weigel, A.E. Young, M.A. Cleveland, and J.C.M. Dekkers, 2014
645   Applied Animal Genomics: Results from the Field. *Annual Review of Animal Biosciences, Vol*
646   *2* 2:105-139.

647 Viana, J.M.S., H.-P. Piepho, and F.F. Silva, 2016 Quantitative genetics theory for genomic
648   selection and efficiency of breeding value prediction in open-pollinated populations. *Scientia*
649   *Agricola* 73 (3):243-251.

29

650 Viana, J.M.S., H.P. Piepho, and F.F. Silva, 2017a Quantitative genetics theory for genomic

651   selection and efficiency of genotypic value prediction in open-pollinated populations. *Scientia*

652   *Agricola* 74 (1):41-50.

653 Viana, J.M.S., F.F. Silva, G.B. Mundim, C.F. Azevedo, and H.U. Jan, 2017b Efficiency of low

654   heritability QTL mapping under high SNP density. *Euphytica* 213 (1).

655 Viana, J.M.S., M.S.F. Valente, F.F. Silva, G.B. Mundim, and G.P. Paes, 2013 Efficacy of

656   population structure analysis with breeding populations and inbred lines. *Genetica* 141 (7-

657   9):389-399.

658 Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink *et al.*, 2012 Effectiveness of

659   Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and

660   Environments. *G3-Genes Genomes Genetics* 2 (11):1427-1436.

661 Xu, S., D. Zhu, and Q. Zhang, 2014 Predicting hybrid performance in rice using genomic best linear

662   unbiased prediction. *Proceedings of the National Academy of Sciences of the United States of*

663   *America* 111 (34):12456-12461.

664 Zhao, Y., M. Gowda, W. Liu, T. Wuerschum, H.P. Maurer *et al.*, 2013a Choice of shrinkage

665   parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant*

666   *Breeding* 132 (1):99-106.

667 Zhao, Y., M.F. Mette, and J.C. Reif, 2015 Genomic selection in hybrid breeding. *Plant Breeding*

668   134 (1):1-10.

669 Zhao, Y., J. Zeng, R. Fernando, and J.C. Reif, 2013b Genomic Prediction of Hybrid Wheat

670   Performance. *Crop Science* 53 (3):802.

671

672 **Table 1** Average prediction accuracies of non-assessed single crosses and its standard deviation,

673 assuming single crosses from selected DH lines, 30 and 10% of assessed single crosses, two traits

674 (grain yield - GY, g/plant, and expansion volume - EV, mL/g), two sampling processes of single

675 crosses, four statistical models, three DH lines sampling processes, two genetic models, and three

676 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
			$1\text{-}5/S_0$	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
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_3$	AD[a]	0.7718 ± 0.0161	0.8371 ± 0.0079	0.8888 ± 0.0043
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.7982 ± 0.0140	0.8622 ± 0.0055	0.8973 ± 0.0025
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.7973 ± 0.0124	0.8574 ± 0.0070	0.8978 ± 0.0019
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.7803 ± 0.0143	0.8435 ± 0.0074	0.8830 ± 0.0024
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.4827 ± 0.0423	0.5000 ± 0.0420	0.5036 ± 0.0465
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.8019 ± 0.0155	0.8656 ± 0.0050	0.9055 ± 0.0020
			$1\text{-}5/S_3$	AD	0.7589 ± 0.0143	0.8424 ± 0.0058	0.9165 ± 0.0027

[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..

