Theor Appl Genet (2014) 127:749–762
DOI 10.1007/s00122-013-2255-x

ORIGINAL PAPER

The impact of population structure on genomic prediction in stratified populations

Zhigang Guo · Dominic M. Tucker · Christopher J. Basten · Harish Gandhi · Elhan Ersoz · Baohong Guo · Zhanyou Xu · Daolong Wang · Gilles Gay

Received: 8 July 2013 / Accepted: 14 December 2013 / Published online: 24 January 2014

Abstract

Key message **Impacts of population structure on the evaluation of genomic heritability and prediction were investigated and quantified using high-density markers in diverse panels in rice and maize.**

Abstract Population structure is an important factor affecting estimation of genomic heritability and assessment of genomic prediction in stratified populations. In this study, our first objective was to assess effects of population structure on estimations of genomic heritability using the diversity panels in rice and maize. Results indicate population structure explained 33 and 7.5 % of genomic heritability for rice and maize, respectively, depending on traits, with the remaining heritability explained by within-subpopulation variation. Estimates of within-subpopulation heritability were higher than that derived from quantitative trait loci identified in genome-wide association studies, suggesting 65 % improvement in genetic gains. The second objective was to evaluate effects of population structure on genomic prediction using cross-validation experiments. When population structure exists in both training and validation sets, correcting for population structure led to a significant decrease in accuracy with genomic prediction. In contrast, when prediction was limited to a specific subpopulation, population structure showed little effect on accuracy and within-subpopulation genetic variance dominated predictions. Finally, effects of genomic heritability on genomic prediction were investigated. Accuracies with genomic prediction increased with genomic heritability in both training and validation sets, with the former showing a slightly greater impact. In summary, our results suggest that the population structure contribution to genomic prediction varies based on prediction strategies, and is also affected by the genetic architectures of traits and populations. In practical breeding, these conclusions may be helpful to better understand and utilize the different genetic resources in genomic prediction.

Communicated by J. Crossa.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2255-x) contains supplementary material, which is available to authorized users.

Z. Guo (✉) · C. J. Basten · D. Wang · G. Gay
Syngenta Biotechnology, Inc., 3054 E Cornwallis Rd., Durham, NC 27709, USA
e-mail: zhigang.guo@syngenta.com

D. M. Tucker
Syngenta, Inc., 12101 Thorps Road, Clinton, IL 61727, USA

H. Gandhi
Syngenta India Ltd., Survey No. 660 and 661, Nuthankal Village, Medchal Mandal, R.R. District 501404, India

E. Ersoz · B. Guo · Z. Xu
Syngenta, Inc., 2369 330th Street, Slater, IA 50244, USA

Introduction

A central question in molecular breeding is to predict breeding values of elite breeding materials, which can be used to measure genetic merits of these materials for complex traits in plant and animal breeding. A widely used approach is marker-assisted selection in which quantitative trait loci (QTL) associated with traits of interest are first identified and then breeding values are predicted for the untested genotypes using a model including these QTL (Lande and Thompson 1990; Bernardo and Yu 2007; Jannink et al. 2010; Nakaya and Isobe 2012). A major limitation of this strategy is the low power of identifying

small-effect QTL, which may jointly explain a considerable proportion of the genetic variation in quantitative traits (Jannink et al. 2010). Often the over-estimation of QTL effects (Beavis 1994) causes another technical concern, namely the decreasing accuracy of predictions.

These challenges motivated developments of genomic prediction where genome-wide markers are simultaneously incorporated into a genomic model in an attempt to capture genetic variation from all the QTL associated with a quantitative trait (Meuwissen et al. 2001). Both simulations and empirical studies in recent years have demonstrated that, with genomic prediction, the prediction accuracy of breeding values can be increased and more genetic gain can be expected in plants (Piyasatian et al. 2007; Bernardo and Yu 2007; Lorenzana and Bernardo 2009; Zhong et al. 2009; de los Campos et al. 2009; Crossa et al. 2010; Heffner et al. 2011; Albrecht et al. 2011; Guo et al. 2012; Zhao et al. 2012; Riedelsheimer et al. 2012, 2013; de Oliveira et al. 2012; Windhausen et al. 2012; Guo et al. 2013; Technow et al. 2013; Crossa et al. 2013), and animals (Legarra et al. 2008; Lee et al.2008; Luan et al. 2009; Hayes et al. 2009; Moser et al. 2009; Rolf et al. 2010; Wolc et al. 2011; Mujibi et al. 2011; Daetwyler et al. 2012; Karoui et al. 2012; Kärkkäinen and Sillanpää 2012; Edriss et al. 2013; Habier et al. 2013).

Population structure, as a property of a pedigree, is an important factor affecting predictions of breeding values with genomic models. Population structure may exist in random or pedigreed populations, owing to geography, natural selection or artificial selection (Yu et al. 2006; Price et al. 2010). Due to different allele frequencies among subpopulations, population structure can produce spurious marker–trait associations in genome-wide association studies (Lander and Schork 1994; Pritchard and Donnelly 2001; Marchini et al. 2004; Price et al. 2010). Consequently, these false associations may inflate estimates of genomic heritability (Visscher et al. 2012) and bias accuracies of genomic predictions (Makowsky et al. 2011; Riedelsheimer et al. 2012; Wray et al. 2013).

Several approaches have been developed to control for population structure in genomic prediction. One approach is to exploit the mean performances of subpopulations to account for population structure. With this approach, population structure may be defined based on known breeding origins of lines (Legarra et al. 2008; Albrecht et al. 2011), the clusters derived from prior pedigrees (Saatchi et al. 2011) or molecular markers (Windhausen et al. 2012). Another approach is to incorporate top principal components from principal component analysis (PCA, Price et al. 2006) as fixed effects into genomic models for correcting for population structure (Yang et al. 2010). However, adding these fixed variables into genomic models raises a concern of double-counting for population structure as these components are derived from a genomic relationship matrix which is already implicitly modeled in genomic models (Janss et al. 2012). To address this issue, a solution was developed by utilizing a re-parameterization of the genomic best linear unbiased prediction (GBLUP) model (Meuwissen et al. 2001), allowing a natural partition of genetic variation between across and within subpopulation (Janss et al. 2012). Given the successful partition of genetic variances in human and wheat studies (Janss et al. 2012), further studies are needed to investigate and quantify impacts of population structure on genomic prediction in breeding populations, aiming to provide useful insight to better understand and utilize different genetic resources in genomic prediction across crops.

Therefore, objectives in this study are threefold: (1) assess the influence of population structure on estimates of genomic heritability for economically important traits based on phenotypic and genotypic data from diversity panels in rice and maize using the reparameterized GBLUP model; (2) empirically evaluate the impact of population structure on the accuracy of genomic prediction using cross-validation experiments for the traits investigated on the above genomic model and (3) investigate effects of genomic heritability in training and validation samples on accuracies with genomic prediction. Our conclusions may aid recommendations for the utility of within- and across-subpopulation genetic variances to increase genetic gains with different genomic selection strategies in stratified and pedigreed breeding populations.

Materials and methods

Two public data sets were used for genomic analysis in this study: (1) rice diversity panel (Zhao et al. 2011), and (2) maize diversity panel (Cook et al. 2012).

