University of São Paulo
"Luiz de Queiroz" College of Agriculture

Influence of multi-trait modeling, dominance, and population structure in genomic prediction of maize hybrids

Danilo Hottis Lyra

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

Danilo Hottis Lyra
Bachelor in Biological Sciences

Influence of multi-trait modeling, dominance, and population structure in genomic prediction of maize hybrids

Advisor:
Prof. Dr. **ROBERTO FRITSCHE-NETO**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Lyra, Danilo Hottis

Influence of multi-trait modelling, dominance, and population structure in genomic prediction of maize hybrids / Danilo Hottis Lyra. - - Piracicaba, 2017.

103 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Milho tropical 2. Efeitos não-aditivos 3. Kernel Gaussiano 4. Variação no número de cópia I. Título

CONTENTS

RESUMO

Influência da modelagem *multi-trait,* dominância, e estruturação populacional na predição genômica em híbridos de milho

Predição genômica de híbridos simples é uma promissora ferramenta no melhoramento de milho, pois permite aumentar os ganhos genéticos por unidade de tempo, principalmente por reduzir o tempo de seleção. Uma estratégia que pode aumentar a acurácia das predições genômicas é realizar esta para múltiplos caracteres considerando os mesmos simultâneamente, ou utilizar índices de seleção, os quais captam a performance dos genótipos tanto em condições ótimas como em condições de estresse. Além disso, fatores como dominância, variantes estruturais, e estruturação populacional podem influenciar a acurácia de estimativas dos valores genéticos genômicos (VGG). Portanto, os objetivos foram aplicar predição genômica em híbridos de milho (*i*) incluindo modelos *multi-trait*, (*ii*) incorporando desvios de dominância e efeitos da variação no número de cópias, e (*iii*) controlando a estruturação populacional. Para isto, dois conjuntos de milho (HELIX e USP) foram utilizados, consistindo de 452 e 906 híbridos simples. Os caracteres avaliados foram produtividade de grãos, altura de planta e espiga, senescência, e quatro índices de seleção. A partir das análises *multi-trait* dos modelos GBLUP e GK, pôde-se concluir que a combinação dos índices é uma alternativa viável, aumentando a acurácia seletiva. Além disso, os resultados sugerem que o melhor método é a predição de híbridos incluindo desvios de dominância, principalmente para caracteres complexos. Observou-se também que incluir efeitos relacionados a variação no número de cópias indica ser adequado, devido ao aumento da acurácia e redução do viés nos modelos de predição genômica. Por outro lado, a acurácia de predição não aumentou quando se adicionou quatro diferentes conjuntos de estruturação como covariáveis fixas no modelo GBLUP. No entanto, usando o escalonamento multidimensional não métrico e o agrupamento do fineSTRUCTURE aumentaram a confiabilidade de estimação do VGG para produtividade de grãos e altura de plantas, respectivamente.

Palavras-chave: Milho tropical; Efeitos não-aditivos; Kernel Gaussiano; Variação no número de cópia

ABSTRACT

Influence of multi-trait modeling, dominance, and population structure in genomic prediction of maize hybrids

Genomic prediction of single-crosses is a promising tool in maize breeding, increasing genetics gains and reducing selection time. A strategy that can increase accuracy is applying multiple-trait genomic prediction using selection indices, which take into account the performance under optimal and stress conditions. Moreover, factors such as dominance, structural variants, and population structure can influence the accuracy of estimates of genomic breeding values (GEBV). Therefore, the objectives were to apply genomic prediction (*i*) including multi-trait models, (*ii*) incorporating dominance deviation and copy number variation effects, and (*iii*) controlling population structure in maize hybrids. Hence, we used two maize datasets (HELIX and USP), consisting of 452 and 906 maize single-crosses. The traits evaluated were grain yield, plant and ear height, stay green, and four selection indices. From multi-trait GBLUP and GK, using the combination of selection indices in MTGP is a viable alternative, increasing the selective accuracy. Furthermore, our results suggest that the best approach is predicting hybrids including dominance deviation, mainly for complex traits. We also observed including copy number variation effects seems to be suitable, due to the increase of prediction accuracies and reduction of model bias. On the other hand, adding four different sets of population structure as fixed covariates to GBLUP did not improve the prediction accuracy for grain yield and plant height. However, using nonmetric multidimensional scaling dimensions and fineSTRUCTURE group clustering increased reliability of the GEBV for GY and PH, respectively.

Keywords: Tropical maize; Non-additive effects; Gaussian kernel; Copy number variation

1. INTRODUCTION

Modern plant breeding comprises combinations of different approaches that include traditional methods, and the use of molecular markers as a "tool" for selecting plants with desirable traits. The last decade has seen tremendous advances in genome-scale data analysis, which was possible due to high-throughput DNA sequencing. In this way, single nucleotide polymorphism (SNPs), representing various regions of all chromosomes, are obtained to be applied in genomic studies (Guo et al. 2016).

Quantitative traits of agricultural importance in plants are influenced by many genomic regions. Thus, whole genome-enabled prediction, such as Genomic Prediction (GP) or Genomic Selection (GS), emerged as a statistical approach to overcome this biological complexity. The main objective of GP, proposed by Meuwissen et al. (2001), is to improve prediction of complex traits based on marker information, increasing precision of selection by generating a genomic estimated breeding value (GEBV) for selection candidates. The accuracies of GP models are most often evaluated by applying validations (independent validation, fold-validation, or jackknife), where all genotypes are randomly divided into training and validation sets (TS, VS). The TS is used to train the prediction model and estimate the marker effects, and by a correlation test using the predicted with the observed values in the validation set, it is possible to obtain the prediction accuracy (PA). The procedure is repeated several times to obtain robust estimates (Zhao et al. 2015).

The most commonly used methods in GP is the genomic best linear unbiased prediction (GBLUP), which utilizes a genomic relationship matrix (GRM) to estimate the genetic merit of an individual. The matrix defines the covariance between individuals based on observed similarity at the genomic level, rather than on expected similarity based on pedigree. Morota and Gianola (2014) reviewed whole-genome regression models using kernel methods to capture non-additive effects, either parametrically (GBLUP) or non-parametrically (Gaussian kernel, GK).

Genomic prediction is superior to phenotypic selection for increasing genetic gains per unit time and shortening the length of the breeding cycle (Heffner et al. 2010). According to Bernardo (2016), GP became a bandwagon in plant breeding in the late 2000s and has been implemented in major seed companies routinely, especially in maize and soybean. In addition, the author argued that similarly with phenotypic selection, applying GP routinely might work reasonably well on average. However, GP still faces challenges in predicting phenotypes of highly polygenic traits due to the complex biological processes, which several factors could influence the estimation of GEBV, such as non-additive effects, population structure, and structural variations (copy number variation). Therefore, the objectives were to apply genomic prediction in maize hybrids (*i*) including multi-trait

models, (*ii*) incorporating dominance deviation and CNV effects, and (*iii*) controlling population structure.

REFERENCES

Bernardo, R. (2016). Bandwagons I, too, have known. Theoretical and Applied Genetics, 129: 2323-2332.

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z. and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theoretical and Applied Genetics, 129(12): 2413-2427.

Heffner, E. L., Lorenz, A. J., Jannink, J. L. and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. Crop Science, 50: 1681-1690.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157: 1819-1829.

Morota, G. and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. Frontiers in genetics, 5: 363.

Zhao, Y., Mette, M. F. and Reif, J. C. (2015). Genomic selection in hybrid breeding. Plant breeding, 134(1): 1-10.

2. MULTI-TRAIT GENOMIC PREDICTION FOR NITROGEN RESPONSE INDICES IN TROPICAL MAIZE HYBRIDS

ABSTRACT

In maize breeding, genomic prediction may be an efficient tool for selecting single-crosses evaluated under abiotic stress conditions. In addition, a promising strategy is applying multiple-trait genomic prediction using selection indices (SIs), increasing genetics gains and reducing time per cycles. In this study, we aimed (*i*) to compare accuracy of single- and multi-trait genomic prediction (STGP; MTGP) in two maize datasets, (*ii*) to evaluate prediction of four selection indices that could contribute to the selection of tropical maize hybrids under contrasting nitrogen conditions, and (*iii*) to compare the use of linear (GBLUP) and nonlinear (RKHS/GK) kernels in STGP and MTGP analyses. For either single-trait GBLUP and RKHS analyses, the highest values obtained of accuracy was 0.40 and 0.41 using harmonic mean (HM), respectively. From multi-trait GBLUP and GK, using the combination of selection indices in MTGP seems to be suitable, increasing the accuracy. Adding grain yield and plant height in MTGP, showed a slight improvement in accuracy compared to STGP. In general, there was a modest benefit of using single-trait RKHS and GK multi-trait, rather than GBLUP.

Keywords: Abiotic stress; Single-trait genomic prediction; Gaussian kernel; GBLUP; Genomic heritability

Published in *Molecular breeding* on June 8, 2017 as doi 10.1007/s11032-017-0681-1

2.1. INTRODUCTION

Tropical maize is one of the most important crops for smallholder farmers in the Americas, and highly dependent of nitrogen (N) fertilization to increase crop yield (Trachsel et al. 2016). Thus, development of improved single-cross maize hybrids is critical for sustainable grain production (Gong et al. 2015). In breeding programs for abiotic stresses, N-use efficiency (NUE) represents genotypes that tend to have high yields in favorable conditions and low yields under stress; while N-stress tolerance represents genotypes in stressful and optimal conditions that tend to have satisfactory and low grain yield, respectively (Mueller and Vyn 2016). Therefore, selection can be performed directly under stress, indirectly under favorable conditions, or simultaneously under both optimal and stress conditions (Chen et al. 2016; Cecarelli et al. 1998).

Selection indices (SIs) have been used in an attempt to identify the best individuals and to improve the overall genotype performance based on several quantitative traits simultaneously

(Cerón-Rojas et al. 2016; Ceron-Rojas et al. 2015; Dekkers 2007). Specifically for N-use efficiency, a SI was proposed by Craswell and Godwin (1984) and referred as N-agronomic efficiency (NAE). On other hand, Miti et al. (2010) presented low-N tolerance index (LNTI). Furthermore, some SIs as low-N agronomy efficiency (LNAE) (Wu et al. 2011) and harmonic mean (HM) (Jafari et al. 2009) try to take into account the performance under both optimal and stress conditions. Therefore, superior maize hybrids could be selected by applying genomic prediction using SIs as traits, increasing the efficiency of selection and requiring less time than a conventional breeding program (Bernardo 2014; Schulthess et al. 2016).

Genomic prediction (GP), a promising selection method that uses marker and phenotypic information, has been routinely employed in maize hybrid breeding (Crossa et al. 2014; Massman et al. 2013; Riedelsheimer et al. 2012; Cantelmo et al. 2017) and, recently, become an important tool in breeding for abiotic stresses as dehydration (Beyene et al. 2015; Ziyomo and Bernardo 2013), nitrogen (Liu et al. 2016; Fritsche-Neto et al. 2012), and heat (Heslot et al. 2014). However, most studies have used single-trait (ST) analysis to predict genomic estimated breeding values (GEBV) in a stress environment (Crossa et al. 2010; Poland et al. 2012). As far as we know, SIs have not been used for multiple-trait genomic prediction (MTGP) in N stress, and assuming SIs are genetically correlated and presents different heritabilities (Wu et al. 2011), applying MTGP could increase the prediction accuracies (Calus and Veerkamp 2011; He et al. 2016; Wang et al. 2016).

Multiple-trait and multi-environment genomic prediction has been successfully implemented using genomic best linear unbiased prediction (GBLUP) model (Guo et al. 2014). However, depending on the germplasm and genetic architecture of the trait, Bayesian methods, non-linear semiparametric approaches as Reproducing Kernel Hilbert Space (RKHS) and Gaussian kernel (GK) could produce slightly better accuracies (Cuevas et al. 2016; Montesinos-Lopez et al. 2016; Jia and Jannink 2012; Hayashi and Iwata 2013). It is important to notice for RKHS and GK a few parameters must to be chosen correctly, in order to compute a genomic relationship matrix, such as the bandwidth parameter (h), which can be selected based on cross-validation procedure, restricted maximum likelihood, or empirical Bayesian method (Endelman 2011; Pérez-Elizalde et al. 2015; Morota and Gianola 2014). Furthermore, incorporating dominance effects in MTGP could be an efficient strategy to improve accuracy. Wang et al. (2016) and dos Santos et al. (2016) included dominance effects in the multi-trait GBLUP model in rice and maize hybrids, respectively, and found slight improvement in prediction and quality of variance components.

Therefore, our objectives were (i) to compare accuracy of single- and multi-trait genomic prediction for two maize datasets, (ii) to evaluate prediction of four selection indices that could contribute to the selection of tropical maize hybrids under contrasting nitrogen conditions, and (iii) to compare the use of linear (GBLUP) and nonlinear (RKHS/GK) kernels in STGP and MTGP analyses.

2.2. Materials and methods

2.2.1. Phenotypic dataset

Maize dataset I

We used 738 maize single-crosses from a diallel mating design between 49 tropical inbred lines, contrasting for N-use efficiency. The experimental scheme used was an augmented block design (unreplicated trial) consisted of 47 blocks, each with 16 unique hybrids and two checks. Trials were carried out in Anhembi (22°50'51"S, 48°01'06"W, 466 m) and Piracicaba, at São Paulo State, Brazil (22°42'23"S, 47°38'14"W, 535 m), during the second growing season of 2016 cultivated between January to June. In both sites the hybrids were evaluated under two nitrogen (N) levels, low (LN) with 30 kg N ha^{-1}, and normal (NN) with 100 kg N ha^{-1}.

We used plots of seven meters (m) spaced 0.50 m, with sowing density of about 57,000 kernels per hectare, under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, ton ha^{-1}) and plant height (PH, cm). Plots were manually harvested and GY was corrected to 13% moisture, the PH was measured from soil surface to the flag leaf collar on five representative plant within each plot.

For the joint analysis, we used a linear mixed model to calculate best linear unbiased predictions (BLUPs) for the hybrids in each N condition, by fitting the following model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + V\boldsymbol{b} + H\boldsymbol{g} + S\boldsymbol{i} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is a vector of phenotypic values of hybrids; $\boldsymbol{\beta}$ is a vector of fixed effects of site, checks, and site × check; $\boldsymbol{b}$ is block within site, where $\boldsymbol{b} \sim N(0, I\sigma_b^2)$; $\boldsymbol{g}$ is genotypic values of hybrids, where $\boldsymbol{g} \sim N(0, I\sigma_g^2)$; $\boldsymbol{i}$ is interaction sites × hybrids, where $\boldsymbol{i} \sim N(0, I\sigma_{lg}^2)$; $\boldsymbol{\varepsilon}$ is a vector of random residuals from checks, where $\boldsymbol{\varepsilon} \sim N(0, I\sigma_\varepsilon^2)$. σ_ε^2 was jointly estimated based on e sites with t replicated check in each site. X, V, H, and S are the incidence matrices for $\boldsymbol{\beta}$, $\boldsymbol{b}$, $\boldsymbol{g}$, and $\boldsymbol{i}$. Heterogeneous residual variance structure was assumed across sites.

Variance components and entry-mean based heritability (h^2) were obtained for GY under low (GYLN) and normal N (GYNN), and PH under normal N (PH), using

$h^2 = \hat{\sigma}_g^2 / \left(\hat{\sigma}_g^2 + \frac{\hat{\sigma}_{GE}^2}{e} + \frac{\hat{\sigma}_E^2}{re} \right)$, where $\hat{\sigma}_g^2$ is the genetic variance, $\hat{\sigma}_{GE}^2$ is the variance due to G x E interaction, $\hat{\sigma}_E^2$ is the residual variance, e is the number of sites (e = 2), and r is the number of replication (r = 1). The significance of the random effects of hybrid was assessed by the Likelihood Ratio Test (LRT), at 5% probability. Phenotypic analyses were performed using ASReml-R package (Butler et al. 2009).

We plotted the BLUP mean values of the genotypes in the favorable (NN; *y* axis) and under low N conditions (LN; *x* axis). Two straight lines drawn at the mean value of each environment distributes the genotypes in four quadrants: those responsive or sensitive to N stress (above or below the mean value of the *x* axis) and those responsive or not to normal N application (above or below the mean value of the *y* axis). The genotypes responsive to normal N application and with a higher BLUP mean value under low N site are in quadrant I.

Maize dataset II

We used 452 maize single-crosses provided by Helix Sementes®, São Paulo, Brazil. The hybrids represent a partial diallel mating design between 128 tropical inbred lines. The experimental design used was randomized complete block with two replications. Trials were carried out in Ipiaçu (18°40'51"S, 49°49'19"W, 443 m) and Patos de Minas (18°35'02"S, 46°28'10"W, 1067 m), at Minas Gerais, and Sertanópolis (23°02'39"S, 51°03'13"W, 390 m) at Paraná, located in Southeast and South regions of Brazil, during the first growing season of 2014/15 from late September to early February. Two-row plots of 5 m spaced 0.70 m were used, and sowing density was about 63,000 kernels per hectare, under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, ton ha^{-1}) and plant height (PH, cm). Plots were mechanically harvested and adjusted to 13% moisture for GY assessment, and PH measured from soil surface to the flag leaf collar on one representative plant within each plot.

We used a linear mixed model to calculate BLUPs for hybrids, including site as fixed effect, and hybrid and interaction as random effects. Heterogeneous residual variance structure was assumed across sites. Variance components and entry-mean based heritability (h^2) were obtained for GY and PH. The significance of the random effect of hybrid was assessed by LRT at 5% probability. Pearson's phenotypic (r_p) correlation coefficients among GY and PH was calculated. Phenotypic analyses were performed using ASReml-R package.

2.2.2. Genotypic dataset

The genotyping of the 49 and 128 tropical inbred lines was performed by Affymetrix® platform, containing about 614,000 SNPs (Unterseer et al. 2014). Markers with low call rate (<95%), minor allele frequency (MAF, <0.05) and heterozygous loci on at least one individual were removed. Imputation was done based on homozigosity of an individual and marker frequency with missed point. High-quality polymorphic SNPs were used to build the artificial 738 and 452 hybrids genomic matrix, deduced by combining the genotypes from its two parents. MAF was conducted over hybrids markers considering the threshold of 0.05, resulting in a total of 146,670 and 52,700 SNPs, respectively.

2.2.3. Selection indices

The BLUP mean of grain yield of each hybrid under low (GYLN) and normal N (GYNN) condition was used to estimate the selection indices NAE, LNTI, LNAE, and HM. After obtaining the values of selection indices, we carried out the deviance analysis (ANADEV). We also calculated Pearson's phenotypic (r_p) correlation coefficients among four selection indices and adjusted mean of grain yield.

Nitrogen use efficiency was assessed by N-agronomic efficiency (NAE) (Craswell and Godwin 1984), as follows: $NAE_i = \frac{GY_{(NN)_i} - GY_{(LN)_i}}{N_{(NN)} - N_{(LN)}}$, where NAE_i is the N-agronomic efficiency of hybrid i; $GY_{(NN)i}$ is the BLUP mean of grain yield (ton ha^{-1}) in the NN condition of hybrid i; $GY_{(LN)i}$ the BLUP mean of grain yield (ton ha^{-1}) in the LN condition of hybrid i; $N_{(NN)}$ is the amount of nitrogen (ton N ha^{-1}) applied in the NN condition; and $N_{(LN)}$ is the amount of nitrogen (ton N ha^{-1}) applied in the LN condition.

