Mol Breeding (2016) 36:113
DOI 10.1007/s11032-016-0504-9

Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.)

Yansong Ma · Jochen C. Reif · Yong Jiang · Zixiang Wen · Dechun Wang · Zhangxiong Liu · Yong Guo · Shuhong Wei · Shuming Wang · Chunming Yang · Huicai Wang · Chunyan Yang · Weiguo Lu · Ran Xu · Rong Zhou · Ruizhen Wang · Zudong Sun · Huaizhu Chen · Wanhai Zhang · Jian Wu · Guohua Hu · Chunyan Liu · Xiaoyan Luan · Yashu Fu · Tai Guo · Tianfu Han · Mengchen Zhang · Bincheng Sun · Lei Zhang · Weiyuan Chen · Cunxiang Wu · Shi Sun · Baojun Yuan · Xinan Zhou · Dezhi Han · Hongrui Yan · Wenbin Li · Lijuan Qiu

Received: 16 December 2015 / Accepted: 2 June 2016 / Published online: 28 July 2016

Abstract Genomic selection is a promising molecular breeding strategy enhancing genetic gain per unit time. The objectives of our study were to (1) explore the prediction accuracy of genomic selection for plant height and yield per plant in soybean [*Glycine max* (L.) Merr.], (2) discuss the relationship between prediction accuracy and numbers of markers, and (3) evaluate the effect of marker preselection based on different methods on the prediction accuracy. Our study is based on a population of 235 soybean varieties which were evaluated for plant height and yield per plant at multiple locations and genotyped by 5361 single nucleotide polymorphism markers. We applied ridge regression best linear unbiased prediction coupled with fivefold cross-validations and evaluated three strategies of marker preselection. For plant

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0504-9) contains supplementary material, which is available to authorized users.

Y. Ma · W. Li
College of Agriculture, Northeast Agricultural University, Harbin 150030, China

Y. Ma · Z. Liu · Y. Guo · L. Qiu (✉)
The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: qiulijuan@caas.cn

Y. Ma · X. Luan
Soybean Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China

J. C. Reif · Y. Jiang
Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany

Z. Wen · D. Wang
Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

T. Han · C. Wu · S. Sun
Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

S. Wei
Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China

S. Wang · C. Yang
Soybean Research Institute, Jilin Academy of Agricultural Sciences, Changchun 130033, China

H. Wang
Chifeng Institute of Agricultural Sciences, Chifeng 024031, China

height, marker density and marker preselection procedure impacted prediction accuracy only marginally. In contrast, for grain yield, prediction accuracy based on markers selected with a haplotype block analyses-based approach increased by approximately 4 % compared with random or equidistant marker sampling. Thus, applying marker preselection based on haplotype blocks is an interesting option for a cost-efficient implementation of genomic selection for grain yield in soybean breeding.

Keywords Genomic selection · Prediction accuracy · *Glycine max* · Sampling method

Abbreviations

GS	Genomic selection
SNP	Single nucleotide polymorphism
rrBLUP	Ridge regression best linear unbiased prediction
RSM	Random sampling method
HBA	Haplotype block analysis
ESM	Evenly sampling method

Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most important sources of oil and plant protein (Masuda and Goldsmith 2009). Substantial genetic improvements are required for both traits to feed an estimated world population of 9 billion by 2050 (Ray et al. 2013). Genomic selection (GS) is a novel breeding tool accelerating the selection gain per time unit. GS was initially used for animal breeding (Meuwissen et al. 2001), and its potential is currently intensively studied in plant populations (Heffner et al. 2009; Jannink et al. 2010; Nakaya and Isobe 2012). These experimental studies included data of many major crops such as barley (Zhong et al. 2009), wheat (Rutkoski et al. 2011; Zhao et al. 2015; Pérez-Rodríguez et al. 2012; Crossa et al. 2014), maize (Zhao et al. 2012a, b; Bernardo 2013, 2014), rice (Spindel et al. 2015), sunflower (Reif et al. 2013), forage plants (Hayes et al. 2013), sugar beet (Wurschum et al. 2013), and soybean (Bao et al. 2014; Shu et al. 2013). All studies underline the potential of genomic selection as a powerful tool to accelerate selection gain in plant breeding.