Rice diversity panel

We used genotype and phenotype data of a rice diversity panel consisting of 413 inbred lines from 82 countries (Zhao et al. 2011). The panel contained accessions from six subpopulations: *indica* (IND, 87), *temperate japonica* (TEJ, 96), *tropical japonica* (TRJ, 97), *aus* (AUS, 57), *aromatic* (ARO, 14), and *admixed* (ADM, 62). A 44-K chip [44,100 single nucleotide polymorphisms (SNP)] was used for genotyping each inbred in this panel. After filtering SNPs with low call rates (<70 %) and allele frequencies (<0.01), a total of 36,901 high-performing SNP markers were retained for genetic analysis. These SNPs cover approximately 380 Mb of the genome at a density of about 1 SNP per 10 kb across the 12 chromosomes of rice. Each inbred from the diversity panel was evaluated

Table 1 The posterior means and standard deviations of genomic heritabilities for each trait in the rice and maize panels

Panel	Trait	h^{2a}_{gA}	h^{2b}_{gW}	h^2_g	h^2_{QTL}
Rice	Flowering time				
	Flowering time at Arkansas	0.20 ± 0.02 (0.27)	0.54 ± 0.04 (0.73)	0.73 ± 0.04	0.26
	Flowering time at Faridpur	0.16 ± 0.03 (0.39)	0.26 ± 0.05 (0.61)	0.42 ± 0.05	0.05
	Flowering time at Aberdeen	0.03 ± 0.01 (0.04)	0.63 ± 0.06 (0.96)	0.66 ± 0.06	0.50
	FT ratio of Arkansas/Arberdeen	0.10 ± 0.02 (0.22)	0.39 ± 0.06 (0.83)	0.49 ± 0.06	0.39
	FT ratio of Faridpur/Arberdeen	0.07 ± 0.03 (0.17)	0.41 ± 0.07 (0.83)	0.48 ± 0.07	0.26
	Morphology				
	Culm habit	0.39 ± 0.02 (0.49)	0.39 ± 0.04 (0.51)	0.78 ± 0.03	
	Flag leaf length	0.08 ± 0.02 (0.13)	0.56 ± 0.06 (0.87)	0.64 ± 0.06	0.15
	Flag leaf width	0.36 ± 0.02 (0.48)	0.38 ± 0.03 (0.50)	0.74 ± 0.03	0.24
	Yield component				
	Panicle number per plant	0.57 ± 0.02 (0.72)	0.22 ± 0.02 (0.28)	0.80 ± 0.02	0.08
	Plant height	0.32 ± 0.02 (0.39)	0.50 ± 0.03 (0.61)	0.81 ± 0.03	0.22
	Panicle length	0.33 ± 0.02 (0.45)	0.41 ± 0.04 (0.55)	0.73 ± 0.04	0.12
	Primary panicle branch number	0.26 ± 0.03 (0.42)	0.36 ± 0.05 (0.58)	0.62 ± 0.05	0.06
	Seed number per panicle	0.10 ± 0.03 (0.17)	0.50 ± 0.06 (0.83)	0.59 ± 0.06	
	Florets per panicle	0.15 ± 0.03 (0.22)	0.54 ± 0.05 (0.77)	0.69 ± 0.05	0.15
	Panicle fertility	0.14 ± 0.02 (0.29)	0.51 ± 0.05 (0.86)	0.65 ± 0.05	0.07
	Seed morphology				
	Seed length	0.26 ± 0.01 (0.28)	0.65 ± 0.02 (0.72)	0.90 ± 0.02	0.38
	Seed width	0.39 ± 0.02 (0.43)	0.50 ± 0.02 (0.57)	0.89 ± 0.02	0.32
	Seed volume	0.30 ± 0.01 (0.33)	0.60 ± 0.02 (0.67)	0.90 ± 0.02	0.12
	Seed surface area	0.22 ± 0.01 (0.26)	0.68 ± 0.02 (0.74)	0.90 ± 0.02	0.08
	Brown rice seed length	0.31 ± 0.01 (0.34)	0.60 ± 0.02 (0.66)	0.91 ± 0.02	0.24
	Brown rice seed width	0.41 ± 0.02 (0.45)	0.49 ± 0.02 (0.55)	0.90 ± 0.02	0.21
	Brown rice surface area	0.24 ± 0.01 (0.24)	0.66 ± 0.02 (0.71)	0.90 ± 0.02	0.11
	Brown rice volume	0.30 ± 0.02 (0.34)	0.59 ± 0.02 (0.66)	0.89 ± 0.02	0.22
	Seed length/width ratio	0.34 ± 0.02 (0.38)	0.56 ± 0.02 (0.62)	0.90 ± 0.02	0.44
	Brown rice length/width ratio	0.36 ± 0.01 (0.39)	0.56 ± 0.02 (0.61)	0.92 ± 0.02	0.39
	Stress tolerance				
	Straighthead susceptibility	0.30 ± 0.03 (0.41)	0.43 ± 0.04 (0.59)	0.73 ± 0.04	
	Blast resistance	0.29 ± 0.03 (0.40)	0.44 ± 0.04 (0.60)	0.73 ± 0.04	0.26
	Quality				
	Amylose content	0.41 ± 0.02 (0.49)	0.43 ± 0.03 (0.51)	0.85 ± 0.03	0.35
	Alkali spreading value	0.18 ± 0.02 (0.30)	0.41 ± 0.06 (0.70)	0.59 ± 0.06	0.30
	Protein content	0.09 ± 0.02 (0.18)	0.41 ± 0.06 (0.82)	0.50 ± 0.06	0.18
Maize	Kernel composition				
	Starch content	0.03 ± 0.01 (0.05)	0.51 ± 0.09 (0.95)	0.54 ± 0.09	
	Protein content	0.02 ± 0.01 (0.03)	0.49 ± 0.08 (0.97)	0.51 ± 0.08	
	Oil content	0.08 ± 0.02 (0.13)	0.54 ± 0.07 (0.87)	0.62 ± 0.08	

h^2_{gA}, across-subpopulation genomic heritability; h^2_{gW}, within-subpopulation genomic heritability; h^2_g, genomic heritability; h^2_{QTL}, proportion of phenotypic variance explained by QTL obtained from Zhao et al. (2011) in the rice panel

[a] In parentheses is h^2_{gA}/h^2_g

[b] In parentheses is h^2_{gW}/h^2_g

for important agronomic traits in Arkansas over 2 years from 2006 to 2007 with two replicates per year. In the current study, we used the phenotypic data from 30 traits impacting flowering time, plant morphology, yield component, seed morphology, stress tolerance, and quality (Table 1), and phenotypic means of each inbred line across years and replicates were used for our data analysis for each trait.

Maize diversity panel

The maize diversity panel is another valuable pubic breeding resource, which is composed of 282 maize inbred lines capturing a large proportion of the genetic diversity in maize public breeding programs around the world (Flint-Garcia et al. 2005). In the current study, we used genotype and phenotype data of 257 inbred lines from four

subpopulations: *non-stiff stalk* (NSS, 105), *stiff stalk* (SS, 28), *tropical/subtropical* (TS, 63), and *mixed* (MIXED, 61) after excluding genetically distinct sweet-corn and popcorn lines (Cook et al. 2012). A total of 51,741 SNPs (50 K chip) were used to genotype each inbred line, and 48,814 SNPs were retained for analysis after filtering the SNPs with low call rate (<70 %) and allelic frequency (<0.01), covering about 2,058 Mb of the genome at a density of about 1 SNP per 42 kb across the 10 chromosomes of maize. Each inbred from the diversity panel was planted and phenotyped for kernel composition traits starch, protein and oil content (Table 1) in seven locations: five locations (Clayton, NC; Columbia, MO; Aurora, NY; Homestead, FL; and Ponce, PR) in 2006 and two locations (Columbia, MO; and Aurora, NY) in 2007. Phenotypic best linear unbiased predictors from each line across years and locations obtained from mixed model analysis (Cook et al. 2012) were used for subsequent genetic analysis.