In addition, to evaluate N-stress tolerance, the low-N tolerance index (LNTI), described by Miti et al. (2010), was used as follows: $LNTI_i = \left(\frac{GY_{(LN)_i}}{GY_{(NN)_i}}\right) \times 100$, where $LNTI_i$ is the low-N tolerance index of hybrid i; $GY_{(NN)i}$ is the BLUP mean of grain yield in the NN condition of hybrid i; $GY_{(LN)i}$ the BLUP mean of grain yield in the LN condition of hybrid i.

The low-N agronomic efficiency (LNAE), introduced by Wu et al. (2011), was used as follows: $LNAE_i = \left(\frac{GY_{(LN)_i}}{GY_{(NN)_i}}\right) \times GY_{(LN)_i}$, where $LNAE_i$ is the low-N agronomic efficiency of hybrid i; $GY_{(NN)i}$ is the BLUP mean of grain yield in the NN condition of hybrid i; $GY_{(LN)i}$ the BLUP mean of grain yield in the LN condition of hybrid i.

The harmonic mean (HM) was calculated using psych-R package: $HM_i = \frac{2}{\left(\frac{GY_{(NN)i}}{\bar{X}_{(NN)}}\right)^{-1} + \left(\frac{GY_{(LN)i}}{\bar{X}_{(LN)}}\right)^{-1}}$, where HM_i is the harmonic mean of hybrid *i*; $GY_{(NN)i}$ is the BLUP mean of grain yield in the NN condition of hybrid *i*; $GY_{(LN)i}$ is the BLUP mean of grain yield in the LN condition of hybrid *i*; $\bar{X}_{(NN)}$ is the overall BLUP mean of the NN condition; and $\bar{X}_{(LN)}$ is the overall BLUP mean of the LN condition. The expression of HM used was based on the harmonic mean of relative performance of predicted genetic values (Spinelli et al. 2015).

2.2.4. Single- and multi-trait prediction

From the maize dataset I, we used the BLUP mean of GYLN, GYNN, and the combination of the four selection indices to run additive-dominance GBLUP and RKHS for single-trait, and GBLUP and GK for multi-trait. From the maize dataset II, we run the same models described, using BLUP mean of GY and PH.

GBLUP model

Additive-dominance GBLUP for single- and multi-trait (*j* = 1, …, *n* traits) was used by fitting the following model:

$$y = Xb + Z_a a + Z_d d + e \quad (1)$$

where *y* is a vector of BLUP values of hybrids obtained from single-trait or multi-trait, *b* is a vector of fixed effects, *a* is a vector of additive genetic effects of the individuals, *d* is the vector of dominance effects, and *e* is a vector of random residuals. *X*, Z_a and Z_d are the incidence matrices for *b*, *a*, and *d*. The distributions assumed were $a \sim N(0, \sigma_a^2 G_a)$, $d \sim N(0, \sigma_d^2 G_d)$, and $e \sim N(0, \sigma_e^2 I_m)$. G_a and G_d are the additive and dominance genomic relationship matrix (GRM), following the equation: $G_a = \frac{W_A W_A'}{2\sum_{i=1}^{n} p_i(1-p_i)}$ and $G_d = \frac{W_D W_D'}{4\sum_{i=1}^{n} (p_i(1-p_i))^2}$, where p_i is frequency of one allele of the locus *i* and *W* is the matrix of incidence of markers (Da et al. 2014; VanRaden 2008). The W_A matrix was coded as 0 for homozygote A_1A_1, 1 for heterozygote A_1A_2 and 2 for homozygote A_2A_2, for W_D was considered 0 for both homozygotes and 1 to heterozygote.

All variance components were determined using Bayesian generalized linear regression (BGLR) (Perez and de los Campos 2014) and MultiTrait Model (MTM) package (de los Campos and Grüneberg 2016). We reported mean estimates and standard deviations of the additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2). Moreover, we calculated in ASReml-R the genetic (r_g) correlation coefficients of traits in the maize dataset I and II, following the equation: $COV_{12} / \sqrt{\sigma_{g_{11}} \sigma_{g_{22}}}$, where COV_{12} is the SNP additive genetic covariance for multiple traits; $\sigma_{g_{11}}$ and $\sigma_{g_{22}}$ is the SNP additive variance associated to each trait.

RKHS and GK models

Additive-dominance RKHS and GK was used for single- and multi-trait model (j = 1,…, n traits), respectively, by fitting the following model:

$$y = Xb + Z_a a + Z_d d + e \tag{2}$$

where y is a vector of BLUP values of hybrids obtained from single-trait or multi-trait,, b is a vector of fixed effects, a is a vector of additive genetic effects of the individuals, d is the vector of dominance effects, and e is a vector of random residuals. X, Z_a and Z_d are the incidence matrices for b, a, and d. The distributions assumed were $a \sim N(0, \sigma_a^2 K_a)$, $d \sim N(0, \sigma_d^2 K_d)$, and $\varepsilon \sim N(0, \sigma_e^2 I_m)$. K_a and K_d are the additive and dominance symmetric semipositive definite matrix representing the covariance of the genetic values, following the equation: $K_a = exp(-hd_{a_i}^2 / q_{0.05})$ and $K_d = exp(-hd_{d_i}^2 / q_{0.05})$, where h is a bandwidth parameter, estimated from the Bayesian method (Cuevas et al. 2016); $d_{a_i}^2$ and $d_{d_i}^2$ are the squared Euclidean distance based on a centered and standardized additive and dominance incidence matrix, respectively, between individuals i (Morota et al. 2014); and $q_{0.05}$ is the fifth percentile of the same distance. The bandwidth parameter (h) for RKHS was estimated using one trait, and for GK we averaged the h of two traits.

We reported posterior mean estimates and standard deviations of the additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2), using BGLR-R and MTM-R. We used a total of 30,000 MCMC iterations, 5,000 for burn-in, and 5 for thinning. In the multivariate model, we assumed an unstructured genetic and residual covariance matrix. The degree of freedom hyperparameters of the scaled inverse chi-square distributions and the scale parameters were all set to the total number of traits used, which in our case was two.

Validation and prediction accuracy

From GP models, we evaluated prediction accuracy (r_{MP}), correlation between BLUP values and predicted phenotypic values of the hybrids, from fifty replications, randomly sampling 75% of the hybrids to form the training set (TS) and the rest as validation set (VS). We reported the average correlation and approximate *P*-values for pairwise comparisons of prediction accuracy between single- and multi-trait models based on paired *t*-tests applied after using Fisher's Z transformation.

2.3. RESULTS

Prediction of SIs

From the phenotypic analysis, it was found significant differences between hybrids by the LRT ($P<0.05$) for GYLN, GYNN, and PH. Entry-mean based heritability was 0.43 for GYLN, 0.59 for GYNN, and 0.83 for PH, reflecting good accuracy of phenotypic evaluation. The average of BLUP mean for GYLN was 5.30, ranging from 3.14 to 8.12 ton ha^{-1}, and for GYNN was 6.54, ranging from 3.56 to 10.09 ton ha^{-1} (Fig. 1), with a reduction of 18.96% under NN compared to LN condition. We identified 273 hybrids in quadrant I (Fig. 1a) responsive to normal N condition and presented higher mean in the unfavorable environment (low N condition). In quadrant III, 258 hybrids presented low grain yield evaluated in low and normal condition of N.

To highlight the importance of the selection based on the environment mean, Fig. 1b shows the relation between the means of the ten best maize hybrid selected based on the mean under low N condition (mean LN), on the mean of all environments (mean ENV) and in the normal N condition (mean NN). The mean LN of the best hybrids selected was superior under low N and inferior under a favorable environment. The hybrids selected based on the means of the environments, presented lower-rate and higher-rate performance in the unfavorable and favorable conditions, respectively.

From the selection indices, the hybrid effect was significant for NAE, LNAE, LNTI, and HM. The NAE mean was 17.68, ranging from -25.73 to 45.98 ton ha^{-1}. The LNTI mean was 81.74 %, and varied from 55.62 to 149.21 %. The LNAE mean was 4.35, and ranged from 2.02 to 8.14 ton ha^{-1}, while HM was 0.99, and varied from 0.59 to 1.53 among the genotypes (Supplemental Fig. S1).

Phenotypic correlations (r_p) of GYLN showed significant positive correlations with all other traits (r_p=0.25 - 0.89), except for NAE (Table 1; Supplemental Fig. S1). Significant positive

correlations were found among GYNN with all selection indices, except for LNTI. NAE was negatively correlated with LNTI and LNAE. From the genetic correlations (r_g) we found GYLN negatively correlated with LNTI, and LNAE non-significantly correlated with NAE (r_g=0.16), reflecting differences compared to phenotypic correlation. We found that GYNN correlated significantly (P<0.05) with PH (r_p = 0.45; r_g = 0.75).

From single- and multi-trait GBLUB analyses, we obtained the values of additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2) for each trait and model (Table 2). Estimates of variance components and genomic heritability varied considerable among single- and multi-trait model. Single-trait h_g^2 ranged from 0.24 to 0.38, and for multi-trait h_g^2 from 0.39 to 0.88. The highest heritability estimate for the selection indices in single- and multi-trait was 0.38 for HM and 0.87 for NAE.

From single-trait RKHS and multi-trait GK, we found considerable differences in estimates of variance components and genomic heritability among traits and models (Table 3). It is important to highlight, for both models, dominance variance was higher for all traits, and h_g^2 ranged from 0.28 to 0.54 for single-trait, and 0.39 to 0.88 for multi-trait. The highest heritability estimate for single-trait was obtained for HM, and multi-trait was NAE.

From the single-trait GBLUP model, the r_{MP} varied from 0.12 to 0.40, with the highest value obtained for GYNN (0.39) and HM (0.40) (Fig. 2a). From multi-trait GBLUP model, the value of r_{MP} remained the same for HM (0.40) and reduced for NAE (0.14) and LNAE (0.12) (Fig. 3a), when combined to each one. However, increased for LNTI (0.16) combined with HM (0.40), NAE (0.21) with LNTI (0.16), and LNTI (0.16) with LNAE (0.15).

For RKHS model, we found the highest value of r_{MP} for GYNN (0.41) and HM (0.41) (Fig. 2b). From GK model, the value of r_{MP} remained the same for the combination GYLN and GYNN, and increased for NAE (0.21) combined with LNTI (0.15), NAE (0.22) with LNAE (0.14), and LNTI (0.18) with LNAE (0.14) (Fig. 3b). The r_{MP} of GK was slightly higher compared to GBLUP for some traits, such as HM combined to NAE, LNTI, and LNAE (Fig. 3).

Prediction of GY and PH

From the phenotypic analysis of maize dataset II, it was found significant genetic differences between hybrids by LRT (P<0.05), for both traits. Entry-mean based heritability was 0.62 and 0.86 for GY and PH, respectively. The average of BLUP mean for GY was 6.6, and varied from 1.53 to 10.08 ton ha^{-1}, and for PH was 240, varying from 185 to 277 cm. The correlation showed that GY correlated significantly (P<0.01) with PH (r_p = 0.40; r_g = 0.53).

From single-trait GBLUP model, estimates of genomic broad-sense heritability and prediction accuracy was 0.38; 0.39 (GY dataset I), 0.71; 0.55 (PH dataset I), 0.77; 0.70 (GY dataset II), and 0.87; 0.80 (PH dataset II) (Fig. 4a). For multi-trait GBLUP model, a significant increase was identified for GY in dataset I (738H). On the other hand, a slight increase of r_{MP} was observed for GY in dataset II (452H) (Fig. 4a). For single-trait RKHS model, estimates of r_{MP} was similar to GBLUP, except for GY 738H (Fig. 4b). For multi-trait GK model, a significant increase also was identified for GY in dataset I, from 0.41 to 0.43 (P<0.05), and for dataset II, a slight increase of r_{MP} was observed for GY and PH.

2.4. DISCUSSION

In the present study, we identified superior maize hybrids with higher mean in low N and responsive to normal N condition (Fig. 1 - quadrant I). Tropical maize germplasm generally presents N genetic variability due to farming systems with low mineral nutrient availability, demonstrating potential to identify best individuals through high general combining ability and/or selection indices (Kumar et al. 2016; Trachsel et al. 2016). In our work, we used four selection indices and found genetic variation from significant effect of hybrid. Recent studies have demonstrated the efficiency of SIs to identify *N stress tolerant* and *N-use efficient* genotypes (Abdel-Ghani et al. 2013; Granato et al. 2014; Khan and Mohammad 2016).

We identified significant positive genetic correlation of HM (r_g = 0.91; r_g = 0.98) and LNAE (r_g = 0.87; r_g = 0.56) with grain yield under low and normal N sites (Table 1), respectively, indicating the high performance of these indices to identify genotypes in both conditions. On the other hand, NAE was non-significant negative and positive correlated (r_g = -0.01; r_g = 0.88) and LNTI was negative associated with GYLN and GYNN, respectively. These findings, according to Wu et al. (2011), could result in the selection of genotypes with great difference between low and normal N conditions. In contrary, HM is highly associated to the environment (arithimetic) mean and geometric mean, and is recommended to select best genotypes adapted to stress and non-stress conditions (Jafari et al. 2009), which is in agreement to our study (Table 1, Fig. 1b).

We applied single- and multi-trait genomic prediction using SIs to identify superior maize hybrids. For single-trait analysis, we found a high genomic broad sense heritability for HM in GBLUP (h_g^2 = 0.38) and RKHS (h_g^2 = 0.54) (Table 2), compared to the remaining indices, reflecting in a feasible trait for selection. In contrast, we identified lower genomic heritability for GY under low N availability (Table 2). Likewise, Beyene et al. (2015) found reduction in the estimate of heritability between grain yield under water-limited and water-sufficient sites, across a

diverse set of bi-parental maize populations. Several authors argued that unfavorable conditions interfere in the metabolism of growth (Cooper et al. 2014; Harfouche et al. 2014), reducing the expression of genetic variability, and thus affecting heritability.

For single-trait RKHS analysis, we found higher dominance variances for all traits (Table 3), which is in agreement with those of Wang et al. (2016) and dos Santos et al. (2016), that adding this effect substantially contributed to rice and maize hybrid prediction, respectively. In terms of prediction accuracy (r_{MP}), we observed considerable differences between GYLN (0.33; 0.33) and GYNN (0.38; 0.41), for single-trait GBLUP and RKHS models, respectively. Similar reduction of r_{MP} were observed by Crossa et al. (2010) and Poland et al. (2012) for yield in maize and wheat, under water-limited and water-sufficient environments. In addition, Ziyomo and Bernardo (2013) also observed lower r_{MP} of grain yield under water-limited (0.46) and well-watered (0.48) sites, in a population of maize inbred lines. According to the authors, unfavorable conditions could affect heritability, and consequently, interfere in the estimates of prediction accuracy.

For the selection indices, HM presented higher estimate of r_{MP} (0.40; 0.41) in both single-trait models tested, suggesting selection could be performed on this trait in predictive analysis. Similarly, previous studies have shown that prediction using arithmetic mean of two environments is better than single-environment (Wang et al. 2016). We observed a slight increase of r_{MP} for GYNN, NAE, LNTI, and HM, using single-trait RKHS relative to GBLUP (Fig. 2). Similar findings were observed by Crossa et al. (2010) and Jiang and Reif (2015), showing better model performance for RKHS, which could be related to the genetic architecture of the traits modeled in the kernels.

Our main strategy was to combine different selection indices in the MTGP analyses, attempting to explore the correlation between traits. We identified a slight increase of r_{MP} comparing single- to multi-trait GBLUP (Fig. 2a; Fig. 3a), and single- RKHS to multi-trait GK (Fig. 2b; Fig. 3b) models. Several studies showed the same trend, which the success of the MTGP is highly dependent of genetically correlated traits, and a target trait with low heritability (Jia and Jannink 2012; Schulthess et al. 2016; Wang et al. 2016). In contrast, Schulthess et al. (2016) used SIs in MTGP for grain yield and protein content in rye, and recommended to perform single-trait prediction. In addition, a few authors argue the benefits of using MTGP over STGP, due to the computational issues and real gains (Alimi et al. 2013; Lee and van der Werf 2016). In our multi-trait analyses, we found an increase of r_{MP} for the SIs with low genomic heritability when combined to moderate h_g^2 (Table 3), for example, combining HM (h_g^2 = 0.54) and LNAE (h_g^2 = 0.28), resulted in an increase of r_{MP} from 0.13 to 0.15 for LNAE, and for HM it remained the

same. Similar results was obtained by Hayashi and Iwata (2013) and Guo et al. (2014), that found accurate GEBV prediction for low heritability traits correlated with high heritability traits.

From the MTGP analyses, GK outperformed GBLUP, showing higher or equal prediction accuracy for all traits, and lower standard errors. Recently, additive GK models were applied by Cuevas et al. (2016) and Cuevas et al. (2017) for multi-environments in wheat and maize datasets, and the authors attributed the better performance relative to GBLUP, due to at least two reasons, (i) more flexibility of the kernels to model complex marker main effects and (ii) marker-specific interaction effects. Liu et al. (2016) working with genomic prediction for rice NUE breeding, found slight higher r_{MP} of GK compared to GBLUP.

From the prediction of grain yield and plant height (Fig. 4), we used two different datasets applying single- and multi-trait GP. A considerable difference was observed in the broad sense genomic heritability and prediction accuracy of both datasets for GY (dataset I, h_g^2 = 0.38, r_{MP} = 0.39) and GY (dataset II, h_g^2 = 0.77, r_{MP} = 0.70), and to PH (dataset I, h_g^2 = 0.71, r_{MP} = 0.55) and PH (dataset II, h_g^2 = 0.87, r_{MP} = 0.80) from GBLUP model. We observed that higher genomic heritability is associated to higher prediction accuracy. However, according to de los Campos et al. (2015) prediction and genomic heritability are two different problems, and warned about the true proportion of variance that can be explained by a regression on markers.

It is important to highlight dataset I has more individuals than dataset II, and we observed lower r_{MP} in this dataset for GY and PH (Fig. 4). According to several authors, increasing the training set size can improve the prediction accuracies (Lian et al. 2014; Mendes and de Souza 2016). On the other hand, simulations showed, in some cases, small TSs can be just as accurate as larger TS (Habier et al. 2009). Besides TS size, several factors can affect accuracy such as number of markers, trait heritability, effective population size, and relationship between test and validation set (Daetwyler et al. 2008; Habier et al. 2007; Riedelsheimer et al. 2013). In addition, measures as prediction error variance and coefficient of determination are well adopted to optimize TS (Isidro et al. 2015; Rincent et al. 2012). In our study, phenotypic precision (experimental design) may have led to the lower r_{MP} of dataset I.

We also observed a significant increase of r_{MP} for GY in dataset I when combined to PH, in multi-trait GBLUP and GK relative to single-trait GBLUP and RKHS, respectively (Fig. 4). On contrast to dataset II, we observed a non-significant increase of r_{MP}, for GY and PH comparing single- and multi-trait models. Similar results was observed by dos Santos et al. (2016), which did not find any improvements in r_{MP} between plant height and kernel weight using multivariate models. The authors argued this result was not expected, since both traits are moderately correlated. Bao et al. (2015) also found no benefits of using MTGP over STGP

models in four traits of soybean. However, these discouraging results was argued by Schulthess et al. (2016), emphasizing MTGP could also be applied in a way to achieve more cycles of selection by unit of time similarly as STGP.