Information on the level of prediction accuracy of genomic selection is crucial to integrate this new tool into applied plant breeding programs. GS prediction accuracy is affected by many factors (Zhong et al. 2009; Calus et al. 2008; Solberg et al. 2008; Zhao et al. 2012a, b; Habier et al. 2007). Thereby, the number of markers is one factor to successfully integrate GS in applied plant breeding programs. A high number of markers facilitate to capture most of the linkage

C. Yang · M. Zhang
Institution of Cereal and Oil Crops Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang 050031, China

W. Lu
Economic Crops Institute, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China

R. Xu
Crop Research Institute, Shandong Academy of Agricultural Sciences, Jinan 250010, China

R. Zhou · X. Zhou
Oil Crop Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China

R. Wang
Institute of Crop Sciences, Jiangxi Academy of Agricultural Sciences, Nanchang 330200, China

Z. Sun · H. Chen
Institute of Economical Crops, Guangxi Academy of Agricultural Sciences, Nanning 530007, China

W. Zhang · B. Sun
Hulun Buir Institute of Agricultural Sciences, Hulun Buir 021000, China

J. Wu · D. Han · H. Yan
Heihe Branch Institute, Heilongjiang Academy of Agricultural Sciences, Heihe 164300, China

G. Hu · C. Liu
The Crop Research and Breeding Center of Land-Reclamation, Harbin 150090, Heilongjiang, China

Y. Fu · W. Chen
Suihua Branch Institute, Heilongjiang Academy of Agricultural Sciences, Suihua 152052, China

T. Guo
Jiamusi Branch Institute, Heilongjiang Academy of Agricultural Sciences, Jiamusi 154007, China

L. Zhang
Crop Institute, Anhui Academy of Agricultural Sciences, Hefei 230031, Anhui, China

information between QTL and SNP (Solberg et al. 2008; Meuwissen et al. 2001). Nevertheless, large number of markers increases costs and more importantly can create problems due to collinearity among markers. Moreover, as GS also exploits relatedness (Habier et al. 2007, 2010), it is pivotal to have a balanced set of markers allowing to portray reliably the relationship matrix (Liu et al. 2015; Habier et al. 2010).

Soybean is suitable for genomic selection because of moderated genome size and rapid progress on soybean genome sequencing (Schmutz et al. 2010) and re-sequencing (Lam et al. 2010; Li et al. 2013). Moreover, SNP markers have been developed which are distributed throughout the soybean genome (Song et al. 2013) accelerating the application of GS. Shu et al. (2013) used 288 soybean varieties and 79 sequence-characterized amplified region (SCAR) markers and illustrated the potential of whole-genome prediction of hundred-seed weight. Bao et al. (2014) used 282 elite soybean lines, which were fingerprinted with 1536 single nucleotide polymorphism (SNP) markers, and highlighted the prospective of genomic selection for improving resistance to soybean cyst nematode (SCN). All previous research showed that genomic selection was an effective procedure in soybean breeding. However, results on genomic selection in soybean on complex traits such as yield are to the best of our knowledge still missing.

The objectives of this study were to apply ridge regression best linear unbiased prediction in a population of 235 soybean varieties fingerprinted with 5361 genome-wide distributed SNPs in order to (1) explore the genomic prediction accuracy for plant height and yield per plant, (2) discuss the relationship between prediction accuracy and numbers of markers, and (3) evaluate the effect of marker preselection based on different methods on the prediction accuracy.

Materials and methods

Field trials

Our study comprised phenotypic data of 235 soybean varieties provided by the National Key Facility for Crop Gene Resources and Genetic Improvement (NFCIR), Institute of Crop Science, Chinese Academy of Agricultural Science. Out of the 235 varieties, 185 were North Spring soybean (NSs) and 50 HuangHuai summer soybean (HHSs) lines. The 235 varieties were evaluated in replicated field trials in 23 locations in Northeast China and in the HuangHuai region in the year 2011 (Supplementary Table S1). The experimental designs were randomized complete block designs with two replications. Plots consisted of three rows with 3 m in length and 0.2 m apart. Fertility and pest management were performed following standard management recommendations. Plant height (cm) and yield per plant (g) were determined in each location following standard protocols (Qiu et al. 2006).

Phenotypic data analyses

Variance components and heritability of plant height and yield per plant were estimated using the lme4 package implemented in the software package R (Bates et al. 2014). The following mixed linear model was fitted:

$$y_{ij} = \mu + L_i + G_j + e_{ij},$$

where y_{ij} is the average phenotypic value for ith line at jth location, μ is the population mean, L_i and G_j refer to the effect of jth location and ith line, respectively, and e_{ij} denotes the random residual term. Variance components were estimated assuming random location and genotype effects. The best linear unbiased estimation (BLUE) of each line was determined using the same model mentioned above by assuming fixed genotypic effect and random location effects. The difference of target traits average between NSs subsets and HHSs subsets was evaluated applying a t test using PASW statistics.