Models for estimation of genomic heritability and marker effects

In general, a GBLUP model may be used to estimate marker effects in genomic prediction based on genotypic and phenotypic data from a breeding population (Meuwissen et al. 2001). In the GBLUP model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{e}, \tag{1}$$

$\mathbf{y}$ is an $n \times 1$ vector of phenotypic data with n lines; $\mathbf{1}$ is an $n \times 1$ vector of ones; μ is the overall mean; $\mathbf{X}$ is an $n \times m$ marker genotype matrix with m the total number of markers; $\mathbf{b}$ is an $m \times 1$ vector of marker effects; and $\mathbf{e}$ is an $n \times 1$ vector of residuals following a normal distribution $N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$ with $\mathbf{0}$ an $n \times 1$ vector of zeros, $\mathbf{I}_n$ an $n \times n$ identity matrix, and σ_e^2 the residual variance. Marker effects $\mathbf{b}$ are assumed to be random effects following a normal distribution $N(\mathbf{0}, \mathbf{I}_m\sigma_b^2)$ with $\mathbf{0}$ an $m \times 1$ vector of zeros, $\mathbf{I}_m$ an $m \times m$ identity matrix, and σ_b^2 the common genetic variance for each of the m markers. The marker genotype matrix $\mathbf{X}$ is derived from observed marker information $\mathbf{M}$ according to VanRaden (2008) as

$$\mathbf{X} = \mathbf{M} - \mathbf{P}$$

where $\mathbf{M}$ is an $n \times m$ matrix from n lines and m markers with each element from each column defined as -1, 0, and 1 for the homozygous, heterozygous, and other homozygous genotypes; p_i is the frequency of the second allele at locus i, and $\mathbf{P}$ is an $n \times m$ matrix with $2(p_i - 0.5)$ for column i ($i = 1, 2, \ldots, m$) representing the mean of genotypes for the corresponding column of $\mathbf{M}$ from the unselected base population.

Given model (1), the genetic value $\mathbf{g}$ of each line may be written as

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{Xb}.$$

The genetic covariance matrix of $\mathbf{g}$ can be expressed as

$$\mathrm{Var}(\mathbf{g}) = \mathbf{XX}^{\mathrm{T}}\sigma_b^2 = \frac{\mathbf{XX}^{\mathrm{T}}}{\sum_{i=1}^{m} 2p_i(1-p_i)}\sigma_g^2 = \mathbf{G}\sigma_g^2$$

where $\mathbf{G}$ is an $n \times n$ symmetric, non-negative definite genomic relationship matrix, and σ_g^2 the total genetic variance of all the markers. Although the genomic relationship matrix $\mathbf{G}$ in GBLUP may implicitly capture genetic variation from population structure, family structure, admixture, genetic differences between full sibs within a family, and genetic diversity between unrelated individuals (Makowsky et al. 2011; Janss et al. 2012; Bastiaansen et al. 2012; Crossa et al. 2013; Habier et al. 2013), it is difficult to directly use this model to differentiate and quantify the genetic variances from each of these components, respectively, without other well-developed simulation studies (i.e., see Habier et al. 2013).

In this study, focusing on investigating impacts of population structure, we utilized a re-parameterization of the GBLUP which was first proposed by de los Campos et al. (2010) and then further developed by Janss et al. (2012) to partition the total genetic variation into across- and within-subpopulation variances. After eigenvalue decomposition of $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{UDU}^{\mathrm{T}}$$

where $\mathbf{U}$ is an $n \times (n-1)$ matrix of the eigenvectors of $\mathbf{G}$ with $\mathbf{U}_i$ the column i ($i = 1, 2, \ldots, n-1$) of $\mathbf{U}$ representing the principal component loads, and $\mathbf{D}$ is an $(n-1) \times (n-1)$ diagonal matrix with each diagonal element representing eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{n-1}$ ($\lambda_1 > \lambda_2 > \ldots > \lambda_{n-1}$) of $\mathbf{G}$, model (1) can be reparameterized as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{U}\alpha + \mathbf{e}. \tag{2}$$

In (2) $\boldsymbol{\alpha}$ is an $(n-1) \times 1$ vector of random effects, each having a normal distribution $N(\mathbf{0}, \mathbf{D}\sigma_g^2)$ with $\mathbf{0}$ an $(n-1) \times 1$ matrix of zeros. This model with the principal components as random variables is shown to generate the same distribution as that of model (1) with markers as variables. In comparison with model (1), the advantage of model (2) is to allow a natural separation of across-subpopulation genetic variance σ_{gA}^2 caused by population structure, and within-subpopulation genetic variance σ_{gW}^2 from σ_g^2 as

$$\sigma_{gA}^2 = \frac{1}{n-1}\sum_{i=1}^{d}\alpha_i^2$$

and

$$\sigma^2_{\mathrm{gW}} = \frac{1}{n-1}\sum_{i=d+1}^{n-1}\alpha_i^2$$

where d is the number of dominant principal components used to account for population structure that are independent of the trait of interest. The total genetic variance is

$$\sigma^2_g = \frac{1}{n-1}\sum_{i=1}^{n-1}\alpha_i^2.$$

Accordingly, the genomic heritabilities for across-subpopulation (h^2_{gA}), within-subpopulation (h^2_{gW}), and the whole population (h^2_{g}) can be written as

$$h^2_{\mathrm{gA}} = \frac{\sigma^2_{\mathrm{gA}}}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{e}}},$$

$$h^2_{\mathrm{gW}} = \frac{\sigma^2_{\mathrm{gW}}}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{e}}},$$

and

$$h^2_{\mathrm{g}} = \frac{\sigma^2_{\mathrm{g}}}{\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{e}}}.$$

Note that the above equations are derived in greater detail in Janss et al. (2012).

The above reparameterized GBLUP model was used to evaluate impacts of population structure on genomic heritability which was the first objective in this study. The unknown parameters μ, $\boldsymbol{\alpha}$, σ^2_{gA}, σ^2_{gW}, σ^2_{g}, σ^2_{e}, h^2_{gA}, h^2_{gW}, h^2_{g} were estimated by Markov Chain Monte Carlo (MCMC) using a Gibbs sampler as proposed by de los Campos et al. (2010) and Janss et al. (2012) for each trait using phenotypic and genotypic data from all individuals in the rice or maize population. A total of 25,000 MCMC iterations were run with the first 5,000 iterations discarded for burn-in. The remaining 20,000 iterations were used to estimate means and standard deviations of these parameters.

The second objective in the current study was to evaluate effects of population structure on genomic prediction using cross-validations described later. This required estimations of effects for each marker with GBLUP model (1) based on a training data set, which was a subset of the rice or maize population. To achieve this goal, effects of principal components $\boldsymbol{\alpha}$ were first estimated with model (2) based on phenotypic and genotypic information in the training sample, and marker effects $\mathbf{b}$ in model (1) were then estimated using

$$\mathbf{b} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{U}\mathbf{D}^{-1}\boldsymbol{\alpha}}{\sum_{i=1}^{m} 2p_i(1-p_i)}. \quad (3)$$

In order to account for population structure for estimation of marker effects, the second model GBLUP with correction for population structure (GBLUP-CPS) was also used and directly compared with GBLUP. With GBLUP-CPS, given the $\boldsymbol{\alpha}$ vector estimated from model (2), the first d elements, effects of the top d principal components, were set to zero in an attempt to eliminate effects of population structure on estimating marker effects using Eq. (3). In the current study, differences in prediction accuracy between GBLUP and GBLUP-CPS were used to quantify the influences of population structure on genomic prediction.

Cross-validation and prediction accuracy

Cross-validation methods were used to estimate prediction accuracies with GBLUP and GBLUP-CPS models. We tested two cross-validation methods CV1 and CV2, mimicking different prediction strategies in practical breeding.

CV1 is a stratified fivefold cross-validation design conditional on known population structure (Albrecht et al. 2011). With CV1, all the individuals within each subpopulation were partitioned into mutually exclusive datasets W_1, W_2, W_3, W_4, and W_5 with the similar sample sizes (Supplementary Figure S1A). Though this subdivision was specific to the subpopulation, the same categorical labeling for partition was used across all subpopulations. Following individuals that fell into the same category were combined across all the subpopulations to build five subsets: S_1, S_2, S_3, S_4, and S_5. For example, S_1 consisted of the individuals labeled W_1 across all subpopulations. Four of the five subsets were used to build a training data set for estimating marker effects using models described in previous sections. One subset was used in turn as a validation data set to estimate prediction accuracy which was measured as the correlation between predicted and observed phenotypes in the sample. Additional details for estimating prediction accuracy are discussed later. The CV1 strategy was intended to assess the prediction strategy in which training and validation samples contained similar patterns of population stratification. This could occur if training and validation samples were sampled from a stratified population (Wray et al. 2013).