2.5. CONCLUSION

Our results suggest that the best approach is predicting hybrids based on harmonic mean, since it take into account the performance under optimal and nitrogen stress conditions. Furthermore, the combination of selection indices by multi-trait genomic prediction seems to be suitable, due to the increase of prediction accuracies. However, adding grain yield and plant height in MTGP showed a slight improvement in prediction accuracy compared to single-trait genomic prediction.

The overall performance of the nonlinear GK model was superior relative to multi-trait GBLUP. On other hand, there was a modest benefit of using single-trait RKHS and GK multi-trait, rather than GBLUP.

REFERENCES

Abdel-Ghani AH, Kumar B, Reyes-Matamoros J, Gonzalez-Portilla PJ, Jansen C, San Martin JP, Lee M, Lubberstedt T (2013) Genotypic variation and relationships between seedling and adult plant traits in maize (*Zea mays* L.) inbred lines grown under contrasting nitrogen levels. Euphytica 189:123–133

Alimi NA, Bink MC, Dieleman JA, Magan JJ, Wubs AM, Palloix A, van Eeuwijk FA (2013) Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theor Appl Genet 126 (10):2597-2625

Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54:68-75

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, Gakunga J, Vargas M, Olsen M, Prasanna BM, Banziger M, Crossa J (2015) Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci 55 (1):154-163

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual. Department of Primary Industries, Queensland, Australia

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43 (26):1-14

Cantelmo NF, Von Pinho RV, Balestre M (2017) Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. Mol Breeding 37 (4):51

Cecarelli S, Grando S, Impiglia A (1998) Choice of selection strategy in breeding barley for stress environments. Euphytica 103:307-318

Ceron-Rojas JJ, Crossa J, Arief VN, Basford K, Rutkoski J, Jarquin D, Alvarado G, Beyene Y, Semagn K, DeLacy I (2015) A genomic selection index applied to simulated and real data. G3-Genes Genom Genet 5 (10):2155-2164

Cerón-Rojas JJ, Crossa J, Toledo FH, Sahagún-Castellanos J (2016) A predetermined proportional gains eigen selection index method. Crop Sci 56:2436-2447

Chen K, Camberato JJ, Tuinstra MR, Kumudini SV, Tollenaar M, Vyn TJ (2016) Genetic improvement in density and nitrogen stress tolerance traits over 38 years of commercial maize hybrid release. Field Crop Res 196:438-451

Cooper M, Gho C, Leafgren R, Tang T, Messina C (2014) Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. J Exp Bot 65 (21):6191-6204

Craswell ET, Godwin DC (1984) The efficiency of nitrogen fertilizers applied to cereals grown in different climates. In: Tinker PB, Lauchli A (eds) Advances in plant nutrition Praeger Publishers, New York 1:1-55

Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186 (2):713-724

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 112 (1):48-60

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P, de los Campos G (2017) Bayesian genomic prediction with genotype x environment interaction kernel models. G3-Genes Genom Genet 7:41-53

Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, de los Campos G, Montesinos-López OA, Burgueño J (2016) Genomic prediction of genotype x environment interaction kernel regression models. The plant genome 9 (3):1-20

Da Y, Wang CK, Wang SW, Hu G (2014) Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. Plos One 9 (1):e87666

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One 3 (10):e3395

de los Campos G, Grüneberg A (2016) MTM (Multiple-Trait Model) package. Available at: http://quantgen.github.io/MTM/vignette.html.

de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? Plos Genet 11 (5):e1005048

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124 (6):331-341

dos Santos JPR, Vasconcellos RCD, Pires LPM, Balestre M, Von Pinho RG (2016) Inclusion of dominance effects in the multivariate GBLUP model. Plos One 11 (4):1-21

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

Fritsche-Neto R, DoVale JC, Lanes ECM, Resende MDV, Miranda GV (2012) Genome-wide selection for tropical maize root traits under conditions of nitrogen and phosphorus stress. Acta Scientiarum 34 (4):389-395

Gong FP, Wu XL, Zhang HY, Chen YH, Wang W (2015) Making better maize plants for sustainable grain production in a changing climate. Front Plant Sci 6:1-6

Granato ISC, Bermudez FP, Reis GG, Dovale JC, Miranda GV, Fritsche-Neto R (2014) Index selection of tropical maize genotypes for nitrogen use efficiency. Bragantia 73 (2):153-159

Guo G, Zhao FP, Wang YC, Zhang Y, Du LX, Su GS (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15 (30):1-7

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, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182 (1):343-353

Harfouche A, Meilan R, Altman A (2014) Molecular and physiological responses to abiotic stress in forest trees and their relevance to tree improvement. Tree physiology 34 (11):1181-1198

Hayashi T, Iwata H (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC bioinformatics 14:34

He D, Kuhn D, Parida L (2016) Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. Bioinformatics 32 (12):i37-i43

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127 (2):463-480

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

Jafari A, Paknejad F, Al-Ahmadi MJ (2009) Evaluation of selection indices for drought tolerance of corn (*Zea mays* L.) hybrids. Int J Plant Prod 3 (4):33-38

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192 (4):1513–1522

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201 (2):759-768

Jombart T, Collins C, Kamvar ZN, Lustrik R, Solymos P, Ahmed I, Jombart MT (2015) adegenet: exploratory analysis of genetic and genomic data. R Package Version 201

Khan FU, Mohammad F (2016) Application of stress selection indices for assessment of nitrogen tolerance in wheat (*Triticum aestivum* L.). J Anim Plant Sci 26 (1):201-210

Kumar B, Guleria SK, Khanorkar SM, Dubey RB, Pater J, Kumar V, Parihar CM, Jat SL, Singh V, Yatish KR, Das A, Sekhar JC, Bhati P, Kaur H, Kumar M, Singh AK, Varghese E, Yadav OP (2016) Selection indices to identify maize (Zea mays L.) hybrids adapted under drought-stress and drought-free conditions in a tropical climate. Crop Pasture Sci 67 (10):1087-1095

Lee SH, van der Werf JH (2016) MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32 (9):1420-1422

Lian L, Jacobson A, Zhong S, Bernardo R (2014) Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci 54 (4):1514-1522

Liu ZY, Zhu CS, Jiang Y, Tian YL, Yu J, An HZ, Tang WJ, Sun J, Tang JP, Chen GM, Zhai HQ, Wang CM, Wan JM (2016) Association mapping and genetic dissection of nitrogen use efficiency-related traits in rice (*Oryza sativa* L.). Funct Integr Genomic 16 (3):323-333

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

Mendes MP, de Souza CL (2016) Genomewide prediction of tropical maize single-crosses. Euphytica 209 (3):651-663

Miti F, Tongoona P, Derera J (2010) S_1 selection of local maize landraces for low soil nitrogen tolerance in Zambia. Afr J Plant Sci 4:67-81

Montesinos-Lopez OA, Montesinos-Lopez A, Crossa J, Toledo FH, Perez-Hernandez O, Eskridge KM, Rutkoski J (2016) A Genomic bayesian multi-trait and multi-environment model. G3-Genes Genom Genet 6 (9):2725-2744

Morota G, Boddhireddy P, Vukasinovic N, Gianola D, DeNise S (2014) Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. Front Genet 5:56

Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Front Genet 5:363

Mueller SM, Vyn TJ (2016) Maize plant resilience to N stress and post-silking N capacity changes over time: a review. Front Plant Sci 7 (53):1-38

Pérez-Elizalde S, Cuevas J, Pérez-Rodríguez P, Crossa J (2015) Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. J Agric Biol Environ Stat 20:512–532

Perez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198 (2):483-U463

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

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 (2):217-220

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194 (2):493-503

Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen CC, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192 (2):715-728

Schulthess AW, Wang Y, Miedaner T, Wilde P, Reif JC, Zhao YS (2016) Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor Appl Genet 129 (2):273-287

Spinelli VM, Dias LAS, Rocha RB, Resende MDV (2015) Estimates of genetic parameters with selection within and between half-sib families of *Jatropha curcas* L. Ind Crop Prod 69:355-361

Trachsel S, Leyva M, Lopez M, Suarez EA, Mendoza A, Montiel NG, Macias MS, Burgueno J, San Vicente F (2016) Identification of tropical maize germplasm with tolerance to drought, nitrogen deficiency, and combined heat and drought stresses. Crop Sci 56 (6):3031-3045

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, Bertani C, Davassi A, Mayer KFX, Schon CC (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15 (823):1-15

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91 (11):4414-4423

Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, Hu Z (2016) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity 118:302-310

Wu YS, Liu WG, Li XH, Li MS, Zhang DG, Hao ZF, Weng JF, Xu YB, Bai L, Zhang SH, Xie CX (2011) Low-nitrogen stress tolerance and nitrogen agronomic efficiency among maize inbreds: comparison of multiple indices and evaluation of genetic variation. Euphytica 180 (2):281-290

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28 (24):3326-3328

Ziyomo C, Bernardo R (2013) Drought tolerance in maize: indirect selection through secondary traits versus genomewide selection. Crop Sci 53 (4):1269-1275

FIGURES

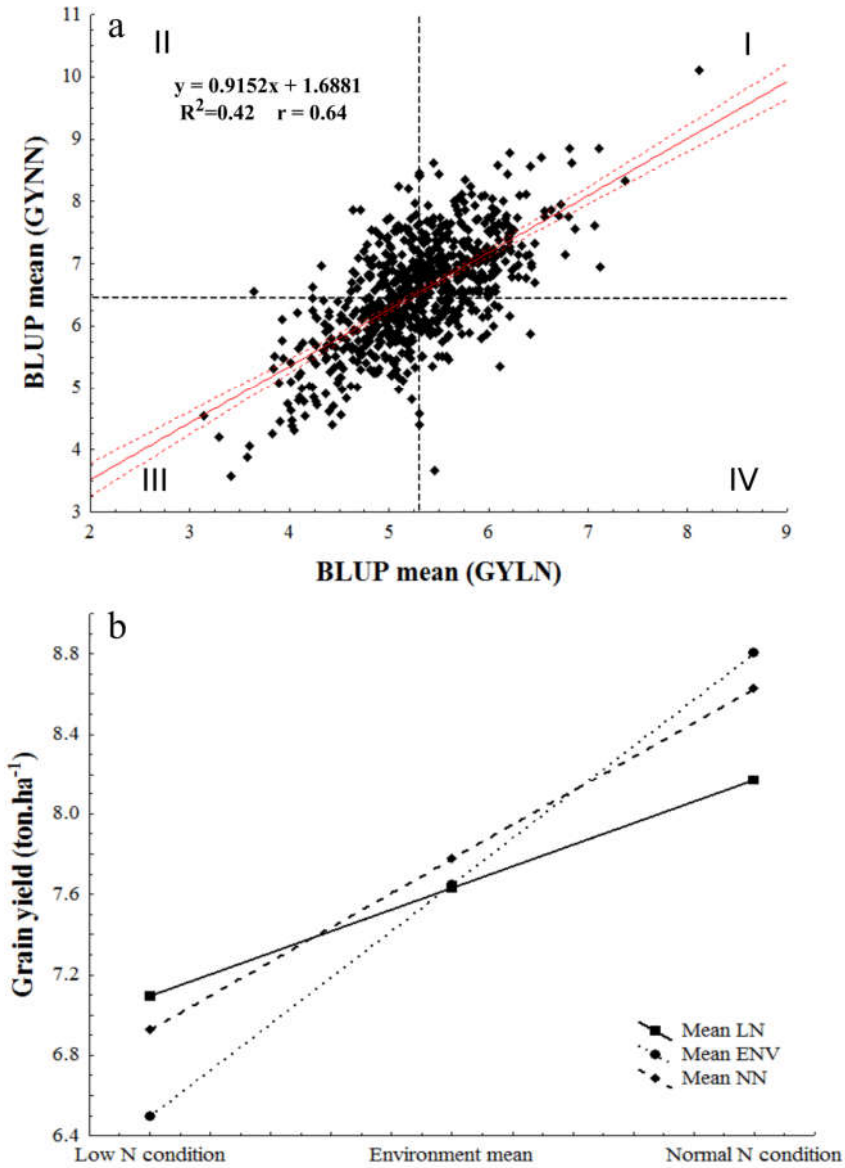


Fig. 1. Performance of 738 tropical maize hybrids under low and normal N condition. (a) Relationship between BLUP mean of grain yield under low (GYLN) and normal N (GYNN) conditions. Dashed black lines represents the mean. Solid red line is the regression slope and 95% confidence interval (red band). Quadrants represents efficient and responsive (I), non-efficient and responsive (II), non-efficient and non-responsive (III), and efficient and non-responsive (IV). (b) Performance of the ten best hybrids selected based on the mean under low N condition (▬), on the mean of all environments (···●···), and in the normal N condition (-·-◆-·-), for grain yield (ton ha^{-1}).

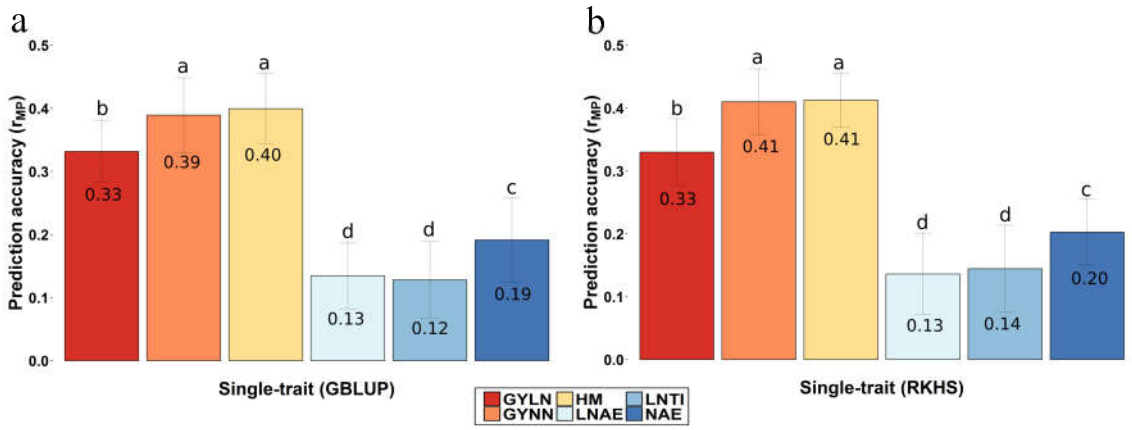


Fig. 2. Barplot of prediction accuracy from single-trait analyses. (a) GBLUP and (b) RKHS, for grain yield (ton ha^{-1}) in low (GYLN) and normal (GYNN) nitrogen, N-agronomic efficiency (NAE, ton ton^{-1} N ha^{-1}), low-N tolerance index (LNTI, %), low-N agronomy efficiency (LNAE, ton ha^{-1}), and harmonic mean (HM). Data are mean ± SD estimated from fifty replications in independent validation. Different letters above bars indicate significant differences (P<0.05) between prediction accuracy for six traits from paired t-tests.

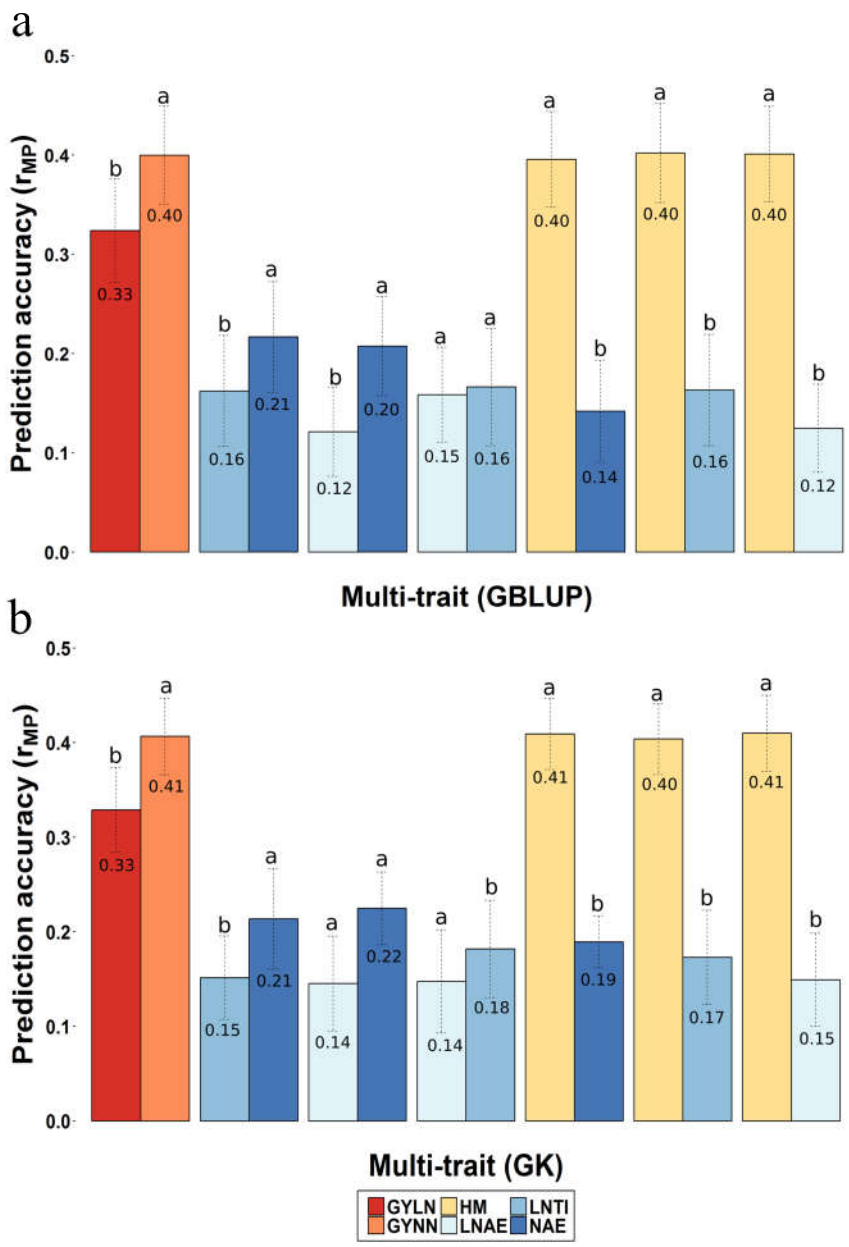


Fig. 3. Barplot of prediction accuracy from multi-trait analyses. (a) GBLUP and (b) GK, for grain yield (ton ha^{-1}) in low (GYLN) and normal (GYNN) nitrogen, N-agronomic efficiency (NAE), low-N tolerance index (LNTI), low-N agronomy efficiency (LNAE), and harmonic mean (HM). Data are mean ± SD estimated from fifty replications in independent validation. Different letters above bars indicate significant differences (P<0.05) between prediction accuracy for the combination of traits in MTGP from paired t-tests.