Genotypic data and linkage disequilibrium analysis

The 235 soybean lines were genotyped with Illumina SoySNP 6 k iSelect BeadChip which comprised 5361 SNPs. These SNPs were chosen from the Illumina SoySNP 50 k iSelect BeadChip (Illumina, San Diego, USA) (Song et al. 2013). We selected SNPs that were located in the proximity of previously described QTLs for various traits. Genotypes are called using the program GenomeStudio (Illumina, San Diego, USA). SNPs with proportion of missing data exceeding 10 % were excluded. For the remaining SNPs, missing

B. Yuan
Zhoukou Institute of Agricultural Sciences, Zhoukou 466001, Henan, China

values were imputed (Poland et al. 2012). Minor allele frequency (MAF) and polymorphism information content (PIC) were estimated using software PowerMarker version 3.0 (http://www.powermarker.net). Linkage disequilibrium parameter (r^2) between SNP pairs was estimated using the statistical software R (Team 2014) (https://www.r-project.org/). Decay of linkage disequilibrium was explored based on the data of estimated r^2 against genetic distance for all SNP pairs, by fitting a curve with the locally weighted polynomial regression method (Cleveland 1979). To evaluate the population structure, principal component analysis (PCA) was performed using genotypic data. PCA was completed using software TASSEL 3.0 (http://www.maizegenetics.net/). The first two principal components were used to examine the presence of subpopulation structure.

Genomic selection and cross-validation

The potential of genomic selection was examined focusing on ridge regression best linear unbiased prediction (RR-BLUP) implemented in the statistical package "rrBLUP" (Endelman 2011). Let n be the number of genotypes and p be the number of markers. The RR-BLUP model has the form, where y is the vector of BLUEs of genotypic values obtained in the phenotypic data analyses, μ refers to the overall mean, α is the vector of additive effects of markers, $X = (x_{ij})$ is the $n \times p$ matrix of markers with x_{ij} being the number of a chosen allele at the jth locus for the ith genotype, and e is the vector of residual terms. In the model, we assumed that marker and residual effects are randomly distributed with $\alpha \sim N(0, I_p\alpha_\alpha^2)$ and, where I_p and I_n denote identity matrices with respective dimensions, $\alpha_\alpha^2 = \alpha_G^2/p$ and note that α_G^2 and α_e^2 were the estimated genotypic and residual variance components in the phenotypic data analyses, and l refers to the number of locations.

We evaluated the prediction accuracy of genomic selection applying fivefold cross-validations. Marker effects were estimated in the training population and the effects were used to predict the genotypic values in the test population. The Pearson product-moment correlation coefficient between the predicted and observed phenotype (r_{MP}) was estimated, and prediction accuracy (r_{GS}) was calculated by standardizing r_{MP} by the square root of the broad-sense heritability. We repeated the procedure 500 times to reduce the sampling error. In addition, we examined the prediction accuracy also within the North Spring soybean (NSs) subpopulation contrasting it with a random subset of the total population with the same sample size.

Sampling strategy of markers

Random sampling method (RSM)

We randomly sampled SNPs to form different subsets. The number of sampled SNPs varied from 5 to 100 % of the total number of SNPs using five percent intervals. Fivefold cross-validation was applied to study the accuracy of genomic selection with the different subsets. 500 replicates were explored to eliminate sampling error.

Haplotype block analysis (HBA)

Haplotype analysis was completed using Haploview 4.2 software based on the population of all 235 soybean lines. Haplotype blocks were defined following previous suggestions (Gabriel et al. 2002). The 5361 SNPs were classified after haplotype block analysis into SNPs belonging to haplotype blocks and SNPs not forming haplotype blocks. We selected then randomly one SNP per haplotype block plus SNPs not forming haplotype blocks. This data were then again used in combination with fivefold cross-validation to study the accuracy of genomic selection. 500 replicates were explored to eliminate sampling error.

Evenly sampling method (ESM)

The same numbers of SNPs as used in the haplotype block analyses were selected evenly according to their position around genome. Fivefold cross-validation and 500 replicates were explored to evaluate the prediction accuracy of target traits according to previous scenarios.

Results

Extensive phenotyping revealed large genetic variation for plant height and grain yield

We observed for both traits, plant height and grain yield per plant, a significant ($P < 0.01$) and broad

genetic variation for the assayed 235 soybean varieties. Lines belonging to the HuangHuai summer group (HHSs) displayed significantly ($P < 0.01$) higher plant height and larger grain yield per plant as compared to North Spring (NSs) lines (Table 1). Heritability estimates of plant height and yield per plant amounted to 0.96 and 0.63, respectively, (Table 1).

Analysis of linkage disequilibrium identified haplotype blocks comprising up to 22 SNPs

Linkage disequilibrium between pairs of SNPs declined sharply to $r^2 = 0.1$ at around 1000 kb (Fig. 1). We identify 357 haplotype blocks across the 20 soybean chromosomes, which comprised a total of 2164 SNPs. The remaining 3197 SNPs, which were not forming haplotype blocks, were defined as "SNPs". The number of SNPs composing haplotype blocks ranged from 2 to 22 and the percentage of SNPs assigned to haplotype blocks in every chromosome ranged from 1.28 % (chromosome 1) to 67.31 % (chromosome 9), respectively, (Fig. 2).