CV2 focuses on predicting performances of individuals within a subpopulation. Three prediction schemes were used in terms of assembly of training samples. The first scheme, within-subpopulation prediction (WP), was a typical fivefold cross-validation design based on the subdivision of W_1, W_2, W_3, W_4, and W_5 within a specific subpopulation obtained from CV1. Each of W_1–W_5 data sets

was tested as a validation sample in turn with the remaining data sets within the subpopulation combined to build the training sample (Supplementary Figure S1B). Prediction accuracy was then calculated as the correlation between predicted and observed phenotype in the validation sample. In comparison to CV1, both training and validation sets in WP were from the same subpopulation, with no across-subpopulation information included.

The second scheme in CV2 was across-subpopulation prediction (AP), where all individuals from other subpopulations were used to build training samples to predict performances of individuals from the same validation set with WP. For example, if W_5 in the maize tropical-subtropical (TS) subpopulation was used as a validation sample, then the training sample would consist of all the non-TS individuals in the panel (Supplementary Figure S1B). For comparison, we also performed AP for the whole test subpopulation (APW), rather than a specific fold within it. In the previous example, this could mean using all non-TS as a training sample to predict all TS individuals (Supplementary Figure S1C).

The third scheme in CV2 was combined prediction (CP), where the validation set was onefold of one subpopulation: e.g., W_5 in the maize TS subpopulation, and training set were all the remaining individuals in the panel, a combination of training samples from WP and AP (Supplementary Figure S1B). The population structure was retained in the training samples, and its impact was evaluated on genomic prediction, similar to CV1. In summary, except for APW, WP, AP and CP all predicted onefold of one subpopulation (Supplementary Figure S1B), allowing us to determine the optimal scheme for predictions. We excluded validations for subpopulations with sample sizes lower than 60 in the rice and maize populations to avoid large sampling errors.

A key difference between CV1 and CV2 is that population structure is present in validation samples in CV1, but not in CV2. In practical breeding, a predicted population could be either type of validation sample, depending on specific breeding goals and stages. However, given differences between both approaches, we still wanted to explore the relationship between CV1 and CV2. Particularly, we were interested in comparing the performance of GBLUP-CPS in CV1 with the mean performance of WP across all subpopulations in CV2, both of which utilized the within-subpopulation genetic variance, a major genetic resource driving genomic prediction in practical breeding.

A total of 100 replicates were performed for the fivefold cross-validations in the CV1 and CV2 strategies, resulting in 500 estimates of prediction accuracies. At each fold in each replicate, the prediction accuracy was measured as the correlation coefficient between observed phenotypes of individuals in the validation set and their breeding values predicted by

$$\hat{y}_i = \hat{\mu} + \sum_{j=1}^{m} \mathbf{X}_{ij}\hat{\mathbf{b}}_j,$$

where $\hat{y}_i$ is the estimated breeding value of individual i in the validation sample; $\hat{\mu}$ and $\hat{\mathbf{b}}_j$ are the overall mean and marker effects estimated from a training sample using GBLUP and GBLUP-CPS; and $\mathbf{X}_{ij}$ is the genotype of marker j for line i in the validation set. The final reported prediction accuracy was in fact the mean of the 500 predictions generated across replicate runs. Overall accuracies between various models and approaches tested in the study were compared using a pairwise t test ($\alpha = 0.05$) in CV1 and CV2, respectively. Gains or losses in prediction accuracy with one model (e.g., model A) over another (e.g., model B) were calculated using $(R_A - R_B)/R_B$, where R_A represents prediction accuracy with model A, and R_B prediction accuracy with model B.

Detection of population structure using top principal components

One important requirement of the above genomic models was determining the number of top principal components used to capture population structure in the rice and maize diversity panels. In this study, the decision was made mainly based on PCA using genome-wide markers and prior genetic information. Furthermore, in order to interpret the variation of each principal component, one-way ANOVA analysis was performed with principal components as dependent variables and known subpopulation as a categorical explanatory variable (Patterson et al. 2006). Results from this analysis were also used to support the determination of top principal components.

Results

Figure 1 showed the inference of population structure using the top principal components obtained from PCA. First, based on the relative contribution to global molecular variance, the top four principal components from PCA were chosen to represent the population structure in the rice population (Fig. 1a). These components with individual contributions ranging from 2 to 41 % jointly explained 58.7 % of global molecular variance, close to the estimate of F_{st} of 0.56 using genome-wide SNPs. Inspection of these top principal components revealed further information about the population structure. In Fig. 1b, the first principal component (on the abscissa) played a key role in separating *japonica* and *indica* (IND). ARO was clustered between them, but closer to *japonica*, while AUS showed significant overlap with IND. These observations were in agreement

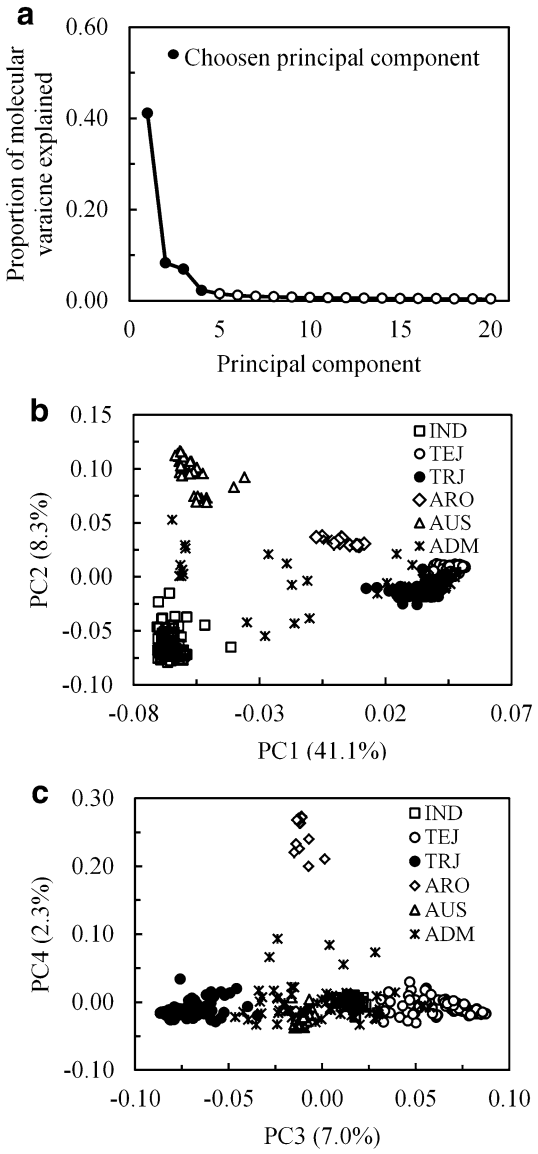

Fig. 1 Population structure derived from the top principal components from PCA in the rice diversity panel. **a** Proportion of molecular variance explained by each of top 20 principal components; **b** the first principal component (PC1) versus the second principal component (PC2); **c** the third principal component (PC3) versus the fourth principal component (PC4). IND, *indica*; TEJ, *temperate japonica*; TRJ, *tropical japonica*; AUS, *aus*; ARO, *aromatic*; ADM, *admixed*