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 (grain yield - GY, g/plant, and expansion volume - EV, mL/g), two sampling processes of single crosses, four statistical models, three DH lines sampling processes, two genetic models, and three parametric coincidence of assessed single crosses

Trait	Samp. proc.	Statistical model	DH lines	Gen. mod.	Coincidence of assessed single crosses		
					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
			$1\text{-}5/S_0$	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
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_3$	AD[a]	0.4881 ± 0.0278	0.5691 ± 0.0229	0.6620 ± 0.0215
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.5651 ± 0.0310	0.6431 ± 0.0164	0.6955 ± 0.0144
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.5650 ± 0.0280	0.6384 ± 0.0204	0.6849 ± 0.0137
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.5531 ± 0.0323	0.6272 ± 0.0194	0.6699 ± 0.0130
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.3268 ± 0.0642	0.3400 ± 0.0691	0.3461 ± 0.0728
			$1\text{-}5/S_3$	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
			$1\text{-}5/S_0$	AD	0.5506 ± 0.0285	0.6337 ± 0.0203	0.6944 ± 0.0141
			$1\text{-}5/S_3$	AD	0.4746 ± 0.0294	0.5843 ± 0.0174	0.7141 ± 0.0171

[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..

(a)　　　　　　　　(b)　　　　　　　　(c)