Population structure analysis revealed presence of genetically distinct subpopulations

After quality filtering, 5275 SNPs were used to explore the population structure of the 235 soybean varieties. The minor allele frequency averaged 0.25 (Fig. 3a) and PIC values averaged 0.27 (Fig. 3b). The first two principle components explained in total 17 % of the molecular variation. The scatter plot using the first two principle components revealed presence of two genetically distinct subpopulations (Fig. 4). Soybean varieties of different ecotypes were separated into two subsets according to the first principle component.

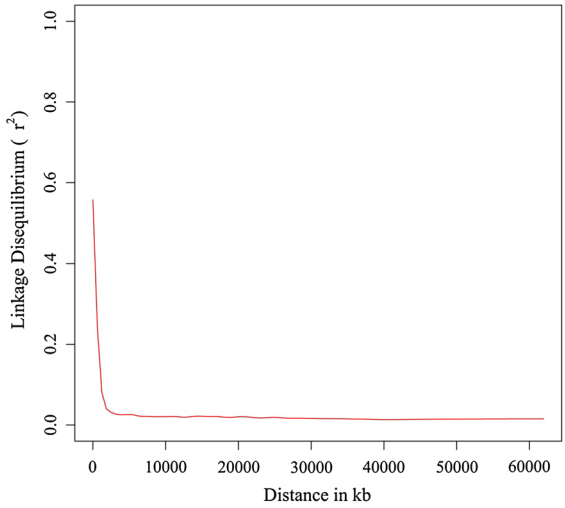

Fig. 1 Decay of linkage disequilibrium (r^2) with physical map distances between markers. The curve was fitted using locally weighted polynomial regression

Genomic prediction accuracies were high for plant height and moderate for grain yield

We used fivefold cross-validation to examine the potential of genome-wide prediction for different soybean traits. The average prediction accuracy was substantially higher for plant height ($r_{GS} = 0.86$) compared to yield per plant ($r_{GS} = 0.47$) (Fig. 5, Table S2). Moreover, the standard deviation of the prediction accuracies was substantially larger for yield per plant compared to plant height (Fig. 5).

Preselection of markers slightly enhanced genomic prediction accuracy for grain yield

We studied the effects of different marker sampling strategies on genomic prediction accuracy for a broad range of marker densities. The marker sampling

Table 1 Genetic variance, broad-sense heritability and contrast of plant height (cm) and yield per plant (g) performances between two subpopulations reflecting different ecotypes

Trait	Genetic variance	Heritability	Mean ± SD		*t* value
			NSs[a]	HHSs[b]	
Plant height	253.33**	0.96	60.26 ± 1.1450	92.37 ± 2.4931	−12.66**
Yield per plant	10.80**	0.63	20.94 ± 0.3289	25.42 ± 0.5174	−6.71**

** Significantly different at 0.01 level probability

[a] North Spring soybean

[b] HuangHuai Summer soybean

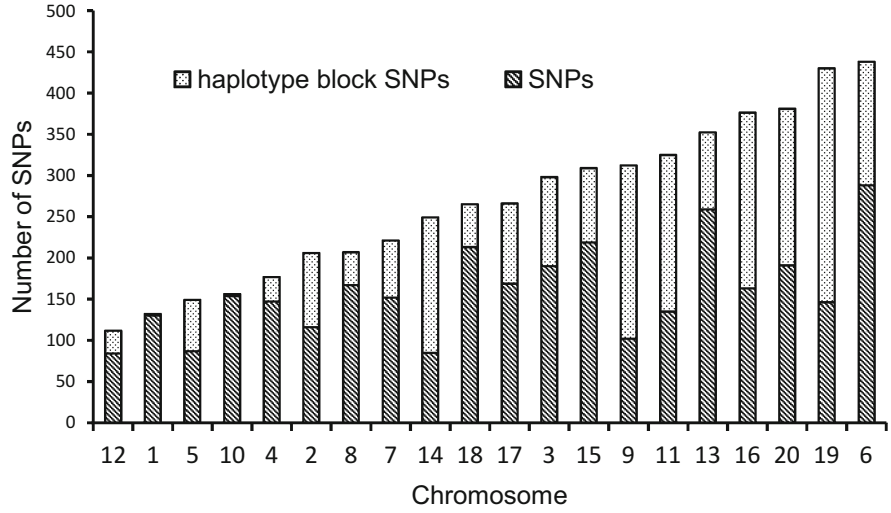

Fig. 2 Distributions of haplotype block SNPs and SNPs for the 20 soybean chromosomes