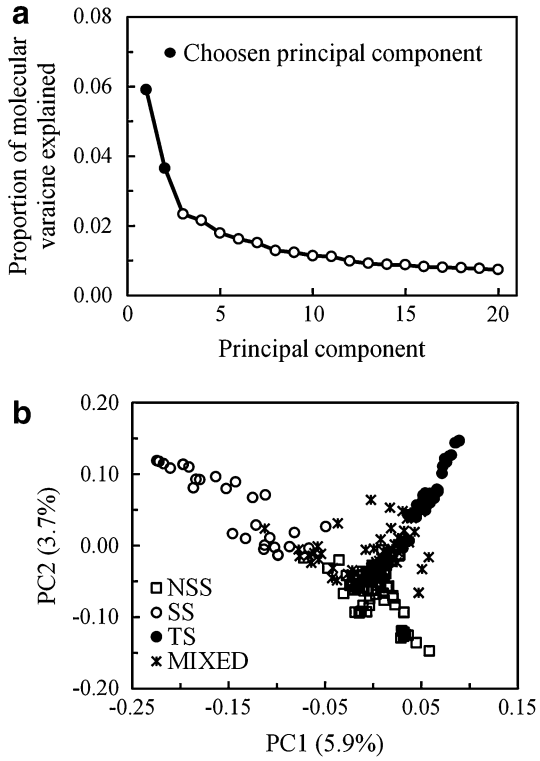

Fig. 2 Population structure derived from the top principal components from PCA in the maize diversity panel. **a** Proportion of molecular variance explained by each of top 20 principal components; **b** the first principal component (PC1) versus the second principal component (PC2). NSS, *non-stiff stalk*; SS, *stiff stalk*; TS, *tropical/subtropical*; MIXED, *mixed*

with the results from Supplementary Figure S2A obtained from one-way ANOVA analysis, where a large difference of the group means was found between TEJ and IND and relatively small difference was seen between IND and AUS or between TRJ and TEJ. The second principal component (on the ordinate) tended to capture the variation between IND and AUS, showing a high level of differentiation in the subpopulation means in Supplementary Figure S2B. With both axes, Fig. 1b provided a clear separation among *japonica* (TEJ and TRJ), IND and AUS, with ADM dispersed between these groups. TEJ and TRJ are genetically close and were grouped together on the right side of the plot. ARO was differentiated from others, but showed a close relationship to *japonica*. This separation was more clearly demonstrated in Fig. 1c, mainly due to the variation from the fourth principal component (on the ordinate) capturing the major difference between ARO and others, which was similarly reflected in Supplementary Figure S2D. In contrast, the abscissa in this figure was mainly used to separate TEJ and TRJ in *japonica*. These results were in agreement with the previous genetic studies about the population structure in rice (Garris et al. 2005; Zhao et al. 2011).

The differentiation of population structure using top principal components in the maize germplasm is shown in Fig. 2. The top two principal components were chosen to capture the population structure, explaining 5.9 and 3.7 %, respectively, of global molecular variance. In Fig. 2b, the first principal component (on the abscissa) tended to group TS and NSS to the right and SS to the left, while the second principal component (on the ordinate) tended to group SS and TS towards the top and NSS near the bottom. The differentiation of these subpopulations was further supported

by results from one-way ANOVA on the basis of each axis (Supplementary Figure S3). Based on the plot with both axes, three genetically different groups NSS, SS and TS were separated, but the differentiation between these groups showed a cline. This differed from discrete clusters shown in the rice panel, likely caused by an increased level of gene flow between these groups in maize (Garris et al. 2005). Similar to the rice result, the MIXED group was dispersed among three major groups SS, NSS and TS, but with a closer relationship to TS and NSS, which was in agreement with Flint-Garcia et al. (2005). It should be noted that the public diversity panel was originally collected to investigate the genetic architecture of traits, and then the genetic differentiation could be less than that in commercial elite lines due to a long-term selection in the latter.

In both cases, although the overall proportions of molecular variation explained by the top principal components were close to or even greater than those reported in the previous studies (Garris et al. 2005; Flint-Garcia et al. 2005; Zhao et al. 2011), more investigations were performed to examine if adding other principal components improved the differentiation of subpopulations. This analysis was particularly important for the maize population where F_{st} was actually estimated at 0.14, higher than the joint contribution of 0.10 from top two principal components, suggesting the need of adding other components. Compared to the significance of the ANOVA p value for each principal component in Supplementary Figures S2 and S3, the p value became insignificant from the fifth principal component in the rice population and the third in the maize population (data not shown). These results served as further evidence to support our decision. Therefore, the number of principal components d used in genomic model (2) was determined to be 4 and 2 in subsequent analysis for rice and maize, respectively.

The next objective was to assess impacts of population structure represented by these top principal components on inferences of genomic heritability based on the entire diversity panel in rice and maize. Table 1 showed estimates of genomic heritabilities for different traits based on the partition of genetic variance in both populations (Supplementary Table S1). In the rice panel, estimates of h^2_g, h^2_{gA}, and h^2_{gW} showed a high level of variability across traits, mainly attributed to the different genetic architectures. Across all traits, estimates averaged 0.74, 0.26, and 0.48 for h^2_g, h^2_{gA}, and h^2_{gW}, respectively. The mean proportion of h^2_{gA} over h^2_g reached 33 % with the remaining 67 % accounted for by h^2_{gW}. Across three kernel composition traits, the estimates of h^2_g, h^2_{gA}, and h^2_{gW} averaged 0.56, 0.04, and 0.52 in maize. The mean proportion of h^2_{gA} over h^2_g was 7.5 %, relatively less than that in rice. Moreover, the estimate of h^2_{gA} was comparable to R^2 obtained from one-way ANOVA with subpopulations as a categorical explanatory variable for phenotypes of each trait (Supplementary Figure S4), indicating a good representation of the population structure using the top principal components chosen. Overall, these results suggested that, although population structure showed a significant impact on genomic heritability, within-subpopulation variation remains a major resource of genetic variance.

Following estimates of genomic heritabilities were compared to that due to QTL identified from genome-wide association studies (Table 1). h^2_{QTL}, the proportion of phenotypic variation explained by all QTL identified for each trait using genome-wide association study in the rice population (Zhao et al. 2011), was lower than or comparable to the estimates of h^2_g and h^2_{gW}. Across all traits, the average h^2_{QTL} was estimated at 0.23, 69 and 54 % less than that of h^2_g and h^2_{gW}, respectively. Large difference was attributable to genetic variances of small-effect QTL which could not be detected by association analysis, but captured by genomic models. This point was more clearly shown in the maize population. Although there was no QTL identified for starch, protein and oil contents using genome-wide association study (Cook et al. 2012), genomic heritability was still estimated to be over 0.50 for h^2_g or h^2_{gW}, reflecting the polygenic architecture of these traits. Given these results, we further estimated genetic gains for the traits tested, according to $g = ih\sigma_g$, where g represented genetic gain, i represented selection intensity, h represented the square root of heritability, and σ_g the standard deviation of genetic variance (Falconer and Mackay 1996). Given the fixed parameter i and genetic variation σ_g, genetic gain was mainly determined by the estimate of heritability. In the rice population, the improvement in genetic gain with h_{gW} over h_{QTL} averaged 65 % across the traits tested. The similar advantage was seen in the maize case where the genetic gain with QTL was little as there were few QTL identified in the previous study (Cook et al. 2012).

Influences of population structure on accuracies of genomic prediction were assessed with two models GBLUP and GBLUP-CPS in CV1. Results from the rice and maize analysis are shown in Table 2. In the rice population, GBLUP-CPS provided significantly lower accuracy than GBLUP for all traits except for flowering time at Aberdeen. Decreases in accuracy ranged from 0 to 62 %, depending on the trait. Across all traits, accuracies with GBLUP-CPS averaged 0.47, 31 % lower than that with GBLUP. Similar trends were observed in maize, but the extent of reduction in accuracy was less than that in rice. The decrease in accuracy was attributed to the correction for population structure, suggesting a significant impact of population structure on genomic prediction. However, even with the reduction, accuracy with GBLUP-CPS reached 69 and 86 % of that with GBLUP for rice and maize, respectively, indicating the dominating role of within-subpopulation genetic variance in CV1.