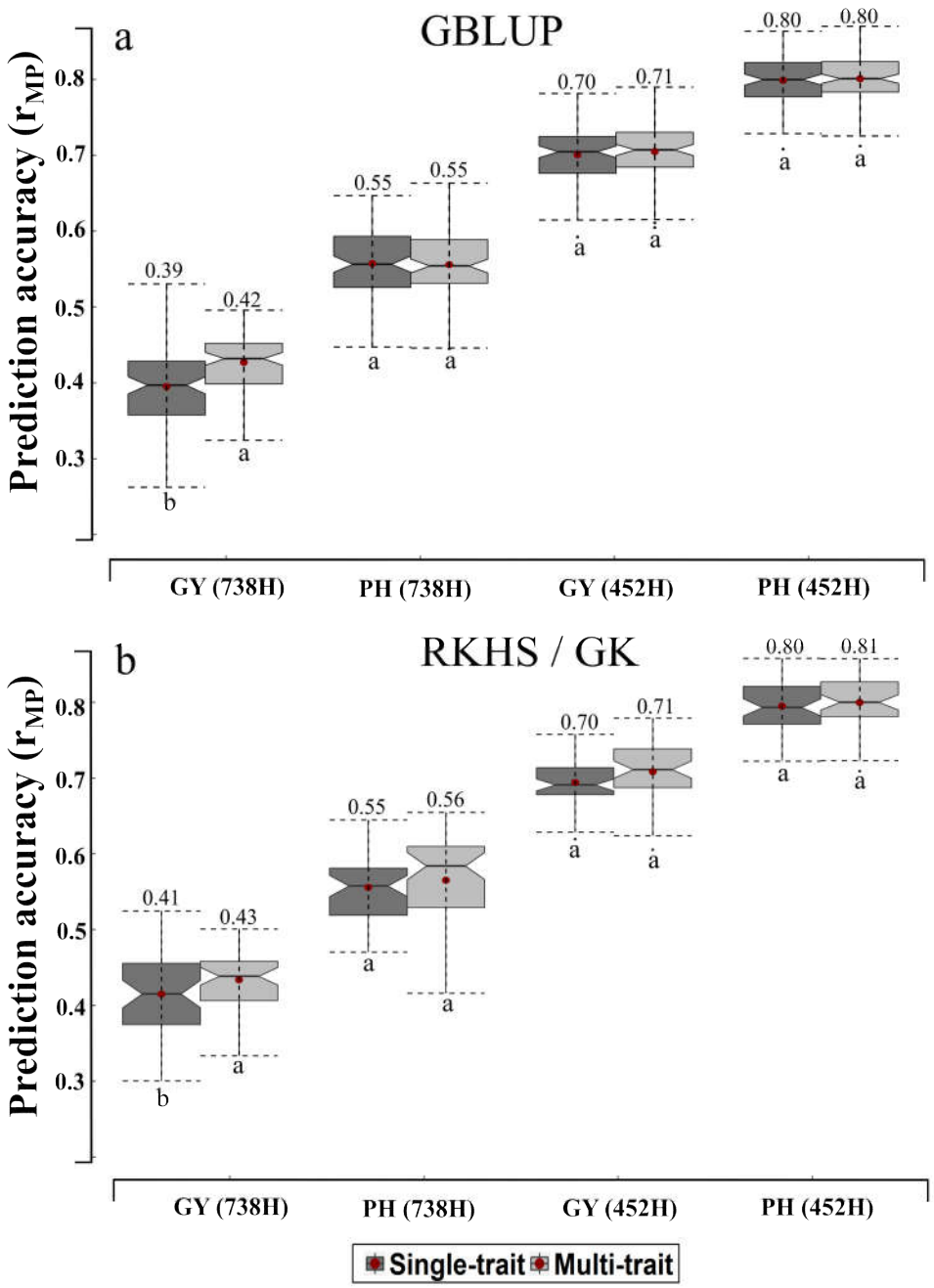
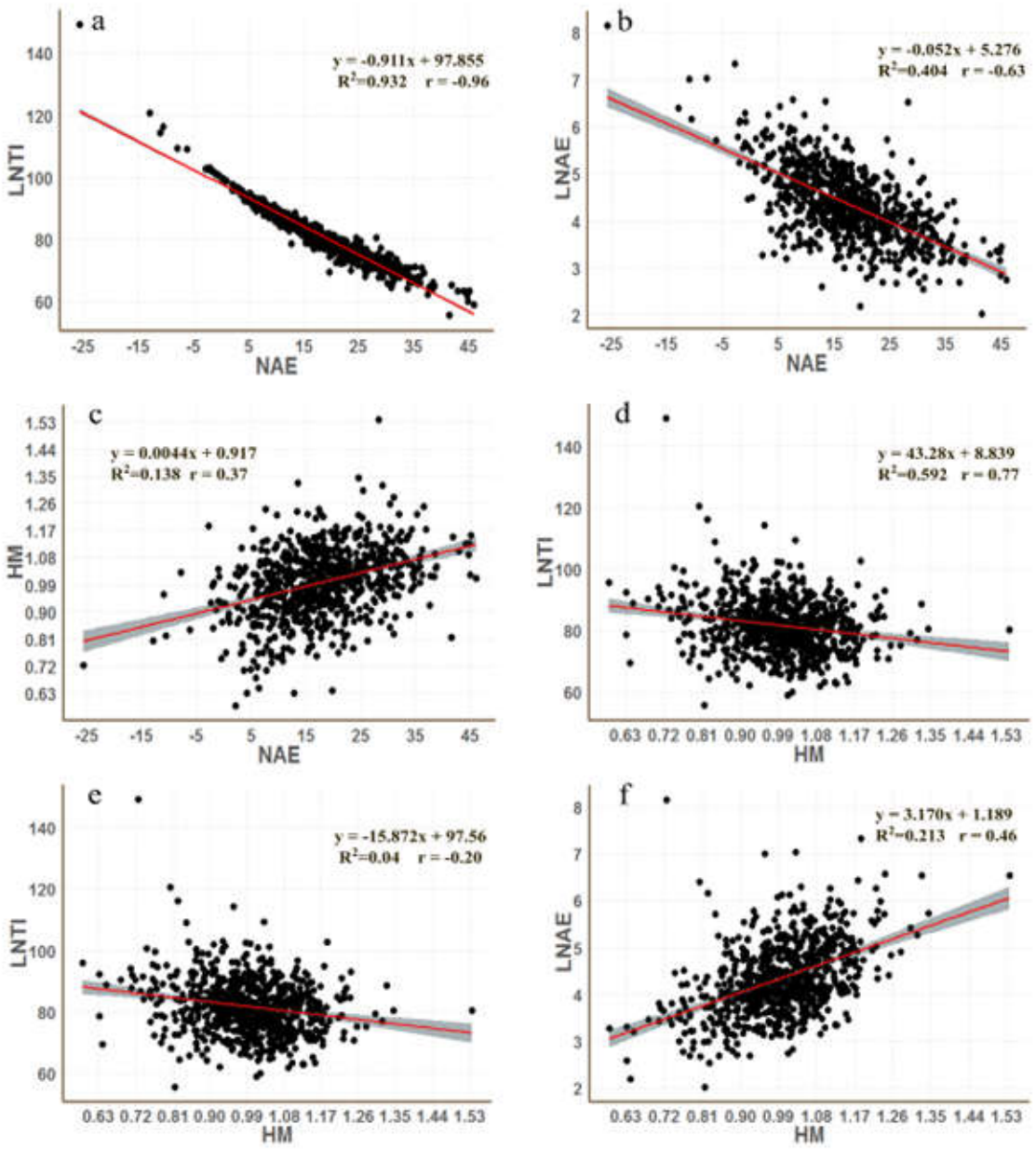


Fig. 4. Boxplot of prediction accuracy. (a) Single- and multi-trait additive-dominance GBLUP, (b) single-RKHS and multi-GK, under dataset I (738 Hybrids) and dataset II (452 Hybrids), for grain yield (GY) and plant height (PH). Red dot and black number are representing the mean. Different letters bellow box indicate significant differences (P<0.05) between prediction accuracy for the combination of traits in STGP and MTGP from paired t-tests.



Supplemental Fig. S1. Scatterplot between the combinations of four selection indices (a-f). N-agronomic efficiency (NAE, ton ton^{-1} N ha^{-1}), low-N tolerance index (LNTI, %), low-N agronomy efficiency (LNAE, ton ha^{-1}), and harmonic mean (HM). Solid red line is the regression slope and 95% confidence interval (light gray band).

TABLES

Table 1 Estimate of phenotypic (r_p, above the diagonal) and genetic (r_g, below the diagonal) correlations

Trait	GYLN	GYNN	NAE	LNTI	LNAE	HM
GYLN	-	0.64**	-0.07*	0.25**	0.80**	0.89**
GYNN	0.93**	-	0.70**	-0.55**	0.08*	0.92**
NAE	-0.01[ns]	0.88**	-	-0.96**	-0.63**	0.37**
LNTI	-0.59**	-0.80**	-0.94**	-	0.77**	-0.20**
LNAE	0.87**	0.56**	0.16[ns]	0.33**	-	0.46**
HM	0.91**	0.98**	0.73**	-0.57**	0.70**	-

Traits are grain yield (ton ha^{-1}) in low (GYLN) and normal (GYNN) nitrogen, N-agronomic efficiency (NAE, ton ton^{-1} N ha^{-1}), low-N tolerance index (LNTI, %), low-N agronomy efficiency (LNAE, ton ha^{-1}), and harmonic mean (HM, ton ha^{-1}).
[ns]Not significant; significant at 5% (*) or 1% (**) level.

Table 2 Estimate of variance components and genetic parameters obtained by single- and multi-trait GBLUP analyses

	Trait	σ_a^2	σ_d^2	σ_ε^2	h_g^2
GBLUP single-trait	GYLN	0.07±0.02[a]	0.05±0.01	0.29±0.02	0.31±0.05
	GYNN	0.19±0.05	0.12±0.03	0.50±0.04	0.38±0.05
	NAE	<0.01±0.00	<0.01±0.00	<0.01±0.00	0.29±0.05
	LNTI	11.61±3.06	15.10±4.44	65.10±5.73	0.29±0.05
	LNAE	0.07±0.02	0.09±0.03	0.51±0.04	0.24±0.04
	HM	0.003±0.00	0.001±0.00	0.008±0.00	0.38±0.05
GBLUP multi-trait	GYLN	0.22±0.08	0.58±0.40	0.29±0.02	0.71±0.09
	GYNN	0.45±0.19	1.38±1.05	0.54±0.05	0.75±0.11
	NAE	0.007±0.00	0.007±0.00	0.003±0.00	0.80±0.01
	LNTI	37.3±8.25	15.6±7.91	52.8±3.42	0.48±0.08
	NAE	0.006±0.00	0.01±0.00	0.003±0.00	0.87±0.01
	LNAE	0.12±0.04	0.29±0.17	0.53±0.04	0.43±0.09
	LNTI	31.2±7.70	6.38±5.03	51.1± 3.60	0.44±0.06
	LNAE	0.11±0.03	0.25±0.12	0.53±0.04	0.39±0.08
	NAE	0.008±0.00	0.01±0.00	0.002±0.00	0.88±0.01
	HM	0.02±0.00	0.04±0.00	0.001±0.00	0.86±0.02
	LNTI	36.9±9.77	14.3±9.77	52.5±3.45	0.50±0.07
	HM	0.01±0.00	0.03±0.00	0.009±0.00	0.85±0.02
	LNAE	0.14±0.05	0.32±0.19	0.54±0.04	0.45±0.10
	HM	0.02±0.00	0.03±0.00	0.01±0.00	0.85±0.02

Traits are grain yield (ton ha^{-1}) in low (GYLN) and normal (GYNN) nitrogen, N-agronomic efficiency (NAE, ton ton^{-1} N ha^{-1}), low-N tolerance index (LNTI, %), low-N agronomy efficiency (LNAE, ton ha^{-1}), and harmonic mean (HM, ton ha^{-1}). Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2).
[a]Data are mean ± standard deviation (SD) estimated from fifty replications in independent validation.

Table 3 Estimate of variance components and genetic parameters obtained by single-trait RKHS and multi-trait GK analyses

	Trait	σ_a^2	σ_d^2	σ_ε^2	h_g^2
RKHS single-trait	GYLN	0.10 ± 0.03[a]	0.11 ± 0.05	0.28 ± 0.02	0.43 ± 0.06
	GYNN	0.22 ± 0.06	0.26 ± 0.13	0.43 ± 0.03	0.52 ± 0.07
	NAE	<0.01 ± 0.00	<0.01 ± 0.00	<0.01 ± 0.00	0.31 ± 0.06
	LNTI	11.89 ± 3.92	17.77 ± 7.03	61.09 ± 5.56	0.32 ± 0.07
	LNAE	0.09 ± 0.03	0.10 ± 0.03	0.49 ± 0.02	0.28 ± 0.05
	HM	0.003 ± 0.00	0.005 ± 0.00	0.007 ± 0.00	0.54 ± 0.07
GK multi-trait	GYLN	0.22 ± 0.08	0.58 ± 0.40	0.29 ± 0.02	0.71 ± 0.09
	GYNN	0.45 ± 0.19	1.38 ± 1.05	0.54 ± 0.05	0.75 ± 0.11
	NAE	0.007 ± 0.00	0.007 ± 0.00	0.003 ± 0.00	0.80 ± 0.01
	LNTI	37.3 ± 8.25	15.6 ± 7.91	52.8 ± 3.42	0.48 ± 0.08
	NAE	0.006 ± 0.00	0.01 ± 0.00	0.003 ± 0.00	0.87 ± 0.01
	LNAE	0.12 ± 0.04	0.29 ± 0.17	0.53 ± 0.04	0.43 ± 0.09
	LNTI	31.2 ± 7.70	6.38 ± 5.03	51.1 ± 3.60	0.44 ± 0.06
	LNAE	0.11 ± 0.03	0.25 ± 0.12	0.53 ± 0.04	0.39 ± 0.08
	NAE	0.008 ± 0.00	0.01 ± 0.00	0.002 ± 0.00	0.88 ± 0.01
	HM	0.02 ± 0.00	0.04 ± 0.00	0.001 ± 0.00	0.86 ± 0.02
	LNTI	36.9 ± 9.77	14.3 ± 9.77	52.5 ± 3.45	0.50 ± 0.07
	HM	0.01 ± 0.00	0.03 ± 0.00	0.009 ± 0.00	0.85 ± 0.02
	LNAE	0.14 ± 0.05	0.32 ± 0.19	0.54 ± 0.04	0.45 ± 0.10
	HM	0.02 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.85 ± 0.02

Traits are grain yield (ton ha^{-1}) in low (GYLN) and normal (GYNN) nitrogen, N-agronomic efficiency (NAE, ton ton^{-1} N ha^{-1}), low-N tolerance index (LNTI, %), low-N agronomy efficiency (LNAE, ton ha^{-1}), and harmonic mean (HM, ton ha^{-1}). Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2).

[a]Data are mean ± SD estimated from fifty replications in independent validation.

3. MODELING DOMINANCE AND COPY NUMBER VARIATION CAN IMPROVE PREDICTION ACCURACY OF MAIZE HYBRIDS

ABSTRACT

In maize breeding, heterosis is the main factor to increase crop yield in single-cross hybrids and is a result of non-additive effects from the action of multiple loci. In addition, complementation of allelic variation and gene content also may be a valuable contributor, especially, in the cases of higher content of copy gains in inbred lines. Hence, we aimed (i) to compare the accuracy of additive, dominance, and additive-dominance genomic prediction models in two maize hybrids datasets, (ii) to evaluate prediction including the copy number variation (CNV) effects into GP models, and (iii) to compare the use of linear (NOIA) and nonlinear (GK) kernels in these analyses. In HELIX dataset, we observed a pronounced increase in accuracy between $NOIA_a$ (0.57) and $NOIA_{ad}$ (0.75), showing the contribution of dominance. For USP dataset, predicting with CNV model was the best approach regarding accuracy and bias for grain yield (GY), plant height (PH) and stay green (SG), outperforming even the additive-dominance models. Moreover, we observed a significant positive phenotypic correlation of small magnitude between PH and SG with copy number gain, showing a tendency of association between some copies with height and senescence of plants. The model GK_a demonstrated a superiority compared to $NOIA_a$ for GY in both datasets, confirming the performance in modeling complex non-additive effects. Our results suggest predicting hybrids including dominance effects led to slightly higher estimates of accuracy, mainly for complex traits. Furthermore, exploring information of copy variants in prediction models could lead to better estimates of genomic breeding values.

Keywords: Non-additive effects; Copy gain and loss; NOIA model

3.1. INTRODUCTION

In industry and public plant breeding programs, genomic prediction (GP) schemes are well established to increase genetic gains and reduce extensive field trials (Lado *et al*, 2017). In addition, it works reasonably well if applied routinely, in the same way as phenotypic selection (Bernardo, 2016; Lian *et al*, 2014). Recently, studies related to maize hybrid breeding have demonstrated that prediction accuracies can be improved by modeling genotype × environment interaction (Sousa *et al*, 2017) and non-additive effects (Santos *et al*, 2015). However, even having switched from a bandwagon to a consolidated methodology, GP still faces challenges in predicting phenotypes of highly polygenic traits due to the complex biological processes.

In the context of classical quantitative genetics regarding non-additive effects, several works have been published in animal (Miglior *et al*, 1995) and plant (Bernardo, 1996) science. Newly, with the advance of genomic technologies, estimation of variance components based on

SNPs became more accurate and precise (Da *et al*, 2014; Vitezica *et al*, 2013), even not reflecting the biological effect of the genes (Huang and Mackay, 2016). In genomic prediction, it is common to use additive models, but some studies report low accuracy for characters that have heterosis or relevant epistasis (Jiang and Reif, 2015). Thus, it would be useful to exploit non-additive effects in the GP model to increase the accuracy of selection. Furthermore, the inclusion models should also be a topic of interest for the application of GP in predicting heterotic transgressive phenotypes (Zhao et al, 2015). For instance, a simulated and real study in *Eucalyptus* breeding was reported by Denis and Bouvet (2013) and Bouvet *et al* (2016), respectively, where the authors showed that including dominance effects from a genomic realized relationship matrices (GRM) performed better for clone selection only when dominance effects were preponderant. Wang *et al* (2016), dos Santos *et al* (2016), and Resende *et al* (2017) included dominance effects in the GBLUP model in rice, maize, and *Eucalyptus* hybrids, respectively, and noticed a slight increase in accuracy.

Therefore, several prediction models adding dominance deviation have been proposed including parametric, non-parametric, and Bayesian approaches (dos Santos *et al*, 2016; Morota *et al*, 2014). For example, the natural and orthogonal interactions (NOIA) approach (Alvarez-Castro and Carlborg, 2007; Vitezica *et al*, 2017) removes the assumption of Hardy-Weinberg equilibrium, showing high performance for single-crosses prediction. Also, nonlinear Gaussian kernel (GK) is a viable alternative method to account for small complex non-additive effects without a direct modeling (Morota and Gianola, 2014). It is important to notice that we use molecular markers to measure relatedness and build GRM, regardless of parameterization. Consequently, other information could be used to account for different aspects of the genome, such as gene expression and metabolic abundance information (Guo *et al*, 2016; Riedelsheimer *et al*, 2012; Xu *et al*, 2017), or even imprinting effects (Jiang *et al*, 2017; Lopes *et al*, 2015). In this context, on maize genome, several factors may influence the prediction for quantitative traits (Rodgers-Melnick *et al*, 2016), leading to a biased estimation of genomic breeding value, such as *genomic imprinting*, *epigenetic regions*, *transposons*, and *copy number variations*.

Copy number variants (CNVs) are duplications/deletions of large DNA segments (>1000 base pair DNA segments) in comparison with a reference genome (Samelak-Czajka *et al*, 2017). In maize, structural genomic variation such as copy gain (duplication) and copy loss (deletion) may have different effects on the gene dosage and the phenotype, influencing directly in the gene expression (Springer *et al*, 2009). For instance, according to Swanson-Wagner *et al* (2010), over 10% of the ~32,500 genes surveyed in the maize panel exhibited CNVs relative to the B73 reference genome. Nowadays, with the high-throughput genomics techniques, it is possible to

identify CNVs from SNP data, and several algorithms are available (Abyzov et al, 2011; Mayrhofer et al, 2016). Even though the most common approach to detecting genomic intraspecific variation is SNP, CNV information could be used to capture larger variation in the genomic analysis (Manching et al, 2017). Thus, in breeding populations, CNVs can be tracked with high precision, increasing the accuracy of selection. For example, the *Rhg1* locus of soybean is a CNV of multiple genomic units, each containing four genes, which confers resistance to soybean cyst nematode (Lee *et al*, 2016).