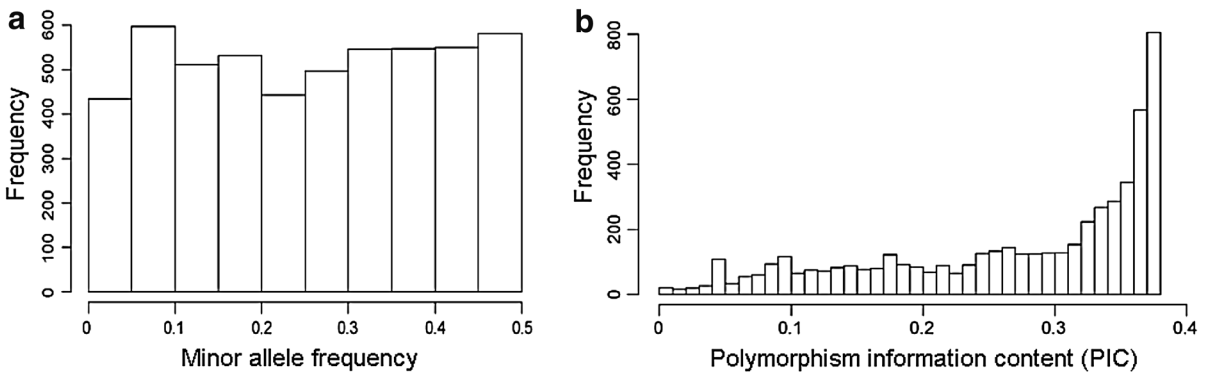

Fig. 3 **a** Histogram of minor allele frequency and **b** polymorphism information content of 5275 SNPs

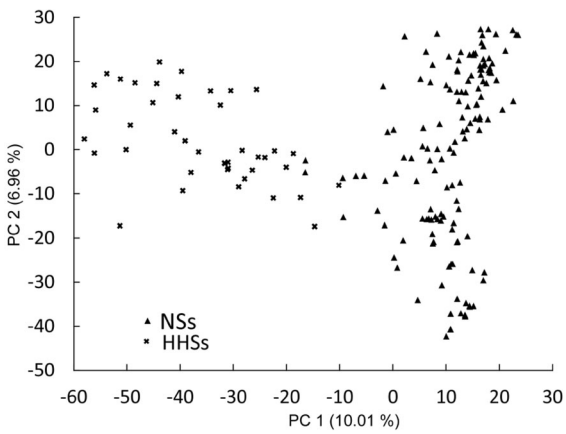

Fig. 4 Scatter plots of the first two principal components (PC) for 235 soybean varieties clustered into North Spring soybean (NSs) and Huanghuai Summer soybean (HHSs) subpopulations

strategies were a random sampling method (RSM), a haplotype block analysis-based sampling (HBA), and evenly sampling method (ESM). Using a step of 250 SNPs, 265 to 5015 SNPs were randomly selected for RSM in order to estimate the prediction accuracies (Supplementary Table S2). In contrast, for HBA we selected one SNP for each of the 357 identified haplotype blocks. These SNPs were combined with the remaining 3197 "SNPs". From this data set, we randomly selected 172 to 3554 SNPs with a step of 178 SNPs and examined the prediction accuracy for the target traits (Supplementary Table S2). We also selected from 172 to 2664 SNPs evenly around genome with a step of 178 SNPs for ESM strategy and evaluated the prediction accuracies (Supplementary Table S2). Generally, prediction accuracies for

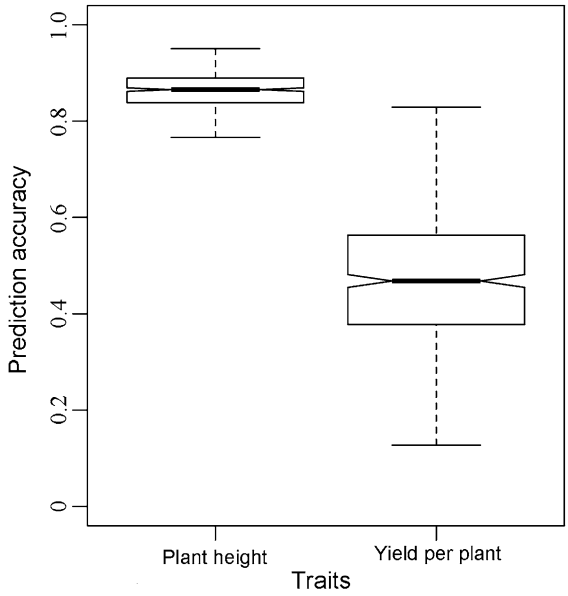

Fig. 5 Box-Whisker plots of cross-validated prediction accuracies of plant height and yield per plant, with the method of ridge regression best linear unbiased prediction

both plant height and yield per plant increased with increasing number of SNPs for both sampling strategies (Fig. 6, Supplementary Table S2). Haplotype block analysis-based sampling facilitated highest prediction accuracies for both target traits. Randomly sampling method improved the prediction accuracy slightly compared with ESM. For yield per plant, prediction accuracy based on markers selected with HBA increased by 3.66 and 4.10 % compared with the RSM and ESM strategies, respectively. In contrast, for plant height, prediction accuracies were comparable for all marker selection strategies.