Given the above results in CV1, the effect of population structure was further evaluated in CV2. Accuracies of

Table 2 The means and standard deviations of prediction accuracies with GBLUP and GBLUP-CPS models in CV1 for each trait in the rice and maize panels

Panel	Trait	GBLUP	GBLUP-CPS
Rice	Flowering time		
	Flowering time at Arkansas	0.66 ± 0.08	0.49* ± 0.09
	Flowering time at Faridpur	0.49 ± 0.10	0.30* ± 0.12
	Flowering time at Aberdeen	0.57 ± 0.09	0.57 ± 0.08
	FT ratio of Arkansas/Arberdeen	0.54 ± 0.09	0.46* ± 0.10
	FT ratio of Faridpur/Arberdeen	0.47 ± 0.10	0.41* ± 0.10
	Morphology		
	Culm habit	0.70 ± 0.06	0.33* ± 0.11
	Flag leaf length	0.50 ± 0.08	0.43* ± 0.08
	Flag leaf width	0.75 ± 0.05	0.48* ± 0.08
	Yield component		
	Panicle number per plant	0.82 ± 0.04	0.31* ± 0.12
	Plant height	0.75 ± 0.06	0.50* ± 0.08
	Panicle length	0.66 ± 0.06	0.34* ± 0.11
	Primary panicle branch number	0.63 ± 0.06	0.37* ± 0.11
	Seed number per panicle	0.57 ± 0.08	0.49* ± 0.08
	Florets per panicle	0.65 ± 0.07	0.53* ± 0.08
	Panicle fertility	0.54 ± 0.08	0.38* ± 0.11
	Seed morphology		
	Seed length	0.75 ± 0.05	0.57* ± 0.08
	Seed width	0.84 ± 0.04	0.55* ± 0.09
	Seed volume	0.81 ± 0.05	0.61* ± 0.08
	Seed surface area	0.78 ± 0.05	0.63* ± 0.08
	Brown rice seed length	0.79 ± 0.04	0.57* ± 0.08
	Brown rice seed width	0.84 ± 0.04	0.54* ± 0.09
	Brown rice surface area	0.77 ± 0.05	0.60* ± 0.08
	Brown rice volume	0.82 ± 0.04	0.62* ± 0.08
	Seed length/width ratio	0.80 ± 0.04	0.54* ± 0.09
	Brown rice length/width ratio	0.82 ± 0.04	0.56* ± 0.09
	Stress tolerance		
	Straighthead susceptibility	0.72 ± 0.05	0.48* ± 0.10
	Blast resistance	0.69 ± 0.06	0.43* ± 0.09
	Quality		
	Amylose content	0.80 ± 0.05	0.44* ± 0.13
	Alkali spreading value	0.51 ± 0.10	0.30* ± 0.14
	Protein content	0.44 ± 0.09	0.34* ± 0.10
Maize	Kernel composition		
	Starch content	0.34 ± 0.10	0.31* ± 0.11
	Protein content	0.29 ± 0.11	0.28* ± 0.11
	Oil content	0.42 ± 0.09	0.31* ± 0.10

* Indicates the accuracy with GBLUP-CPS differs significantly from that with GBLUP at $\alpha = 0.05$

predictions with CV2 for each subpopulation are shown in Table 3. Prediction accuracy in each cell in this table was the average of prediction accuracies over all traits for the rice and maize panels (Supplementary Table S2 to S5), respectively. WP provided prediction accuracy comparable to that with CP, both of which gave accuracy greater than that with AP. As expected, APW provided prediction accuracies close to that with AP, due to small differences in validation samples with same training samples. In both populations, the difference in accuracy between GBLUP and GBLUP-CPS was tiny in AP and CP, indicating the little effect of population structure in these methods in CV2.

Estimates of genomic heritability in training and validation data sets on accuracies of genomic predictions were examined in rice due to a large number of traits tested in comparison to maize data (Fig. 3). Note that the genomic heritability and prediction used for each trait in this figure were actually the means of estimates across 100 replicates of fivefold cross-validations. Because of the same cross-validation strategy applied for each trait, averages of genetic relationship between training and validation samples were same for all the traits tested in this study. Given this, across all traits, prediction with GBLUP in CV1 increased with h_g^2 in training and validation data sets with correlation coefficients of 0.94 and 0.93, respectively (Fig. 3a). Similar trends were seen for the relationship between h_{gW}^2 and accuracies with GBLUP-CPS, but to a lower level of correlation (0.85 and 0.70) likely due to the correction for population structure (Fig. 3b). The impact of population structure on genomic prediction was more clearly shown in Fig. 3c, where the reduction in accuracy with GBLUP-CPS over GBLUP increased with h_{gA}^2 in training and validation sets. Due to the negative correlation observed in Fig. 3c, these losses decreased with increasing h_{gW}^2 (Fig. 3d). Overall, based on the results of CV1, prediction accuracy increased with the estimate of genomic heritability in both training and validation sets, with the former showing a slightly greater impact. A similar conclusion was drawn from the results of WP in CV2 (Supplementary Figure S5), suggesting that this result would not rely on specific cross-validation designs.

Finally, the connection between CV1 and CV2 was investigated. We wanted to test if the overall prediction with GBLUP-CPS in CV1 was close to the average accuracy with WP across subpopulations in CV2, both of which were driven by within-subpopulation genetic variance. Recall that only relatively large subpopulations with sample sizes greater than 60 were tested originally in CV2 to ensure meaningful predictions. In order to test this hypothesis, even at the risk of increasing sampling error, this limitation was relaxed to 20, allowing conducting fivefold cross-validation with WP in each subpopulation (except for ARO). In our preliminary studies, due to missing phenotypic data, the actual sample size of ARO was even lower than 10 for some traits tested, generating very poor results with WP. The accuracy with GBLUP-CPS in CV1 was generally close to the mean accuracy with WP across

Table 3 Average prediction accuracies of GBLUP and GBLUP-CPS models with WP, AP, APW, and CP for subpopulations in CV2 across all traits based on the rice and maize diversity panels

Population	Subpopulation	WP	AP		APW		CP	
		GBLUP	GBLUP	GBLUP-CPS	GBLUP	GBLUP-CPS	GBLUP	GBLUP-CPS
Rice	IND	0.54	0.21	0.24	0.21	0.24	0.55	0.55
	TRJ	0.49	0.28	0.26	0.28	0.27	0.50	0.50
	TEJ	0.54	0.33	0.30	0.33	0.29	0.55	0.52
Maize	NSS	0.26	0.03	0.04	0.03	0.04	0.22	0.23
	TS	0.29	−0.07	−0.07	−0.07	−0.07	0.26	0.26

WP, within-subpopulation prediction; AP, across-subpopulation prediction; APW, across-subpopulation prediction for the whole test subpopulation; CP, combined prediction; IND, *indica*; TRJ, *tropical japonica*; TEJ, *temperate japonica*; NSS, *non-stiff stalk*; TS, *tropical/subtropical*