However, at this point, no experimental data exist in maize breeding regarding the ability to predict the breeding values including copy number variation effects, or even testing non-additive effects in a real large maize population under contrasting N condition. Thus, our objectives were (i) to compare the accuracy of additive, dominance, and additive-dominance GP models in two maize hybrids datasets, (ii) to evaluate prediction including CNV effects into GP models, and (iii) to compare the use of linear (NOIA) and nonlinear (GK) kernels in GP analyses.

3.2. Materials and methods

3.2.1. Phenotypic data

USP dataset

We used 906 maize single-crosses from a full diallel mating design between 49 tropical inbred lines, contrasting for N-use efficiency (Mendonça *et al*, 2017). The experimental scheme used was augmented blocks (unreplicated trial) with two checks. The trials were carried out in Anhembi (22°50'51"S, 48°01'06"W, 466 m) and Piracicaba (22°42'23"S, 47°38'14"W, 535 m), at São Paulo State, Brazil, during the second growing season (January to May) of 2016 and 2017. In both sites, the hybrids were evaluated under two nitrogen (N) application levels, low (LN) with 30 kg N ha^{-1}, and ideal (IN) with 100 kg N ha^{-1}. Plots of seven meters (m) spaced 0.50 m were used under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, ton ha^{-1}), plant height (PH, cm), and stay green (SG). No stand correction was performed as the effect of genotype in plant number was verified by generalized linear modeling. Plots were manually harvested, and GY was corrected to 13% moisture. The PH was measured from soil surface to the flag leaf collar on five representative plants within each plot. SG was visually measured using a scale ranging from one (no senescence) to five (complete senescence).

We used ASReml-R (Butler *et al*, 2009) to perform a joint analysis to obtain best linear unbiased predictions (BLUPs) for the hybrids in each N condition, by fitting the following model:

$$y = X\beta + Vb + Hg + Si + \varepsilon$$

where y is a vector of phenotypic values of hybrids; β is a vector of fixed effects of environments (site and year), and checks; b is block within environment, where $b \sim N(0, I\sigma_b^2)$; g is genotypic values of hybrids, where $g \sim N(0, I\sigma_g^2)$; i is interaction environment × hybrids, where $i \sim N(0, I\sigma_{lg}^2)$; $\boldsymbol{\varepsilon}$ is a vector of random residuals from checks, where $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$. σ_ε^2 was jointly estimated based on *e* environment with *t* replicated check in each environment. *X*, *V*, *H*, and *S* are the incidence matrices for *β*, *b*, *g*, and *i*. Heterogeneous residual variance structure was assumed across environments.

In an attempt to predict genotypes under both optimal and N stress conditions, we used the selection index harmonic mean (HM) (Jafari *et al*, 2009) following the equation: $HM_i = \frac{2\times\left(GY_{(IN)_i}\times GY_{(LN)_i}\right)}{GY_{(IN)_i}+GY_{(LN)_i}}$, where HM_i is the harmonic mean of hybrid *i*; $GY_{(IN)_i}$ is the BLUP mean of grain yield in the IN condition of hybrid *i*; $GY_{(LN)_i}$is the BLUP for grain yield in the LN condition of hybrid *i*.

HELIX dataset

We used 452 maize single-crosses provided by Helix Sementes®, São Paulo, Brazil. The hybrids represent a partial diallel mating design between 128 tropical inbred lines. The experimental design used was a randomized complete block with two replications. Trials were carried out in Ipiaçu (18°40'51"S, 49°49'19"W, 443 m) and Patos de Minas (18°35'02"S, 46°28'10"W, 1067 m), in Minas Gerais State. Also, in Sertanópolis (23°02'39"S, 51°03'13"W, 390 m), in Paraná State, Nova Mutum (13°05'S, 56°05' W, 460 m), and Sorriso (12°32'S, 55°42'W, 365 m), in Mato Grosso State. These sites are in Southeastern, Southern, and Western of Brazil, respectively. The trials were conducted during the first growing season of 2014/15 from late September to early February. Two-row plots of 5 m spaced 0.70 m were used, and sowing density was about 63,000 kernels per hectare, under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, ton ha^{-1}), plant height (PH, cm), and ear height (EH, cm). The dataset presented a genotypic imbalance. Plots were mechanically harvested and adjusted to 13% moisture for GY assessment. PH/EH was measured from soil surface to the flag leaf collar on one representative plant within each plot (company criteria). No stand

correction was performed as the effect of genotype in plant number was verified by generalized linear modeling.

We used a linear mixed model in ASReml-R to predict the BLUPs for hybrids, considering the site as a fixed effect, and hybrid and the genotype by environment as random effects. Heterogeneous residual variance structure was assumed across sites. Entry-mean based heritability and phenotypic correlations among all trials are available on Sousa *et al* (2017).

3.2.2. Genotypic dataset

The genotyping of the 49 and 128 tropical inbred lines was performed using the Affymetrix Axiom® platform, containing about 614,000 SNPs (Unterseer *et al*, 2014). Markers with low call rate (<95%) and heterozygous loci on at least one individual were removed. In USP dataset, remaining missing data were imputed with Synbreed-R (Wimmer *et al*, 2012) and SNPs in LD with a pairwise r^2 value greater than 0.9 were removed using SNPRelate-R (Zheng *et al*, 2012). For HELIX dataset, imputation was done based on homozygosity of an individual and marker frequency with missed point using snpReady-R (Granato and Fritsche-Neto, 2017). High-quality polymorphic SNPs were used to build the artificial 906 and 452 hybrids genomic matrix, deduced by combining the genotypes from the parents. Afterwards, minor allele frequency (MAF) was conducted over hybrids markers considering the threshold of 0.05, resulting in a total of 34,571 and 52,700 SNPs, respectively. The frequency of heterozygous was estimated for individuals and markers using GAPIT-R (Lipka *et al*, 2012).

3.2.3. CNV calling

Raw Affymetrix (660K Axiom Maize) CEL files from 49 and 128 inbred lines were pre-processed in Axiom Analysis Suite Software, separately, to generate normalized signal intensity data and genotype calls. Afterwards, the $\log_2$ DNA copy number ratios and B-allele frequency (BAF) values for each sample were made using the reference files in the Axiom CNV Tool Software (Supplementary Figure S1). Copy number variation (gain and loss) was detected using Nexus Trial software v. 9.0 (Biodiscovery, El Segundo CA, USA) which uses a BAM (multiscale reference) method using a Hidden Markov Model (HMM) to segment the genome. We removed gains and losses smaller than 300 Kb, and used maize ZmB73 v.5 as a reference genome. We identified presence and absence of copy gain (CG) and copy loss (CL) in each chromosome for all samples in the two datasets.

We built the artificial 906 and 452 hybrids genomic matrix, deduced by combining the genotypes from its two parents, based on the codification of CNV as '1' for copy gain, '0' for no copy, and '-1' for copy loss. The final matrix was composed of a total of 321 and 283 CNVs, respectively. Furthermore, to visualize the genetic differences between inbred lines, a Neighbor-Joining Tree (NJT) was generated based on the Euclidean distance of the centered and standardized CNV matrix. We also calculated in ASReml-R the phenotypic (r_p) and genetic (r_g) correlation coefficients between traits and CG, CL, and gain and loss ratio (G/L), following the equation: $COV_{12}/\sqrt{\sigma_{a_1}\sigma_{a_2}}$, where COV_{12} is the additive genetic covariance; σ_{a_1} and σ_{a_2} are additive variances associated with each trait. Also, we measured the association between CG and CL in the inbred lines and the hybrids by linear regression.

We showed four ways to represent the formation of copy number variation in a single-cross hybrid (H_{12}) derived from inbred lines 1 (L_1) and 2 (L_2) (Figure 1). The first is a partial copy gain, showing L_1 with CG and L_2 only with the original/reference copy (Figure 1A). The second, a complete copy gain, with CG in the same region of the two inbred lines (Figure 1B). The third is a partial copy loss, with the line L_2 showing a copy loss (Figure 1C); and the last, showing a copy loss and gain, in different lines (Figure 1 D).

3.2.4. Prediction models

NOIA model

Additive-dominance NOIA (Vitezica *et al*, 2017) was used by fitting the following model:

$$g = Xb + Z_a a + Z_d d + e \quad (1)$$

where g is a vector of BLUP values of the n hybrids, b is a vector of fixed effects, a is a vector of additive genetic effects on the individuals, d is the vector of dominance effects, and e is a vector of random residuals. X and Z are the incidence matrices for b, a, and d. The distributions assumed were $a \sim N(0, \sigma_a^2 G_a)$, $d \sim N(0, \sigma_d^2 G_d)$, and $e \sim N(0, \sigma_e^2 I_n)$. G_a and G_d are the additive and dominance genomic relationship matrix (GRM), following the equation: $G_a = \frac{W_A W_A^T}{tr(W_A W_A^T)/m}$ and $G_d = \frac{W_D W_D^T}{tr(W_D W_D^T)/m}$, where m is the number of markers. The incidence matrices W_A and W_D were designed following:

$$W_A = \begin{cases} -(-p_{A_1A_2} - 2p_{A_2A_2}) \\ -(1 - p_{A_1A_2} - 2p_{A_2A_2}) \\ -(2 - p_{A_1A_2} - 2p_{A_2A_2}) \end{cases} \text{for genotypes} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

$$W_D = \begin{cases} -\frac{2p_{A_1A_2}p_{A_2A_2}}{p_{A_1A_1}+p_{A_2A_2}-(p_{A_1A_1}-p_{A_2A_2})^2} \\ \frac{4p_{A_1A_1}p_{A_2A_2}}{p_{A_1A_1}+p_{A_2A_2}-(p_{A_1A_1}-p_{A_2A_2})^2} \\ -\frac{2p_{A_1A_1}p_{A_1A_2}}{p_{A_1A_1}+p_{A_2A_2}-(p_{A_1A_1}-p_{A_2A_2})^2} \end{cases} \text{ for genotypes } \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

The W_A matrix was coded as 2 for homozygote A_1A_1, 1 for heterozygote $A_1A_2$and 0 for homozygote A_2A_2, for W_D was considered 0 for both homozygotes and 1 to the heterozygote. We named the models as $NOIA_a$ (additive), $NOIA_d$ (dominance), and $NOIA_{ad}$ (additive-dominance). For $NOIA_a$, we removed the dominance deviation, and for $NOIA_d$, we removed the additive effects in equation (1).

GK model

Additive-dominance GK was used by fitting the following model:

$$g = Xb + Z_a a + Z_d d + e \quad (2)$$

where g is a vector of BLUP values of the n hybrids, b is a vector of fixed effects, a is a vector of additive genetic effects on the individuals, d is the vector of dominance effects, and e is a vector of random residuals. X and Z are the incidence matrices for b, a, and d. The distributions assumed were $a \sim N(0, \sigma_a^2 K_a)$, $d \sim N(0, \sigma_d^2 K_d)$, and $e \sim N(0, \sigma_e^2 I_n)$. K_a and K_d are the additive and dominance symmetric semi positive definite matrix representing the covariance of the genetic values, following the equation: $K_a = exp\left(-hd_{a_i}^2/q_{0.05}\right)$ and $K_d = exp\left(-hd_{d_i}^2/q_{0.05}\right)$, where h is a bandwidth parameter, estimated from the Bayesian method (Cuevas *et al*, 2016); $d_{a_i}^2$ and $d_{d_i}^2$ are the squared Euclidean distance based on a centered and standardized additive and dominance incidence matrix (equal to equation 1), respectively, between individuals i; and $q_{0.05}$ is the fifth percentile of the same distance. We named the models as GK_a (additive), GK_d (dominance), and GK_{ad} (additive-dominance). However, they do not model additivity and dominance directly, capturing other effects at the same time (Morota *et al*, 2014). For GK_a we removed the dominance deviation, and for GK_d, we removed the additive effects in equation (2).

CNV model

We used CNV effects for predicting phenotypes by fitting the following model:

$$g = Xb + \mathrm{Z}a + e \quad (3)$$

Where g, b, a, and e are same as those defined in the model (1). The distributions assumed were $a \sim N(0, \sigma_c^2 G_c)$ where G_c is the relationship matrix estimated from CNV, and σ_c^2 the variance due to CNV abundances. The GRM G_c is estimated using $G_c = \frac{WW^T}{c}$, where W is a $n \times c$ matrix of a scaled and centered CNV (1 for copy gain, 0 for no copy, and -1 for copy loss) from *n* single-cross hybrid and *c* is the total number of CNV.

3.2.5. Validation and model comparison

From GP models, we evaluated prediction accuracy (r_{MP}), the correlation between BLUP and genomic estimated breeding values, from fifty replications, randomly sampling 75% of the hybrids to form the training set (TS) and the rest of validation set (VS). All prediction analyses were determined using Bayesian Generalized Linear Regression (BGLR) (Perez and de los Campos, 2014). We reported the posterior mean estimates and standard deviations of the SNP additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_e^2), narrow sense genomic heritability ($h_a^2 = \sigma_a^2/\sigma_a^2 + \sigma_e^2$), the proportion of the total phenotypic variance explained by genomic dominance ($d^2 = \sigma_d^2/\sigma_d^2 + \sigma_e^2$), broad sense genomic heritability ($h_{ad}^2 = \sigma_a^2 + \sigma_d^2/\sigma_a^2 + \sigma_d^2 + \sigma_e^2$), and the deviance information criterion ($DIC = D(\bar{\theta}) + 2pD$), where $D(\bar{\theta})$ is the deviance at the posterior mean of the model, and pD is the effective number of parameters. The model with the lowest DIC value presents the best data fit. We used a total of 60,000 MCMC iterations, 15,000 for burn-in, and 5 for thinning. We plotted the genomic relationship matrix of the NOIA, GK, and CNV by a heatmap graph.

3.3. RESULTS

Copy variation

Cluster (NJ tree) analysis revealed levels of diversity for the 49 (USP) and 128 (HELIX) inbred lines (Figure 2), showing ten and nine groups in the cluster, respectively. We found no association ($R^2 < 0.1$) between copy gains and losses in both datasets. Regarding the number of CNV, in the USP dataset we identified 11 CG and 14 CL in the lines, and 20 and 26 in the hybrids for the former and the latter, respectively (Figure 3 A and B). On the other hand, we observed in the HELIX dataset, 6 CG and 13 CL in the inbred lines, and 8.9 CG and 22.9 CL in the single-crosses (Figure 3 C and D).

Despite the low magnitude, the USP dataset in both N conditions showed significant positive r_p between PH and SG with CG (Table 1). On the other hand, for GY, HM, and PH significant negative r_p of low magnitude with CL were observed. Concerning the HELIX dataset, significant positive r_p were found only between GY with CL and GL. Considering the genetic correlations, we did not find any statistical significance between any trait and CNV for either dataset.

Predicting single-crosses using dominance deviation and CNV

Regarding the USP dataset, the hybrids average performance for GYLN was 5.36, ranging from 2.39 to 7.61 ton ha^{-1}. For GYIN the mean was 5.94, varying from 2.35 to 9.17 ton ha^{-1}, and for HM it was 5.62±0.74 SD (ton ha^{-1}) (Supplementary Figure S2A). The mean for PHLN was 194, varying from 147 to 225 cm, and for PHIN it was 200 from 147.3 to 233.7. In its turn, the mean for SGLN was 3.63±0.28 SD and 3.72±0.43 SD for SGIN (Supplementary Figure S2B). Considering the HELIX dataset, we noticed higher BLUP mean values compared to USP dataset, which were 7.20±0.90 SD for grain yield, 240±15.7 SD plant height, and 128±8.6 SD for ear height. On the other hand, considering the frequency of heterozygosity, USP dataset showed higher values for individuals and markers compared to HELIX (Supplementary Figure S3). Heatmaps graphs revealed several groups according to the GRM, showing more adjustment for GK matrix in both datasets (Supplementary Figure S4; Supplementary Figure S5).

Estimates of variance components and genomic heritability varied considerably among traits and models in USP dataset (Table 2 and 3). It is important to highlight, for both models and N conditions, CNV showed lower DIC values for PH and SG (Table 2). Comparing the additive models, GK was the best relationship kernel for both N conditions and all traits. However, considering the additive-dominance model, NOIA was slightly better than GK for all characters, except for PHLN (Table 2). The worst DIC values was observed for the dominance model. Furthermore, estimates of h_a^2 were higher in additive models compared to d^2, for all traits (Table 3). For GY, PH, and SG in low N, NOIA model h_{ad}^2 reached 0.48, 0.62, and 0.47, respectively. However, considering GK for the same traits the h_{ad}^2 was 0.68, 0.81, and 0.75, respectively. Under ideal N condition, NOIA model h_{ad}^2 reached 0.56, 0.69, and 0.59. and GK h_{ad}^2 were 0.80, 0.86, and 0.81, respectively. The highest heritability estimate for HM was h_{ad}^2=0.72 using GK model. Regarding the prediction accuracy (r_{MP}), the highest values obtained were 0.53 and 0.60 for additive-dominance GK and NOIA for GYLN and GYIN, respectively (Figure 4A). For plant height, the amount of r_{MP} remained the same for additive and additive-dominance NOIA and GK for PHLN (0.70) and PHIN (0.73) (Figure 4B). On the other hand,

r_{MP} for additive models performed slightly better for SG in both N conditions (Figure 4C), reaching values of 0.62 ($NOIA_a$-SGLN) and 0.71 ($NOIA_a$/GK_a-SGIN). Considering the selection index HM, r_{MP} was higher for $NOIA_{ad}$ compared to $NOIA_a$ (Figure 4D). Surprisingly, r_{MP} for dominance $NOIA_d$ was zero for all the traits tested. Nevertheless, the values for GK_d was slightly lower than the other two effects. We highlight the performance the CNV-based model, which outperformed the additive-dominance models for PH and SG in both N condition.

Considering the HELIX dataset, estimates of variance components varied slightly among traits and models (Table 4). The lowest DIC observed for GY was using $NOIA_{ad}$ (489), and for PH and EH was considering GK_{ad} (2378 and 1990, respectively). Differently from USP dataset, CNV presented higher DIC values for all traits. On the other hand, the worst DIC values were obtained when the dominance models were used. Concerning the additive models, the GK kernel outperformed for all traits. The genomic heritability for GY in NOIA model was 0.24 (h_a^2), 0.63 (d^2), and 0.87 (h_{ad}^2); and for GK was 0.10, 0.85, and 0.95 (Figure 5). In $NOIA_d$ and $NOIA_{ad}$ models the proportion of the total phenotypic variance explained by genomic dominance (d^2) was higher in GY compared to PH and EH. In addition, the values of d^2 using GK_d was relatively higher than $NOIA_d$. Regarding the r_{MP}, the highest value was detected for PH, followed by EH and GY. Furthermore, the additive-dominance NOIA and GK were the best models for all traits. Surprisingly, CNV model was one of the worst, for all characters (Figure 6).