Discussion

Population structure impaired the prediction accuracy depending on the target trait

Pronounced population structure has to be considered when evaluating the potential of genomic selection (Hayes et al. 2009; Guo et al. 2014; Isidro et al. 2015). In our study, a total of 235 soybean varieties were sampled reflecting two distinct ecotypes (Fig. 4). Consequently, prediction accuracies within the subpopulations of the two distinct ecotypes are potentially overestimated using cross-validations based on the total population. To study this in more detail, we also estimated the prediction accuracies within the larger subpopulation North Spring (NSs) comprising 185 lines. We found that prediction accuracies decreased by 5.27 and 67.07 % for plant height and yield per plant, respectively, using the North Spring soybean subset compared to the total population using a standardized training population size. Consequently, the population structure substantially influenced the prediction accuracy for yield per plant and has to be considered when interpreting the results. If the wish is to develop soybean varieties for breeding programs specifically designed for the North Spring target environments, the prediction accuracies for yield per plant are upward biased. In contrast, plant height is not affected by subpopulation structure, and thus results of the total population are also applicable for breeding programs specifically targeting North Spring environments.

Genomic selection is a promising tool for soybean breeding

As important agronomic traits, the prediction accuracies of plant height and yield were explored in maize (Zhao et al. 2012a; Riedelsheimer et al. 2012; Crossa et al. 2013), wheat (Heffner et al. 2011; Poland et al. 2012), rye (Wang et al. 2014), barley (Sallam et al. 2015), and rice (Spindel et al. 2015). The previously reported prediction accuracies ranged from 0.34 to 0.85 for plant height and from 0.17 to 0.87 for yield. Our results with prediction accuracies of 0.87 for plant height and 0.49 for yield per plant (Fig. 5) are lying within the range of these previously reported values. The higher prediction accuracies for plant height as compared to yield can be explained by a less complex genetic architecture of plant height than yield (Heffner et al. 2011; Spindel et al. 2015; Sallam et al. 2015).

Different strategies completely or partially relying on genomic selection have been proposed to be implemented into breeding programs (Longin et al. 2015; Bassi et al. 2016). The choice of the most suited strategy thereby depends on the prediction accuracy achieved by the genomic selection models. At early selection stages, many individuals are commonly evaluated at a limited number of locations focusing on negative selection, i.e., disregarding the inferior genotypes (He et al. 2016). Genomic selection is for this early selection stages an interesting alternative if costs of genotyping are comparable to the costs of a