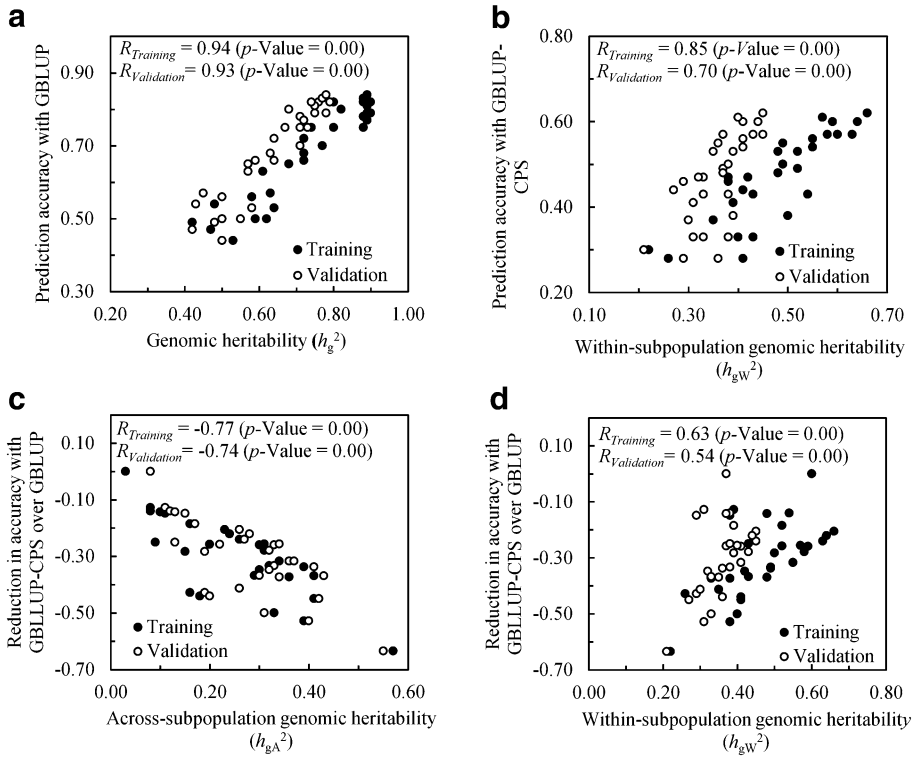

Fig. 3 Relationship between estimates of genomic heritability in training and validation sets and accuracies with genomic predictions for each trait in CV1. **a** Relationship between estimates of genomic heritability h_g^2 in training and validation data sets and accuracies with GBLUP; **b** relationship between estimates of within-subpopulation genomic heritability h_{gW}^2 in training and validation data sets and accuracies with GBLUP-CPS; **c** relationship between estimates of across-subpopulation genomic heritability h_{gA}^2 in training and validation data sets and reduction in accuracy with GBLUP-CPS over GBLUP; **d** relationship between estimates of within-subpopulation genomic heritability h_{gW}^2 in training and validation data sets and reduction in accuracy with GBLUP-CPS over GBLUP. $R_{training}$: correlation coefficient between genomic heritability in training sets and accuracy with genomic prediction (**a**, **b**) or reduction in accuracy with GBLUP-CPS over GBLUP (**c**, **d**); $R_{validation}$: correlation coefficient between genomic heritability in validation sets and accuracy with genomic prediction (**a**, **b**) or reduction in accuracy with GBLUP-CPS over GBLUP (**c**, **d**)

all subpopulations in rice and maize populations (Fig. 4). Although increasing differences were observed for a few traits, likely due to the influence of unbalanced and limited individuals and genetic architecture across subpopulations, the average difference across all traits was equal to or even <0.01 in the rice and maize populations.

Discussion

Population structure causes significant impact on the estimation of genomic heritability, depending on populations and traits. It accounted for 58.7 % of global molecular variation in the rice population, much greater than 9.6 % in the

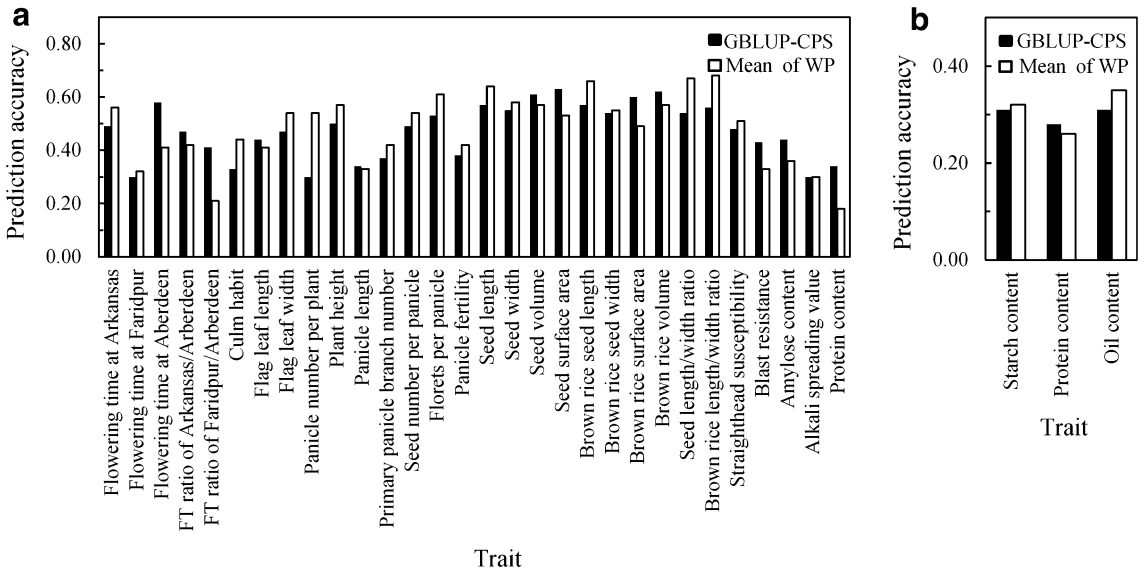

Fig. 4 Comparison of prediction accuracy with GBLUP-CPS in CV1 with the mean accuracy with WP across each subpopulation in CV2. **a** Rice; **b** maize. *WP* within-subpopulation prediction

maize population. The large difference could be caused by two major reasons (Garris et al. 2005). First, less cross-pollination in rice limits the gene flow between subpopulations, leading to a higher level of genetic differentiation than that in maize. The second explanation appeals to the history of domestication of these crops. In comparison to the one putative domestication event in maize history, we currently posit two domestication events in rice, one each for the separation of *indica* and *japonica* from the ancestor species. Given the genetic differentiation based on marker data only, we further investigated trait-dependent genetic variance partitions. Population structure had a significant impact on the estimation of genomic heritability, but the magnitude of this effect was largely different and specific to the trait tested, mainly determined by the genetic architecture of the trait. It should be noted that, in practice, the impact of population structure is expected to be closer to that in the maize population, in comparison to the rice population which should not be common in actual breeding populations.

Although population structure showed a significant impact on estimations of genomic heritability, even reaching a high level as shown in the rice population, the majority of genomic heritability was contributed by within-subpopulation genetic variance. More importantly, the estimate of within-subpopulation genomic heritability h^2_{gW} was found to be greater than the h^2_{QTL} derived from QTL detected by genome-wide association studies, improving genetic gain with 65 % on average. As expected, this advantage could be further increased when comparing h^2_{QTL} with h^2_g as the latter was greater than h^2_{gW} due to the inclusion of genetic variation caused by population structure. In practical breeding, selection is often based on multiple traits depending on goals of breeding programs. This requires not only the estimates of heritability for each trait, but also of genetic correlations between traits in order to build an appropriate selection index. Although more evaluations on building such an index for multiple-trait selection are needed, the benefit of genomic models, likely differing from the results reported in the current study, should be expected in comparison with the QTL-based approach.

The utilization of across- and within-subpopulation genetic variances depends on breeding strategies. The breeding process in plants can be summarized in three steps (Bernardo and Yu 2007; Jonas and de Koning 2013). First, parents are chosen from elite breeding germplasm, and different families are generated by crossing them (Step 1). Then, the performance of progeny in each family is field tested, and superior lines are promoted for further evaluation (Step 2). These advanced lines are re-evaluated across a large number of locations, and the best lines may be released for new varieties (Step 3). Note that lines from Step 3 may be used as parents in Step 1 for the next cycle of crossing and selection. More importantly, they can be used to build a pedigreed and diverse training population to predict the breeding values of the newly generated lines in Steps 2 and 3 for the next breeding cycle, saving expensive and time-consuming phenotyping efforts (Würschum et al. 2013). This motivated the development of the CV1 method in this study. In CV1, similar patterns of population structure exist in training and validation data sets, and population structure can serve as a positive contributor to benefit

predictions of breeding values (Makowsky et al. 2011; Bastiaansen et al. 2012; Crossa et al. 2013; Wray et al. 2013). As a result, we observed a consistently lower accuracy with GBLUP-CPS than GBLUP due to the correction for population structure. This conclusion was in agreement with the decline observed in a maize diversity panel composed of 285 dent lines (Riedelsheimer et al. 2012). Moreover, the reduced prediction with GBLUP-CPS was found to reflect the average level of accuracy with WP in CV2 across all subpopulations. This was not surprising as both methods were driven only by within-subpopulation genetic variance. It should be noted that although both across- and within-subpopulation genetic variances were utilized in CV1, the respective contribution of each for prediction significantly differed across populations and traits. Particularly, for the traits largely impacted by population structure, breeders need more caution because the population structure with the higher genetic merit could overcompensate newly generated breeding populations over time (Crossa et al. 2013).