3.4. DISCUSSION

In our work, phenotypic values between low and ideal N varied considerably for GY, and slightly for PH, and SG (Figure S2), which could be related to physiological issues of N stress (Antonietta *et al*, 2016; Sade *et al*, 2017). In addition, grain yield is negatively associated with stay green under low (r_p = -0.27, P < 0.05) and ideal (r_p = -0.40, P < 0.05) N conditions (results not shown), implying that a longer period of photosynthetic activity increases production. However, an optimal situation is to find genotypes with higher production and precocity (lower canopy senescence). According to Antonietta *et al* (2016) maize genotypes could present higher post-silking N uptake, increasing productivity, and SG is not always linked to higher post-silking N accumulation.

We found a high frequency of copy number variants at both datasets of maize inbred lines (Figures 2 and 3), reflecting in a significant contribution to the intra-species genetic variation. In this case, some transposon elements probably contributed to the induction of tandem sequence duplications, which is highly associated to copy gains (Dong *et al*, 2016; Liu *et*

al, 2017; Zhang *et al*, 2013), thus, contributing to the diversity of the material studied. In addition, these lines came from a breeding panel of N stress (USP dataset) and a company based program (HELIX dataset), which passed through a long selection process to become adapted to tropical conditions. We also noticed high mean levels of CG in USP dataset (Figure 3), and this could be related to the selection in extreme abiotic stress conditions, which is in agreement with Dassanayake and Larkin (2017) where gene duplication is one of the possible strategies to handle stress. We showed some cases of CNV in hybrids (Figure 1), but biologically it may be more complex, due to the recombination, some genes deleted or duplicated contributing to the phenotype, size and dosage effects of the copy, and gene action (Zmienko *et al*, 2014). In maize, recent studies indicate that unique small RNAs within the duplicated segments exhibit dosage secondary transcript levels, proving the importance of dosage to the phenotypic variation (Zuo *et al*, 2016).

Based on single-crosses of the USP dataset, we found a significant positive phenotypic correlation between PH and SG with copy number gain, and even of low magnitude, showed a tendency that more CG are related to height and senescence of plants (Table 1). On the other hand, we found a significant negative correlation between CL and GY/PH (USP) and GY (HELIX). It means that lower and less productive plants are associated with high copy losses, suggesting that the deleted DNA segments influenced substantially the phenotype, which can be further studied seeking for genes and regulatory regions. These results show us that the association between total copy number variation and complex traits is still a complicated measurement, mainly, due to the lack of robust method capable of identifying the full range of structural DNA variation. Nevertheless, Chia *et al* (2012) suggest a correlation between genomic regions containing structural variation and QTLs for leaf architecture and resistance to northern and southern leaf blight in a diverse maize panel. In wheat, Wurschum *et al* (2015) showed that a specific copy number has a substantial effect on the fine-tuning of flowering time, empathizing the importance to adaptation.

We did not find any correlation between copy gain and grain yield in either datasets (Table 1). This result was not expected since more duplicated copies could be associated with higher values of production or height, as they are highly polygenic (Belo *et al*, 2010). We agree that this comparison may not be so straightforward, being influenced by many factors such as CNV estimation, or artificially formation of hybrids based on copy number. In the scientific community, a strong hypothesis shows that heterosis could be highly influenced by individuals with high levels of gene duplication. Several studies suggested this evidence (Belo *et al*, 2010; Schnable and Springer, 2013; Springer *et al*, 2009; Swanson-Wagner *et al*, 2010), showing

complementation of allelic variation and variation in gene content and expression patterns, were significant contributors to heterosis.

In the USP dataset we found higher dominance variances and genomic heritabilities (h_a^2, d^2, h_{ad}^2) for all traits in ideal N compared to low N (Table 3). Regarding prediction accuracy, we also detected considerable differences for GY, PH, and SG between both conditions. According to Crossa *et al* (2010) and Ziyomo and Bernardo (2013) unfavorable conditions could affect genomic heritability, and consequently, interfere in the estimates of prediction accuracy. Therefore, selection indices that take into account the performance under both optimal and stress conditions could be an alternative of prediction. In our study, the accuracy of the harmonic mean outperformed the GY in both conditions (Figure 4), which is in agreement with Lyra *et al* (2017) who compared four nitrogen selection indices and found higher r_{MP} for HM. Moreover, we did not find any significant difference between $NOIA_{ad}$ and GK_{ad} related to prediction accuracy for HM (Figure 4). However, we highlighted a slight increase for GY in low and ideal N for $NOIA_a$ (0.50, 0.56) and $NOIA_{ad}$ (0.53, 0.60), GK_a (0.52, 0.59) and GK_{ad} (0.53, 0.60). The r_{MP} was almost the same in additive and additive-dominance models for plant height. On the other hand, additive models showed better performance for stay green. Thus, traits controlled by additive gene action, such as SG (Abdelrahman *et al*, 2017), may not increase accuracy including dominance effects.

Considering the HELIX dataset, we also found higher dominance variances and broad sense genomic heritability (h_{ad}^2) for GY, PH and EH (Figure 5). We identified a considerable increase of r_{MP} for GY comparing $NOIA_{ad}$ (0.75) to $NOIA_a$ (0.57), and GK_{ad} (0.74) to GK_a (0.70) (Figure 6), suggesting the contribution of dominance to boost accuracy. According to Almeida Filho *et al* (2016), additive-dominance models may be improved considerably for traits with large dominance effects. We did not include epistasis, however, several studies of complex traits reported the benefits of including both non-additive effects (Bouvet *et al*, 2016; Munoz *et al*, 2014; Vieira *et al*, 2017). Moreover, high levels of heterozygosity were found in the hybrids and markers (Figure S2), which could contribute to increase r_{MP} in dominance models. It is important to highlight that r_{MP} using GK_a relative to $NOIA_a$ increased significantly, proving the performance of the method to account for marker-specific interaction effects, and complex marker main effects (Cuevas *et al*, 2017; Cuevas *et al*, 2016). Similar results were described in rice (Liu *et al*, 2016) and maize (Sousa *et al*, 2017), where the GK kernel outperformed GBLUP. We also observed that the r_{MP} of $NOIA_d$ for GY was considerable higher in comparison to USP, showing that dominance must be modeled along with additivity. Similarly, Wolfe *et al* (2016) reported lower performance for dominance models when compared to additive in *Cassava* breeding, depending on the trait.

We found values greater than 0.90 for d^2 in all traits and datasets for GK_d models (Table 3, Figure 5), showing higher values compared to $NOIA_d$. Hence, all of the variation wasexplained by dominance, which we suspect that the Gaussian GRM is inflating the estimation of heritability. In contrast, it has been suggested that estimates of h_a^2 could be inflated in the presence of non-additive variation (Zuk *et al*, 2012). Furthermore, we observed NOIA models incorporating dominance in grain yield are more accurate when d^2 is high. For example, for GY_{LN} in $NOIA_d$, d^2 was 0.15 and accuracy was 0.50 and 0.53 to A and A-D. However, for GY in HELIX, d^2 was 0.72 and accuracy was 0.57 and 0.75 to A and A-D. Similar results was reported by Almeida Filho *et al* (2016), which showed an increase of accuracy in additive-dominance models when d^2 was equal to 0.2 for oligogenic and polygenic traits.

Regarding the prediction accuracy, in the USP dataset, we noticed a higher or at least the same value for all traits using copy number variation as a kernel matrix compared to additive, dominance, and additive-dominance models (Figure 4). The performance of the method could be related to the real influence of copy number to the phenotypes, and in differentiating the hybrids concerning structural variation. However, for HELIX dataset, the CNV kernel was the worst for all traits (Figure 6). In this case, the copy number variation, possibly, is not directly affecting causative genes as other effects may do. Although, the prediction was considered high for the traits evaluated, showing that this kind of information could be used in breeding programs. In contrast, the drawback of CNV is the time required to analyze the data, since it did not show any outstanding values of accuracy compared to SNPs markers. In the literature, there is still a lack of information using copy number inside prediction or association models. A similar case was studied by Wurschum *et al* (2017), who worked with wheat genotypes including three KASP markers targeting *Fr-A2* (CNV) as fixed effects in genomic prediction. The authors showed a slight increase in r_{MP} compared to traditional RR-BLUP, and demonstrated a significant contribution of this structural variation to the winter hardiness in wheat. In animal breeding, imprinting information was included in the additive-dominance model in GP, showing significant difference relative to additive model (Jiang *et al*, 2017). Thus, further investigation is necessary to identify the CNVs with high effect on complex traits and combine this information inside prediction models.

3.5. CONCLUSION

Our results suggest that the best approach is predicting single-crosses including dominance effects, mainly for complex traits. Furthermore, including copy number variation

effects seems to be suitable, due to the increase of prediction accuracies and reduction of model bias. Also, the overall performance of the nonlinear gaussian kernel model was superior relative to the natural and orthogonal interactions model.

REFERENCES

Abdelrahman M, El-Sayed M, Jogaiah S, Burritt DJ, Tran LP (2017). The "STAY-GREEN" trait and phytohormone signaling networks in plants under heat stress. *Plant cell reports* **36**(7)**:** 1009-1025.

Abyzov A, Urban AE, Snyder M, Gerstein M (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* **21**(6)**:** 974-984.

Almeida Filho JE, Guimaraes JF, FF ES, de Resende MD, Munoz P, Kirst M *et al* (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb)* **117**(1)**:** 33-41.

Alvarez-Castro JM, Carlborg O (2007). A unified model for functional and statistical epistasis and its application in quantitative trait Loci analysis. *Genetics* **176**(2)**:** 1151-1167.

Antonietta M, Acciaresi HA, Guiamet JJ (2016). Responses to N Deficiency in Stay Green and Non-Stay Green Argentinean Hybrids of Maize. *J Agron Crop Sci* **202**(3)**:** 231-242.

Belo A, Beatty MK, Hondred D, Fengler KA, Li BL, Rafalski A (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor Appl Genet* **120**(2)**:** 355-367.

Bernardo R (1996). Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. *Theor Appl Genet* **93**(7)**:** 1098-1102.

Bernardo R (2016). Bandwagons I, too, have known. *Theor Appl Genet* **129**(12)**:** 2323-2332.

Bouvet JM, Makouanzi G, Cros D, Vigneron P (2016). Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. *Heredity (Edinb)* **116**(2)**:** 146-157.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009). ASReml-R reference manual. *Department of Primary Industries, Queensland, Australia.*

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J *et al* (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* **44**(7)**:** 803-U238.

Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL *et al* (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**(2)**:** 713-724.

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P, de los Campos G (2017). Bayesian genomic prediction with genotype x environment interaction kernel models. *G3-Genes Genom Genet* **7:** 41-53.

Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, de los Campos G *et al* (2016). Genomic prediction of genotype x environment interaction kernel regression models. *The plant genome* **9**(3)**:** 1-20.

Da Y, Wang CK, Wang SW, Hu G (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *Plos One* **9**(1)**:** e87666.

Dassanayake M, Larkin JC (2017). Making Plants Break a Sweat: the Structure, Function, and Evolution of Plant Salt Glands. *Front Plant Sci* **8:** 406.

Denis M, Bouvet JM (2013). Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. *Tree Genet Genomes* **9**(1)**:** 37-51.

Dong J, Feng Y, Kumar D, Zhang W, Zhu T, Luo MC *et al* (2016). Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc Natl Acad Sci U S A* **113**(29)**:** 7949-7956.

dos Santos JPR, Vasconcellos RCD, Pires LPM, Balestre M, Von Pinho RG (2016). Inclusion of dominance effects in the multivariate GBLUP model. *Plos One* **11**(4)**:** 1-21.

Granato ISC, Fritsche-Neto R (2017). snpReady: Preparing genotypic datasets in order to run genomic analysis. *R Package - https://cranr-projectorg/web/packages/snpReady/indexhtml.*

Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* **129**(12)**:** 2413-2427.

Huang W, Mackay TFC (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *Plos Genet* **12**(11).

Jafari A, Paknejad F, Al-Ahmadi MJ (2009). Evaluation of selection indices for drought tolerance of corn (*Zea mays* L.) hybrids. *Int J Plant Prod* **3**(4)**:** 33-38.

Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, Ma L (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *Bmc Genomics* **18**(1)**:** 425.

Jiang Y, Reif JC (2015). Modeling epistasis in genomic selection. *Genetics* **201**(2)**:** 759-768.

Lado B, Battenfield S, Guzman C, Quincke M, Singh RP, Dreisigacker S *et al* (2017). Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. *Plant Genome* **10**(2).

Lee TG, Diers BW, Hudson ME (2016). An efficient method for measuring copy number variation applied to improvement of nematode resistance in soybean. *The Plant journal : for cell and molecular biology* **88**(1)**:** 143-153.

Lian L, Jacobson A, Zhong S, Bernardo R (2014). Genomewide prediction accuracy within 969 maize biparental populations. *Crop Sci* **54**(4)**:** 1514-1522.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ *et al* (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**(18)**:** 2397-2399.

Liu S, Zheng J, Migeon P, Ren J, Hu Y, He C *et al* (2017). Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. *Scientific reports* **7:** 42444.

Liu ZY, Zhu CS, Jiang Y, Tian YL, Yu J, An HZ *et al* (2016). Association mapping and genetic dissection of nitrogen use efficiency-related traits in rice (*Oryza sativa* L.). *Funct Integr Genomic* **16**(3)**:** 323-333.

Lopes MS, Bastiaansen JWM, Janss L, Knol EF, Bovenhuis H (2015). Estimation of Additive, Dominance, and Imprinting Genetic Variance Using Genomic Data. *G3-Genes Genom Genet* **5**(12)**:** 2629-2637.

Lyra DH, Mendonça LF, Galli G, Alves FC, Granato ISC, Fritsche-Neto R (2017). Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Mol Breeding* **37**(80)**:** 1-14.

Manching H, Sengupta S, Hopper KR, Polson SW, Ji Y, Wisser RJ (2017). Phased Genotyping-by-Sequencing Enhances Analysis of Genetic Diversity and Reveals Divergent Copy Number Variants in Maize. *G3* **7**(7)**:** 2161-2170.

Mayrhofer M, Viklund B, Isaksson A (2016). Rawcopy: Improved copy number analysis with Affymetrix arrays. *Scientific reports* **6**.

Mendonça LF, Granato ISC, Alves FC, Morais PPP, Vidotti MS, Fritsche-Neto R (2017). Accuracy and simultaneous selection gains for N-stress tolerance and N-use efficiency in maize tropical lines. *Scientia Agricola* **74**(6)**:** 481-488.

Miglior F, Burnside EB, Dekkers JC (1995). Nonadditive genetic effects and inbreeding depression for somatic cell counts of Holstein cattle. *J Dairy Sci* **78**(5)**:** 1168-1173.

Morota G, Boddhireddy P, Vukasinovic N, Gianola D, DeNise S (2014). Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Front Genet* **5:** 56.

Morota G, Gianola D (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front Genet* **5:** 363.

Munoz PR, Resende MF, Jr., Gezan SA, Resende MD, de Los Campos G, Kirst M *et al* (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* **198**(4)**:** 1759-1768.

Perez P, de los Campos G (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**(2)**:** 483-U463.

Resende R, Resende M, Silva F, Azevedo C, Takahashi E, Silva-Junior O *et al* (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity***:** 1–11.

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

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016). Open chromatin reveals the functional maize genome. *Proc Natl Acad Sci U S A* **113**(22)**:** E3177-3184.

Sade N, del Mar Rubio-Wilhelmi M, Umnajkitikorn K, Blumwald E (2017). Stress-induced senescence and plant tolerance to abiotic stress. *Journal of experimental botany***:** 1-9.

Samelak-Czajka A, Marszalek-Zenczak M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M, Zmienko A (2017). MLPA-Based Analysis of Copy Number Variation in Plant Populations. *Front Plant Sci* **8:** 222.

Santos JP, Pereira HD, Von Pinho RG, Pires LP, Camargos RB, Balestre M (2015). Genome-wide prediction of maize single-cross performance, considering non-additive genetic effects. *Genet Mol Res* **14**(4)**:** 18471-18484.

Schnable PS, Springer NM (2013). Progress toward understanding heterosis in crop plants. *Annual review of plant biology* **64:** 71-88.

Sousa MBE, Cuevas J, Couto EGD, Perez-Rodriguez P, Jarquin D, Fritsche-Neto R *et al* (2017). Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype x Environment Interaction. *G3-Genes Genom Genet* **7**(6)**:** 1995-2014.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y *et al* (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *Plos Genet* **5**(11)**:** e1000734.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D *et al* (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome research* **20**(12)**:** 1689-1699.

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M *et al* (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* **15**(823)**:** 1-15.

Vieira IC, Dos Santos JPR, Pires LPM, Lima BM, Goncalves FMA, Balestre M (2017). Assessing non-additive effects in GBLUP model. *Genet Mol Res* **16**(2).

Vitezica ZG, Legarra A, Toro MA, Varona L (2017). Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. *Genetics* **206**(3)**:** 1297-1307.

Vitezica ZG, Varona L, Legarra A (2013). On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. *Genetics* **195**(4)**:** 1223-+.

Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C *et al* (2016). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity* **118:** 302-310.

Wimmer V, Albrecht T, Auinger HJ, Schon CC (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* **28**(15)**:** 2086-2087.

Wolfe MD, Kulakow P, Rabbi IY, Jannink JL (2016). Marker-Based Estimates Reveal Significant Non-additive Effects in Clonally Propagated Cassava (Manihot esculenta): Implications for the Prediction of Total Genetic Value and the Selection of Varieties. *G3*.

Wurschum T, Boeven PH, Langer SM, Longin CF, Leiser WL (2015). Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *Bmc Genet* **16:** 96.

Wurschum T, Longin CF, Hahn V, Tucker MR, Leiser WL (2017). Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat. *The Plant journal : for cell and molecular biology* **89**(4)**:** 764-773.

Xu Y, Xu C, Xu S (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity***:** 1–11.

Zhang J, Zuo T, Peterson T (2013). Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. *Plos Genet* **9**(8)**:** e1003691.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**(24)**:** 3326-3328.

Ziyomo C, Bernardo R (2013). Drought tolerance in maize: indirect selection through secondary traits versus genomewide selection. *Crop Sci* **53**(4)**:** 1269-1275.

Zmienko A, Samelak A, Kozlowski P, Figlerowicz M (2014). Copy number polymorphism in plant genomes. *Theor Appl Genet* **127**(1)**:** 1-18.

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**(4)**:** 1193-1198.

Zuo T, Zhang J, Lithio A, Dash S, Weber DF, Wise R *et al* (2016). Genes and Small RNA Transcripts Exhibit Dosage-Dependent Expression Pattern in Maize Copy-Number Alterations. *Genetics* **203**(3)**:** 1133-1147.