(d)　　　　　　　　(e)
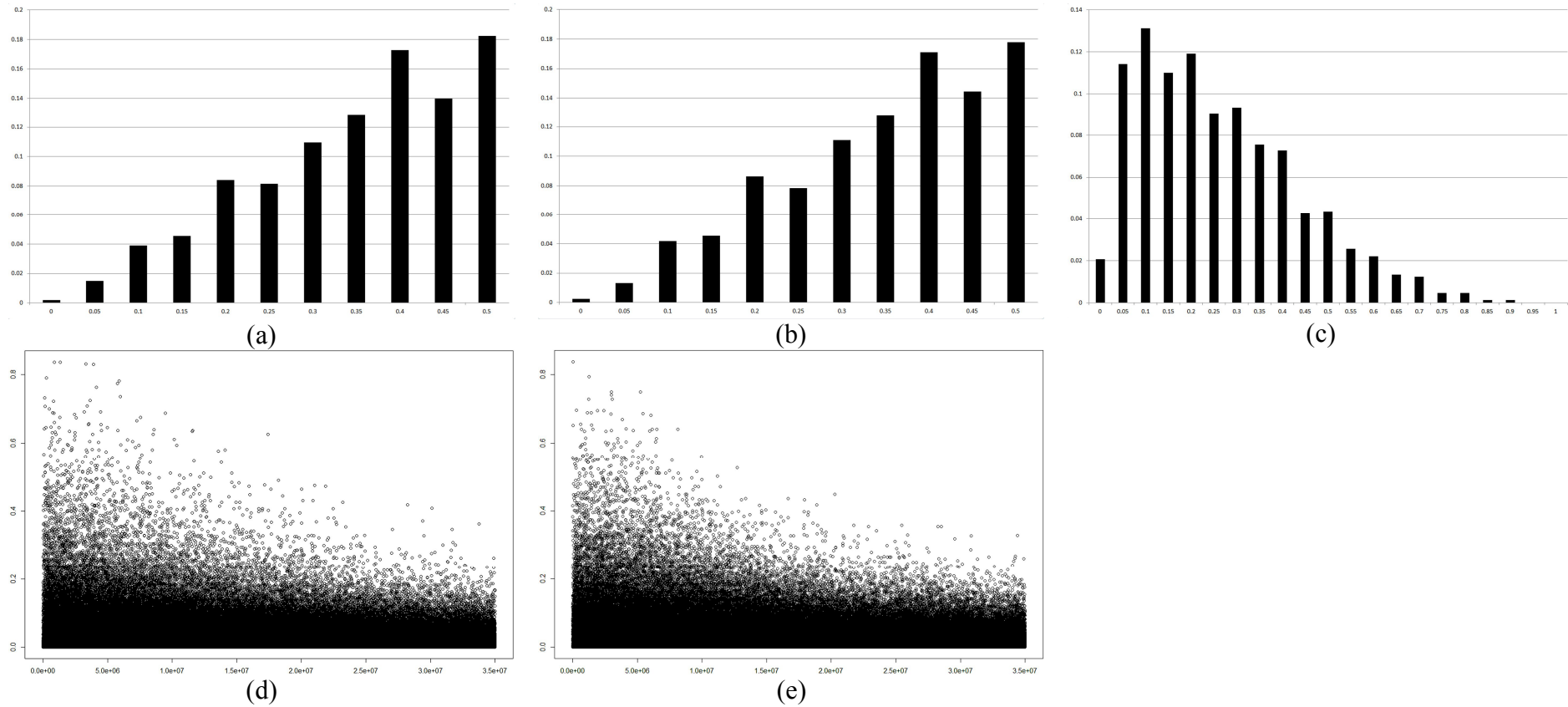
Figure 1 Frequency distribution of the MAF in the groups of selected DH lines (a and b) and the absolute value of the difference between a SNP allele frequency (c), and LD (r^2) in relation to distance (cM) in the two groups of selected DH lines (d and e), regarding SNPs in chromosome 1 separated by zero to 35 cM, assuming one DH line per S_0 plant.
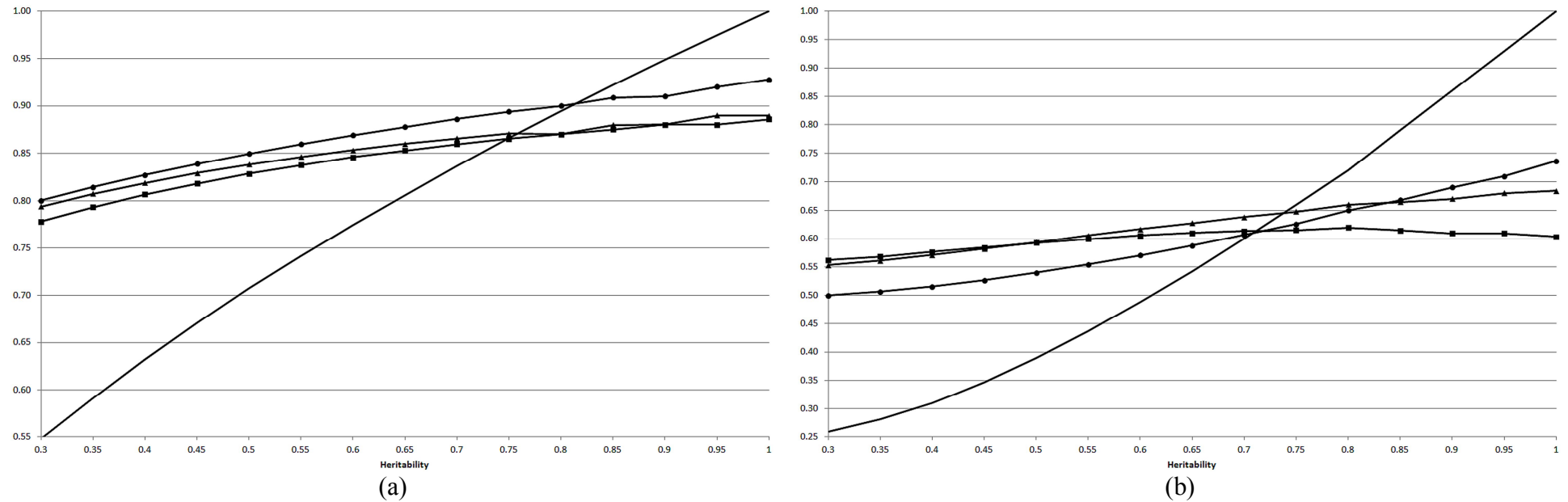
Figure 2 Predicted accuracies (a) and coincidence indexes (b) for untested single crosses (square: $1/S_0$; triangle: $1-5/S_0$; circle: $1-5/S_3$), and parametric accuracies and coincidence indexes for tested single crosses (continuous line), assuming our model, average SNP density of 0.1 cM, training set size of 30%, positive dominance (grain yield), additive-dominance model, and sampling of single crosses.