In contrast, prediction in CV2 relies mainly on within-subpopulation genetic information. In practice, CV2 can be deployed in a way similar to CV1, but with no population structure in the predicted population. In this case, WP could be the main use for CV2, utilizing the within-subpopulation genetic variation. However, a relevant question concerns whether it is possible to improve accuracy by adding genetic information from other subpopulations. Our results indicated no consistent gains with CP over WP, suggesting little benefit from using other subpopulations. This conclusion was further supported by the decreased accuracy from AP and APW. The poor prediction may be due to the genetic heterogeneity caused by differences in linkage phases between QTL and markers (Riedelsheimer et al. 2013). Interactions between QTL and genetic backgrounds could be another reason. Similar results were also found in mice (Legarra et al. 2008) and other experiments in maize (Zhao et al. 2011; Windhausen et al. 2012; Guo et al. 2013; Technow et al. 2013; Crossa et al. 2013). It was reported that CP improved accuracy by introducing genetic variation from other subpopulations and compensating for limited sample sizes (Riedelsheimer et al. 2013). However, benefits were insignificant in the CP approach tested in this study, likely attributable to different populations and traits used. Although population structure exists in training samples in AP and CP, it showed little impact on prediction in CV2.

While it has been shown that genomic relationship plays a key role in determining prediction accuracies (Habier et al. 2007), our study suggested that prediction was also affected by genomic heritability in training and validation samples. First, we found accuracy increased with the genomic heritability in training samples, largely attributed to improved estimations of marker effects by introducing more genetic variation into those samples. A similar trend was seen in validation samples, where the higher heritability means less environmental noise, an indication of more accurate prediction in these samples. Given these results being consistent with previous studies (Bernardo and Yu 2007; Villumsen et al. 2008; Guo et al. 2012), our study also suggested that genomic heritability showed a slightly greater impact in training samples than in validation samples, highlighting the importance of increasing and maintaining sufficient genetic variation in training populations in practical genomic selection. Given the consistent conclusions obtained from different cross-validation schemes tested in this study, we expect that this finding may be used as a general guidance in developing genomic selection breeding strategies. In addition, as expected, reductions in prediction accuracy due to the correction for population structure were seen to increase with across-subpopulation heritability. This was mainly contributed by the across-subpopulation genetic variance caused by population structure. In practice, to avoid the decrease in prediction accuracy, one may consider making crosses using parental lines from genetically distant subpopulations for the traits highly affected by population structure.

Finally, it should be noted that the above conclusions are derived from additive genetic models using cross-validation. Although a high level of estimates of genomic heritability and predicted accuracy were achieved with these models, it is still worthwhile to explore the utilization of other genetic effects such as epistasis and genotype-by-environment interactions. More studies will be needed in the future not only for the extension and modification of the current reparameterized GBLUP model to accommodate these effects, but also for the preparation and evaluation of the quality of data set which is suitable for these analyses. Furthermore, the accuracy obtained from this study was based on cross-validations using the inbred lines from the rice and maize populations. More studies are needed to evaluate the predicted accuracy for offspring derived from these inbred lines at different generations (Jonas and de Koning 2013). This requires the construction of a genotyping platform for both parents and offspring at each cycle of the breeding process. Nonetheless, results from this study are encouraging and meaningful to deepen our understanding on the utilization of across- and within-subpopulation genetic variation for different traits and populations in different prediction and breeding strategies.

Acknowledgments The authors of the current manuscript would like to thank researchers and institutions who contributed to the development of the rice and maize diversity panels. In addition, the authors would like to express gratitude to the editor and three anonymous reviewers for their detailed input in assessment and improvement of the manuscript.

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

References

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

Bastiaansen J, Coster A, Calus M, Van Arendonk J, Bovenhuis H (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol 44:3

Beavis WD (1994) QTL analysis: power, precision and accuracy. In: Paterson AH (ed) Molecular dissection of complex traits. CRC Press, Boca Raton, pp 145–162

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Predictions of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2013) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. doi:10.1038/hdy.2013.16

Daetwyler HD, Swan AA, van der Werf JHJ, Hayes BJ (2012) Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet Sel Evol 44:33

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

de los Campos G, Gianola D, Rosa G, Weige K, Crossa J (2010) Semiparametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92:295–308

de Oliveira EJ, de Resende DV, da Silva Santos V, Ferreira CF, Oliveira GAF, da Silva MS, de Oliveira LA, Aguilar-Vildoso GI (2012) Genome-wide selection in cassava. Euphytica 187:263–276

Edriss V, Fernando RL, Su GS, Lund MS, Guldbrandtsen B (2013) The effect of using genealogy-based haplotypes for genomic prediction. Genet Sel Evol 45:5

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Prentice Hall, London

Flint-Garcia SA, Thuillet AC, Yu JM, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Guo Z, Tucker D, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124:261–275

Guo Z, Tucker D, Wang D, Basten C, Ersoz E, Briggs W, Lu J, Li M, Gay G (2013) Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. G3 3:263–272

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

Habier D, Fernando RL, Garrick DJ (2013) Genomic-BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607

Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Heffner EL, Jannink JL, Iwata H, Souza E, Sorrells ME (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597–2606

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Janss LG, de los Campos G, Sheehan N, Sorensen D (2012) Inferences from genomic models in stratified populations. Genetics 192:693–704

Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31(9):497–504

Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191:969–987

Karoui S, Carabaño MJ, Díaz C, Legarra A (2012) Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol 44:39

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4(10):e1000231

Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen TH (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119–1126

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7(4):e1002051

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517

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

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56

Mujibi FDN, Nkumah JD, Durunna ON, Stothard P, Mah J, Wang Z, Basarab J, Plastow G, Crews DH Jr, Moore SS (2011) Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. J Dairy Sci 89:3353–3361

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

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:2074–2093

Piyasatian N, Fernando R, Dekkers JCM (2007) Genomic selection for marker-assisted improvement in line crosses. Theor Appl Genet 115:665–674

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal component analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60:227–237

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected bi-parental maize populations. Genetics. doi:10.1534/genetics.113.150227

Rolf MM, Taylor JF, Schnabel RD, Mckay S, McClure M, Northcutt S, Kerley M, Weaber R (2010) Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet 11:24

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J et al (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol 43:1–16

Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3:197–203

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Villumsen TM, Janss L, Lund MS (2008) The importance of haplotype length and heritability using genomic selection in dairy cattle. J Anim Breed Genet 126:3–13

Visscher PM, Yang J, Goddard MEA (2012) A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res Hum Genet 13:517–524

Windhausen VS, Atlin CN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic predictions of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436

Wolc A, Stricker C, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Habier D, Fernardo R, Garrick D, Lamont SJ, Dekkers JCM (2011) Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genet Sel Evol 43:5

Wray NR, Yang J, Hayes BJ, Price AL, Michael E, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14(7):507–515

Würschum T, Reif JC, Kraft T, Janssen G, Zhao YS (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14:85

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhao KY, Tung CW, Eizenga GC, Wright MH, Ali L, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467

Zhao YS, Gowda M, Liu WX, Würschum T, Maurer HP, Longin FH, Ranc N, Reif JC (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776

Zhong SQ, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364