FIGURES

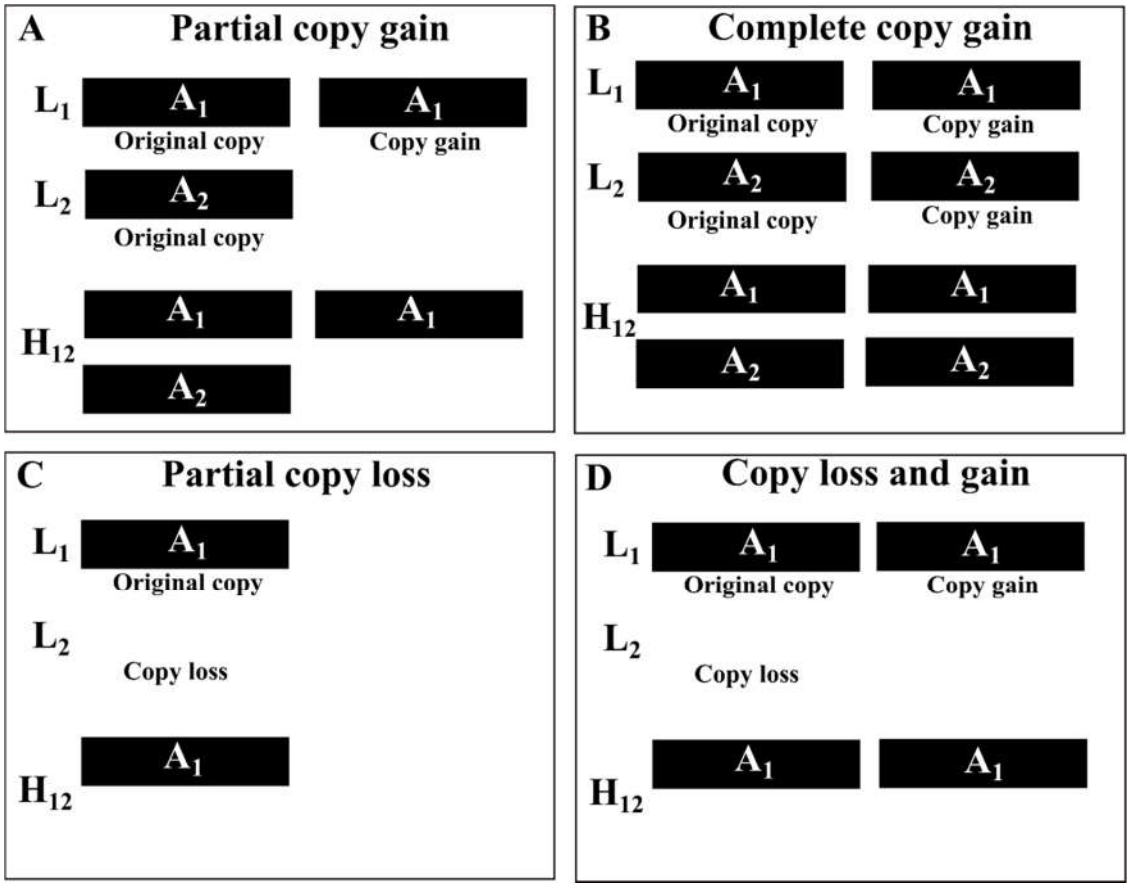


Figure 1. Schematic representation of copy number variation considering one gene with two alleles (A1/A2) in complete dominance, for two inbred lines (L1/L2) and one derived hybrid (H12). (A) Partial copy gain, (B) complete copy gain, (C) partial copy loss, and (D) copy loss/gain.

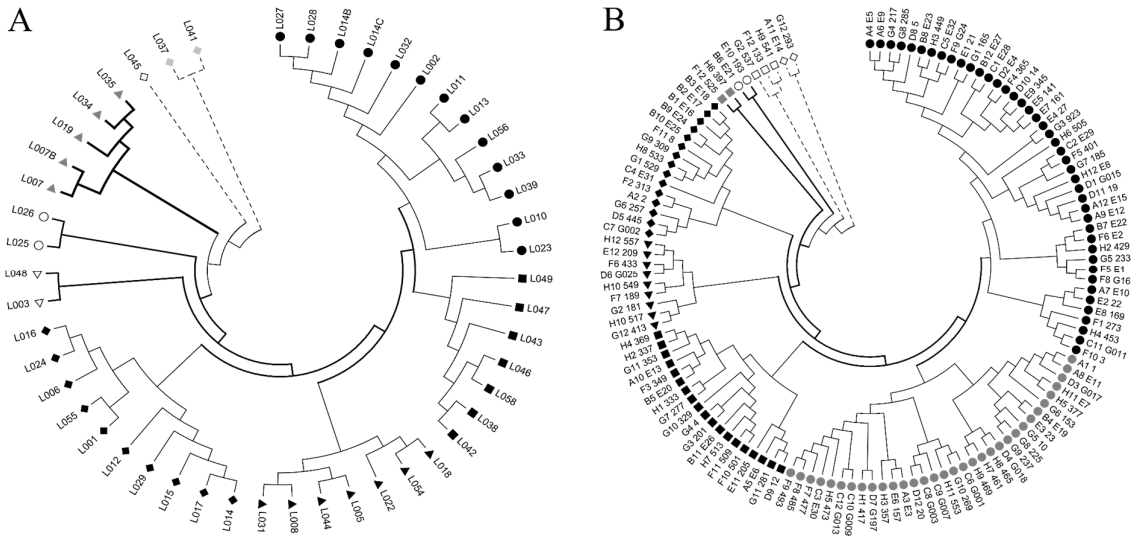


Figure 2. Circular Neighbor-Joining Tree. Euclidean' distance from copy number variation in (A) 49 and (B) 128 inbred maize lines.

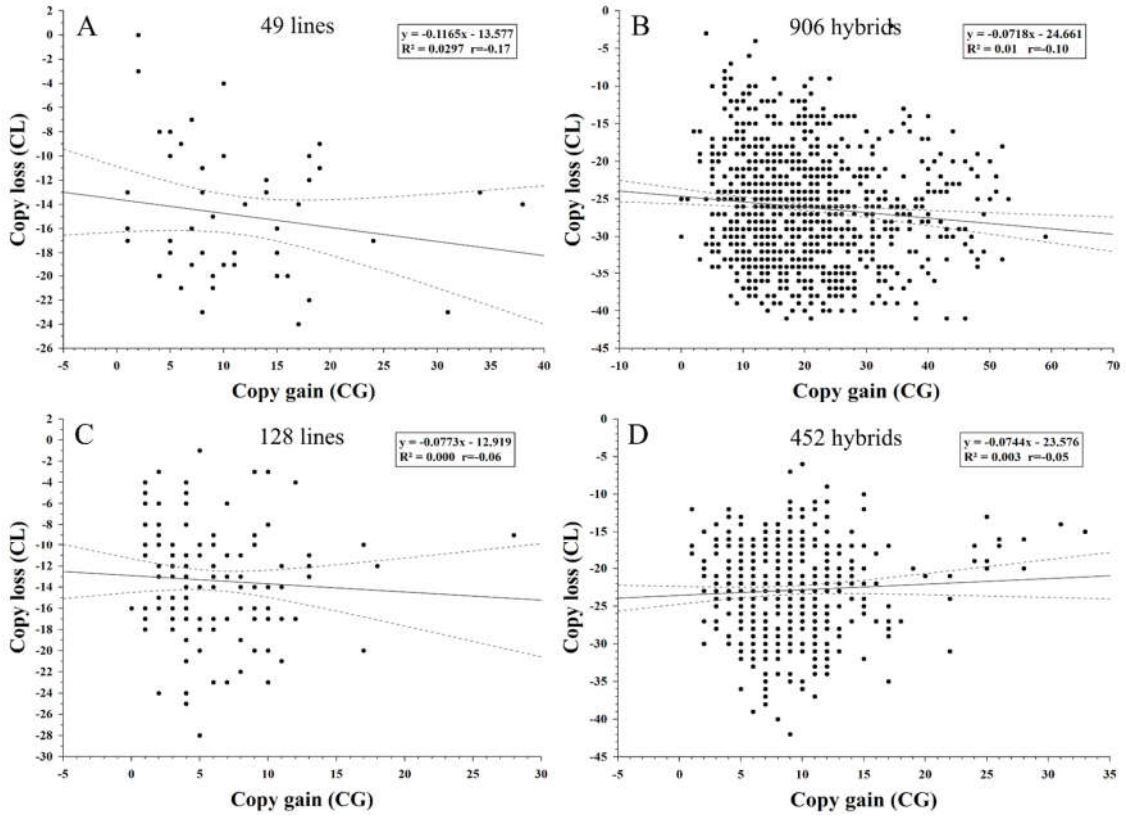


Figure 3. Scatterplot of copy number gain vs. copy number loss. USP dataset consisting of (A) 49 inbred lines and (B) 906 hybrids. HELIX dataset composed of (C) 128 inbred lines and (D) 452 single-crosses. The gray line is the regression slope and 95% confidence interval (light gray dotted line).

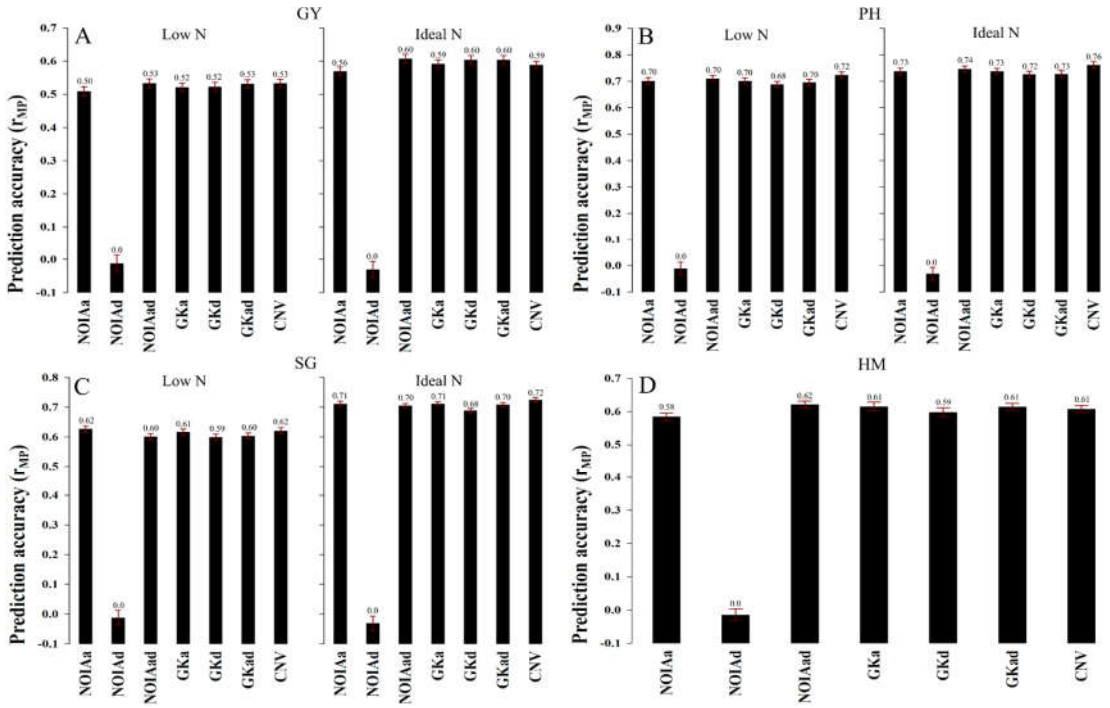


Figure 4. Barplot of prediction accuracy from 906 maize single-crosses. Traits evaluated in low (LN) and ideal (IN) nitrogen are (A) grain yield (GY, ton ha^{-1}), (B) plant height (PH, cm), (C) stay green (SG), and (D) harmonic mean (HM, ton ha^{-1}). Black numbers above bars represent the mean± 95% confidence intervals estimated from fifty replications in independent validation. Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV.

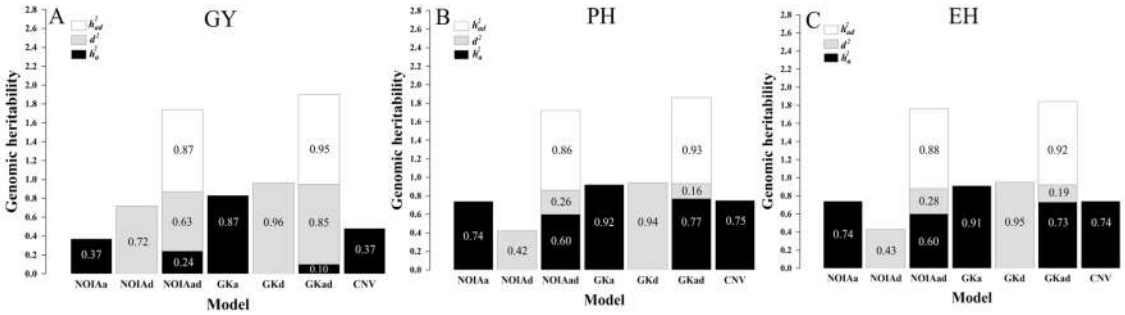


Figure 5. Estimate of genomic heritability from 452 maize single-crosses. Traits are (A) grain yield (GY, ton ha^{-1}), (B) plant height (PH, cm), and (C) ear height (PH, cm). Narrow sense genomic heritability (h_a^2), proportion of the total phenotypic variance explained by dominance (d^2), and broad sense genomic heritability (h_{ad}^2). Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV. Numbers inside plot represent the mean estimated from fifty replications in independent validation.

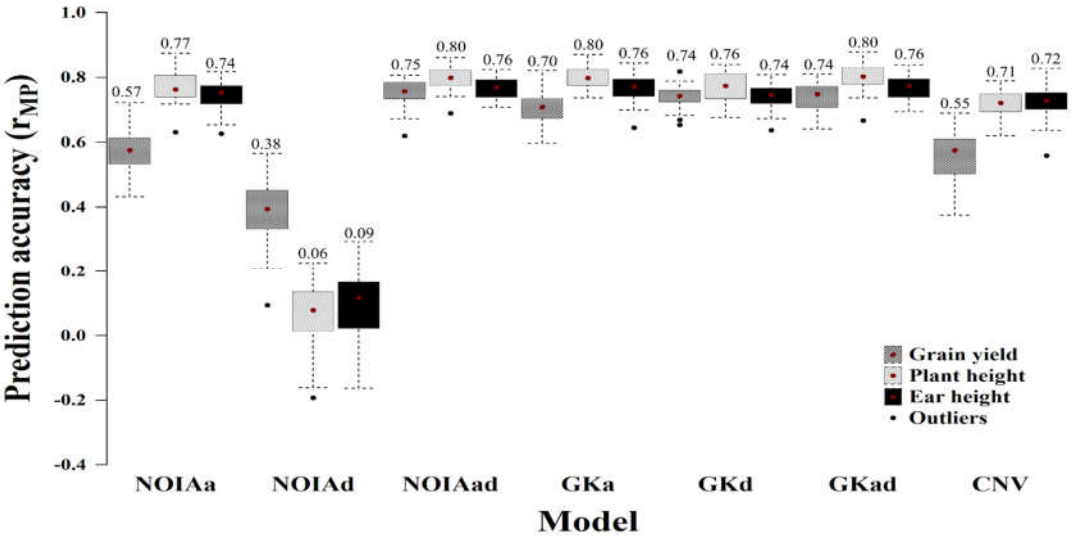


Figure 6. Boxplot of prediction accuracy from 452 maize single-crosses. Traits are grain yield (GY, ton ha^{-1}), plant height (PH, cm), and ear height (PH, cm). Red dots and black numbers represent the mean estimated from fifty replications in independent validation. Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV.

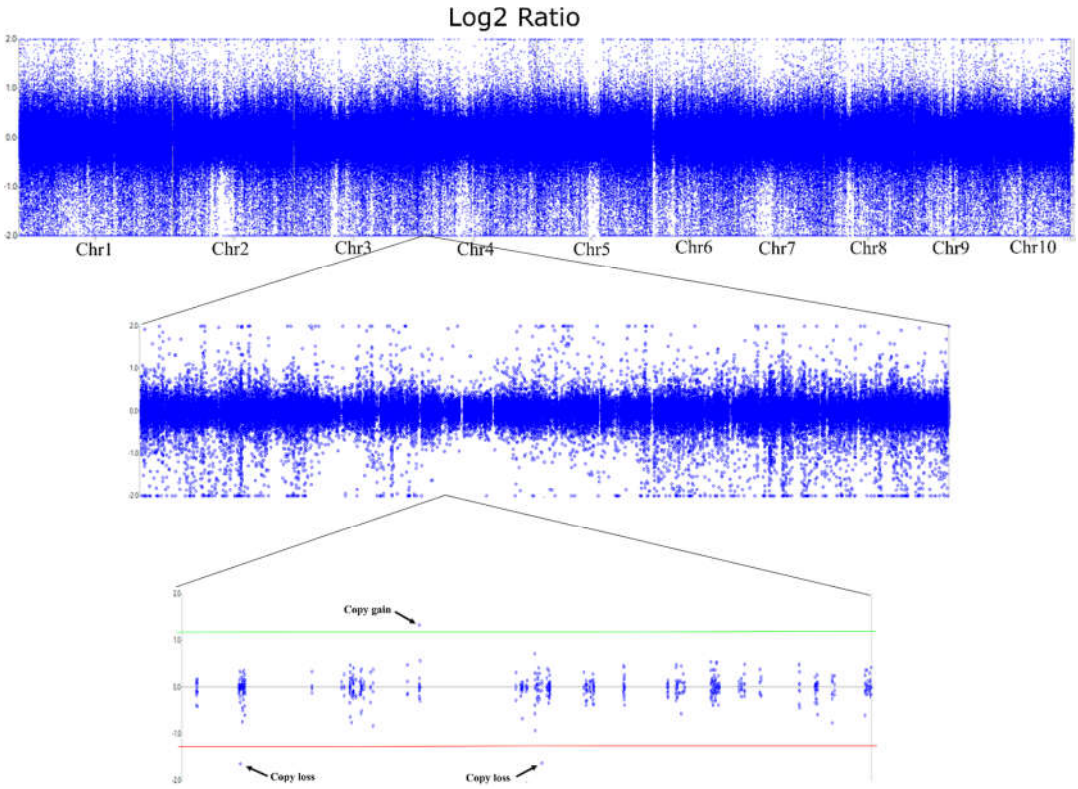


Figure S1. Example of copy number estimation based on the log2 ratios from Axiom CNV Tool Software. USP dataset (49 inbred lines) was used for the analysis. Each probe is represented as a small blue dot along the length of the ten chromosomes. Two horizontal lines (thresholds) was determined for copy gain (green) and copy loss (red).

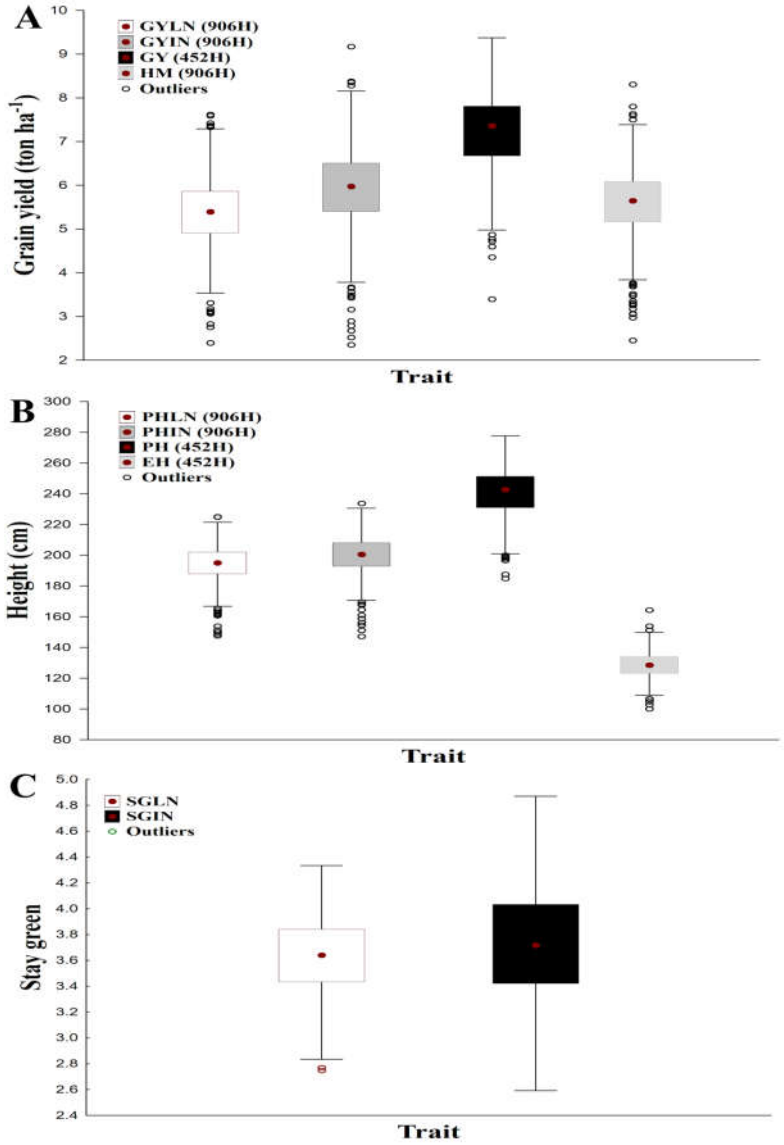


Figure S2. Boxplot of phenotypic traits. (A) Grain yield (GY, ton ha^{-1}) and harmonic mean (HM, ton ha^{-1}), (B) plant and ear height (PH/EH, cm), and (C) stay green (SG) for USP and HELIX datasets.