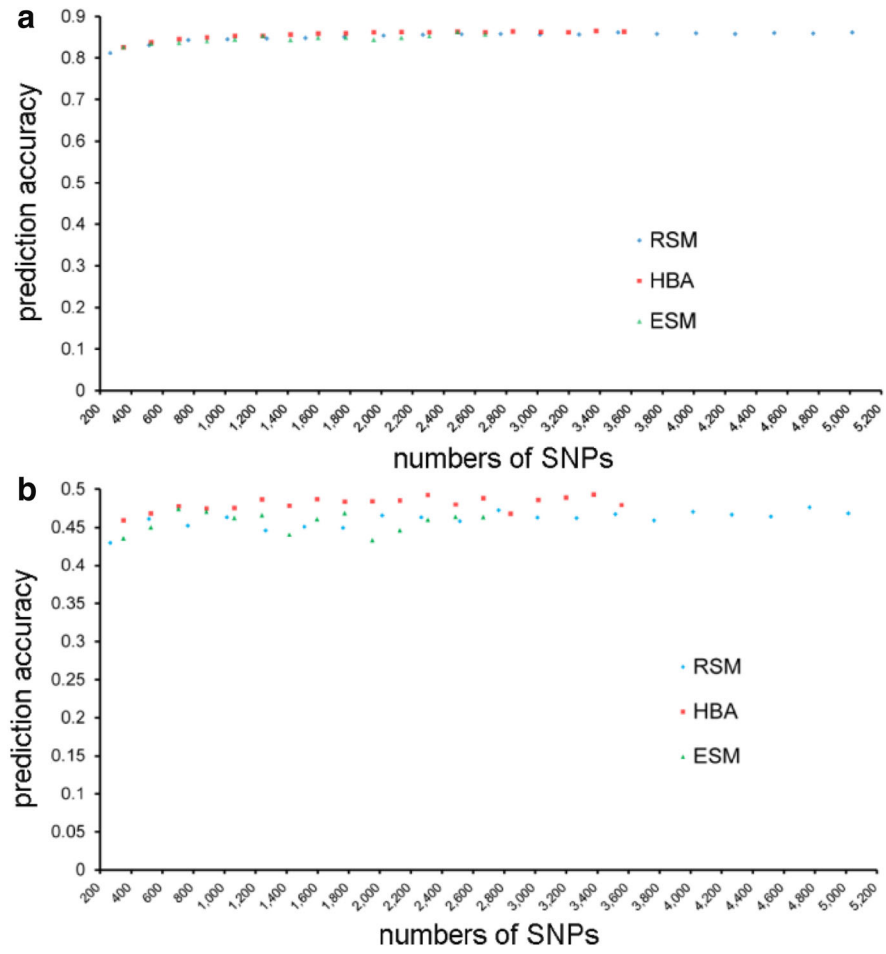

Fig. 6 Cross-validated prediction accuracies of ridge regression best linear unbiased prediction based on three marker sampling strategies for plant height (**a**) and yield per plant (**b**). Marker subsets were selected using a random sampling (RSM), a haplotype block-based sampling strategy (HBA), and evenly sampling method (ESM)

single location yield trial (Heffner et al. 2010). We observed for grain yield a prediction accuracy of 0.47 in our study corresponding to field trials conducted at 3–4 locations (Supplementary Table S2, Fig. 5). Consequently, genomic selection is for yield per plant an interesting alternative for negative selection, thus, replacing early stages of selection in soybean breeding. This trend of favoring genomic selection for negative selection of grain yield has been also observed for other crops such as wheat (He et al. 2016).

Breeding programs exclusively based on genomic predictions focusing also on positive selection, i.e., identifying the best genotype, were only recommended if high prediction accuracies can be achieved by the genomic selection models (Longin et al. 2015). The observed prediction accuracy for plant height amounted to 0.86 in our study (Supplementary Table S2, Fig. 5). Thus, plant height can be reliably predicted based on genomic selection alone.

Effects of marker sampling strategy on genomic prediction accuracies

Meuwissen (Meuwissen 2009) showed in a simulation study that to take advantages of high marker densities, comprehensive training data sets exhibiting a large effective population size are required. Elite soybean breeding populations, however, display often a limited effective population size (St Martin 1982). In this case, marker density may be reduced with only marginal loss in prediction accuracies for an economic implementation of genomic selection. We compared in our study different strategies to reduce the marker density. Our findings show that the marker sampling strategy impacted the prediction accuracies only marginally for plant height (Fig. 6a). In contrast, for grain yield, prediction accuracies based on markers selected with HBA increased by approximately 4 % compared with the two alternative strategies examined in our study

(Fig. 6b). Thus, applying marker preselection based on haplotype blocks is an interesting option for a cost-efficient implementation of genomic selection for grain yield in soybean breeding.

Acknowledgments This work was supported by The 13th Five-Year National Breeding Program for Precise Identification and Germplasm Enhancement of Economic Crops; Plant Gemplasm Conservation of the Chinese Ministry of Agriculture [NB06-070401-(22-27)-05; NB07-2130135-(25-30)-06; NB08-2130135-(25-31)-06; NB2010-2130135-25-05], The National Transgenic Major Program of China (2014ZX08004001), and The Agricultural Science and Technology Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences.

References

Bao Y, Vuong T, Meinhardt C, Tiffin P, Denny R, Chen S, Nguyen HT, Orf JH, Young ND (2014) Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance. Plant Genome. doi:10.3835/plantgenome2013.11.0039

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242:23–36. doi:10.1016/j.plantsci.2015.08.021

Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, Dai B, Eigen C, Rcpp L (2014) Package 'lme4'. R Foundation for Statistical Computing, Vienna

Bernardo R (2013) Genomewide markers as cofactors for precision mapping of quantitative trait loci. Theor Appl Genet 126(4):999–1009. doi:10.1007/s00122-012-2032-2

Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54(1):68–75. doi:10.2135/cropsci2013.05.0315

Calus MP, Meuwissen TH, de Roos AP, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1):553–561. doi:10.1534/genetics.107.080838

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am stat Assoc 74(368):829–836

Crossa J, Beyene Y, Kassa S, Perez P, Hickey JM, Chen C, de los Campos G, Burgueno J, Windhausen VS, Buckler E, Jannink JL, Lopez Cruz MA, Babu R (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 3(11):1903–1926. doi:10.1534/g3.113.008227

Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Ceron-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112(1):48–60. doi:10.1038/hdy.2013.16

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3):250–255. doi:10.3835/plantgenome2011.08.0024

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296(5576):2225–2229. doi:10.1126/science.1069424

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127(3):749–762. doi:10.1007/s00122-013-2255-x

Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397. doi:10.1534/genetics.107.081190

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42(1):5. doi:10.1186/1297-9686-42-5

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41(1):51. doi:10.1186/1297-9686-41-51

Hayes BJ, Cogan NOI, Pembleton LW, Goddard ME, Wang J, Spangenberg GC, Forster JW, Rognli OA (2013) Prospects for genomic selection in forage plant species. Plant Breed 132(2):133–143. doi:10.1111/pbr.12037

He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129(3):641–651. doi:10.1007/s00122-015-2655-1

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49(1):1. doi:10.2135/cropsci2008.08.0512

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci 50(5):1681. doi:10.2135/cropsci2009.11.0662

Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4(1):65–75

Isidro J, Jannink JL, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128(1):145–158. doi:10.1007/s00122-014-2418-4

Jannink JL (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35. doi:10.1186/1297-9686-42-35

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42(12):1053–1059. doi:10.1038/ng.715

Li YH, Zhao SC, Ma JX, Li D, Yan L, Li J, Qi XT, Guo XS, Zhang L, He WM, Chang RZ, Liang QS, Guo Y, Ye C, Wang XB, Tao Y, Guan RX, Wang JY, Liu YL, Jin LG,

Zhang XQ, Liu ZX, Zhang LJ, Chen J, Wang KJ, Nielsen R, Li RQ, Chen PY, Li WB, Reif JC, Purugganan M, Wang J, Zhang MC, Wang J, Qiu LJ (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genom 14(1):579. doi:10.1186/1471-2164-14-579

Liu H, Zhou H, Wu Y, Li X, Zhao J, Zuo T, Zhang X, Zhang Y, Liu S, Shen Y, Lin H, Zhang Z, Huang K, Lubberstedt T, Pan G (2015) The impact of genetic relationship and linkage disequilibrium on genomic selection. PLoS One 10(7):e0132379. doi:10.1371/journal.pone.0132379

Longin CF, Mi X, Wurschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet. doi:10.1007/s00122-015-2505-1

Masuda T, Goldsmith PD (2009) World soybean production: area harvested, yield, and long-term projections. Int Food Agribus Manag Rev 12(4):143–162

Meuwissen T (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110(6):1303–1316. doi:10.1093/aob/mcs109

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 2(12):1595–1605

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5(3):103–113

Qiu L, Chang R, Liu Z, Guan R, Li Y (2006) Descriptors and data standard for soybean (*Glycine* spp.). China Agriculture Press, Beijing

Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PLoS One 8(6):e66428. doi:10.1371/journal.pone.0066428

Reif JC, Zhao YS, Wurschum T, Gowda M, Hahn V (2013) Genomic prediction of sunflower hybrid performance. Plant Breed 132(1):107–114. doi:10.1111/Pbr.12007

Riedelsheimer C, Technow F, Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genom 13(1):452. doi:10.1186/1471-2164-13-452

Rutkoski JE, Heffner EL, Sorrells ME (2011) Genomic selection for durable stem rust resistance in wheat. Euphytica 179(1):161–173. doi:10.1007/s10681-010-0301-1

Sallam AH, Endelman JB, Jannink JL, Smith KP (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. Plant Genome 8(1). doi:10.3835/plantgenome2014.05.0020

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278):178–183. doi:10.1038/nature08670

Shu YJ, Yu DS, Wang D, Bai X, Zhu YM, Guo CH (2013) Genomic selection of seed weight based on low-density SCAR markers in soybean. GMR 12(3):2178–2188. doi:10.4238/2013.July.3.2

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH (2008) Genomic selection using different marker types and densities. J Anim Sci 86(10):2447–2454. doi:10.2527/jas.2007-0010

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8(1):e54985. doi:10.1371/journal.pone.0054985

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink J-L, McCouch SR, Mauricio R (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11(2):e1004982–e1004982

St Martin S (1982) Effective population size for the soybean improvement program in maturity groups 00 to IV. Crop Sci 22(1):151–152

Team RC (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genom 15(1):556

Wurschum T, Reif JC, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14:85. doi:10.1186/1471-2156-14-85

Zhao Y, Gowda M, Liu W, Wurschum T, Maurer HP, Longin FH, Ranc N, Reif JC (2012a) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124(4):769–776. doi:10.1007/s00122-011-1745-y

Zhao Y, Gowda M, Longin FH, Wurschum T, Ranc N, Reif JC (2012b) Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theor Appl Genet 125(4):707–713. doi:10.1007/s00122-012-1862-2

Zhao Y, Li Z, Liu G, Jiang Y, Maurer HP, Wurschum T, Mock HP, Matros A, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Gowda M, Longin CF, Reif JC (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci USA 112(51):15624–15629. doi:10.1073/pnas.1514547112

Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics 182(1):355–364. doi:10.1534/genetics.108.098277