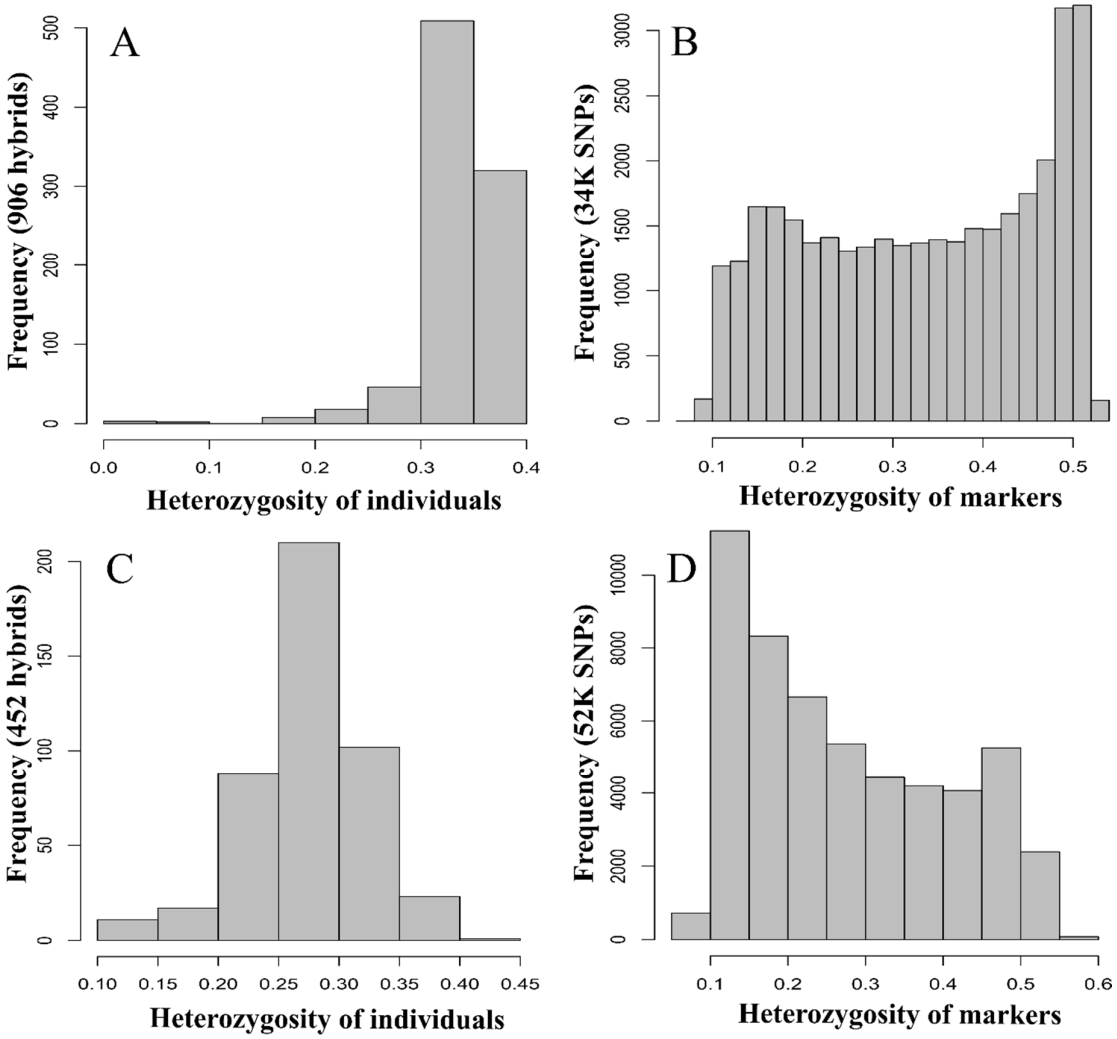


Figure S3. Frequency of heterozygous for individuals and markers. USP dataset with (A) 906 hybrids and (B) 34K SNPs. HELIX dataset with (C) 452 hybrids and (D) 52K SNPs.

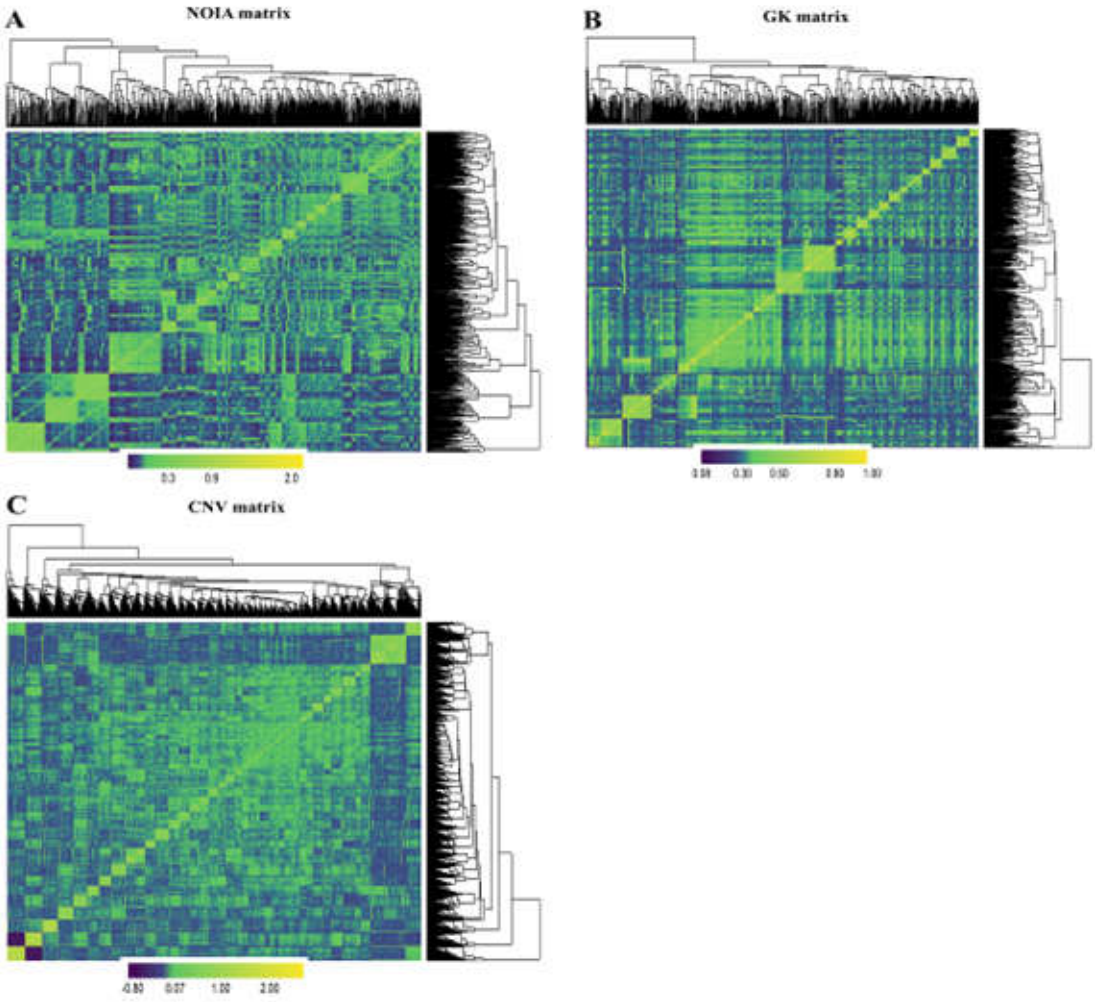


Figure S4. Heatmaps of genomic relationship matrix (GRM) from 906 maize single-crosses. (A) NOIA matrix, (B) GK matrix from grain yield ideal N, (C) CNV matrix.

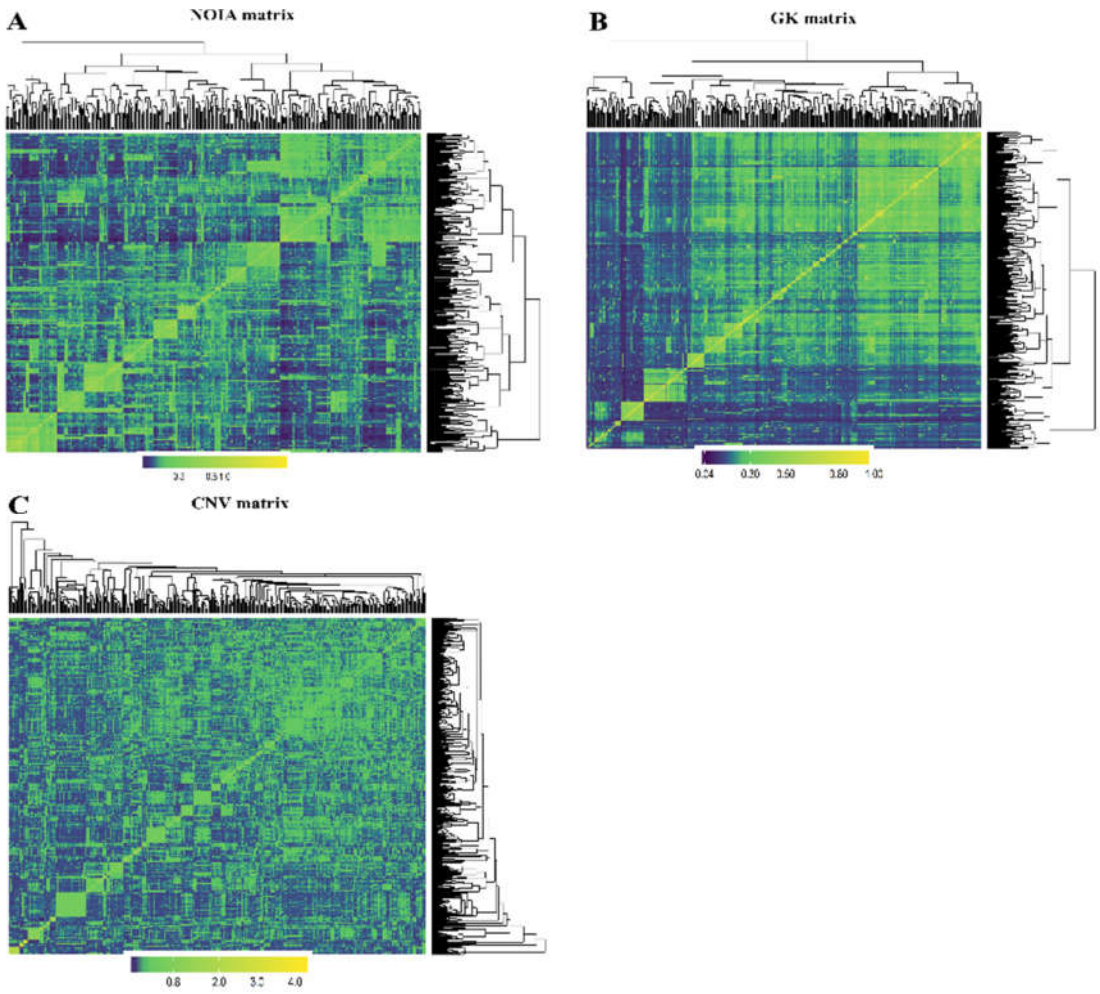


Figure S5. Heatmaps of genomic relationship matrix (GRM) from 452 maize single-crosses. (A) NOIA matrix, (B) GK matrix from grain yield, (C) CNV matrix.

TABLES

Table 1. Estimate of phenotypic (r_p) and genetic (r_g) correlations

Dataset	Trait	CG		CL		G/L	
		r_p	r_g	r_p	r_g	r_p	r_g
USP	GYLN	0.05^{ns}	0.16^{ns}	-0.06 *	0.08^{ns}	0.00^{ns}	0.18^{ns}
	GYIN	0.05^{ns}	0.11^{ns}	-0.12**	0.03^{ns}	-0.02^{ns}	0.12^{ns}
	PHLN	0.06*	0.01^{ns}	-0.18**	-0.08 ^{ns}	-0.05^{ns}	-0.01^{ns}
	PHIN	0.06*	0.03^{ns}	-0.19**	-0.08 ^{ns}	-0.05^{ns}	0.00^{ns}
	SGLN	0.10**	0.07^{ns}	-0.02^{ns}	-0.21^{ns}	0.07*	-0.10^{ns}
	SGIN	0.08**	0.06^{ns}	0.04^{ns}	-0.21^{ns}	0.09**	-0.11^{ns}
	HM	0.05^{ns}	0.15^{ns}	-0.09**	0.06^{ns}	0.00^{ns}	0.17^{ns}
HELIX	GY	0.00^{ns}	0.15^{ns}	-0.14**	-0.05^{ns}	-0.10**	0.06^{ns}
	PH	-0.05^{ns}	0.04^{ns}	0.03^{ns}	0.08^{ns}	-0.01^{ns}	-0.03^{ns}
	EH	0.02^{ns}	0.01^{ns}	-0.01^{ns}	-0.03^{ns}	0.00^{ns}	-0.01^{ns}

^{ns}Not significant; significant at 5% (*) or 1% (**) level

Traits evaluated in low (LN) and ideal (IN) nitrogen are grain yield (GY, ton ha^{-1}), plant height (PH, cm), ear height (PH, cm), stay green (SG), and harmonic mean (HM, ton ha^{-1}).

Copy gain (CG), copy loss (CL), and gain and loss ratio (G/L).

Table 2 Estimate of variance components from 906 hybrids (USP dataset)

Site	Model	GY				PH				SG				HM			
		σ_a^2	σ_d^2	σ_e^2	DIC	σ_a^2	σ_d^2	σ_e^2	DIC	σ_a^2	σ_d^2	σ_e^2	DIC	σ_a^2	σ_d^2	σ_e^2	DIC
LN	$NOIA_a$	0.15±0.01	-	0.39±0.01	1331	61.3±3.13	-	63.3±3.87	4791	0.02±0.00	-	0.04±0.00	-93.8	0.20±0.01	-	0.35±0.01	1268
	$NOIA_d$	-	0.08±0.00	0.48±0.01	1539	-	16.3±0.94	120±4.38	5263	-	0.01±0.00	0.07±0.00	261	-	0.08±0.00	0.49±0.01	1541
	$NOIA_{ad}$	0.16±0.01	0.09±0.01	0.28±0.01	1262	62.7±3.97	15.3±1.83	48.5±3.18	4747	0.02±0.00	0.00±0.00	0.03±0.00	-108	0.21±0.01	0.10±0.01	0.24±0.01	1169
	GK_a	0.39±0.04	-	0.29±0.01	1275	181±14.3	-	51.1±3.32	4746	0.09±0.00	-	0.04±0.00	-100	0.53±0.06	-	0.25±0.01	1181
	GK_d	-	3.49±0.32	0.29±0.01	1281	-	873±52.9	44.7±3.75	4762	-	0.48±0.03	0.03±0.00	-113	-	3.96±0.28	0.24±0.01	1196
	GK_{ad}	0.31±0.02	0.49±0.24	0.30±0.01	1278	158±12.4	75.8±22.2	49.3±3.25	4732	0.08±0.00	0.04±0.00	0.04±0.00	-103	0.37±0.05	0.97±0.33	0.24±0.01	1173
	CNV	0.21±0.01	-	0.37±0.01	1304	87.7±5.79	-	55.5±4.54	4716	0.04±0.00	-	0.04±0.00	-91.1	0.30±0.02	-	0.33±0.01	1237
IN	$NOIA_a$	0.26±0.01	-	0.50±0.02	1508	83.4±3.85	-	67.9±4.24	4839	0.07±0.00	-	0.08±0.00	325	-	-	-	-
	$NOIA_d$	-	0.12±0.00	0.67±0.02	1766	-	18.8±0.91	141±4.54	5375	-	0.02±0.00	0.18±0.00	855	-	-	-	-
	$NOIA_{ad}$	0.27±0.01	0.14±0.01	0.33±0.02	1402	85.9±5.22	18.8±2.5	47.7±4.05	4756	0.08±0.00	0.02±0.00	0.07±0.00	312	-	-	-	-
	GK_a	0.69±0.07	-	0.34±0.02	1408	269±21.8	-	51.1±3.82	4766	0.27±0.01	-	0.07±0.00	317	-	-	-	-
	GK_d	-	5.42±0.41	0.34±0.02	1427	-	1089±65.4	44.0±4.34	4779	-	1.24±0.06	0.06±0.00	323	-	-	-	-
	GK_{ad}	0.51±0.06	1.30±0.67	0.33±0.02	1402	225±21.3	122±64.6	49.4±3.78	4757	0.25±0.01	0.09±0.01	0.07±0.00	314	-	-	-	-
	CNV	0.39±0.03	-	0.46±0.02	1465	119±8.87	-	57.9±5.28	4746	0.16±0.00	-	0.08±0.00	286	-	-	-	-

Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_e^2), and deviance information criterion (DIC).
Traits evaluated in low (LN) and ideal (IN) nitrogen are grain yield (GY, ton ha^{-1}), plant height (PH, cm), stay green (SG), and harmonic mean (HM, ton ha^{-1}).
Data are mean ± standard deviation (SD) estimated from fifty replications in independent validation.
Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV.

Table 3 Estimate of genomic heritability from 906 hybrids (USP dataset)

Site	Model	GY			PH			SG			HM		
		h_a^2	d^2	h_{ad}^2	h_a^2	d^2	h_{ad}^2	h_a^2	d^2	h_{ad}^2	h_a^2	d^2	h_{ad}^2
LN	$NOIA_a$	0.28 ± 0.05	-	-	0.48 ± 0.06	-	-	0.33 ± 0.05	-	-	0.34 ± 0.06	-	-
	$NOIA_d$	-	0.15 ± 0.02	-	-	0.12 ± 0.02	-	-	0.12 ± 0.02	-	-	0.15 ± 0.02	-
	$NOIA_{ad}$	0.30 ± 0.06	0.18 ± 0.06	0.48 ± 0.06	0.50 ± 0.06	0.12 ± 0.06	0.62 ± 0.05	0.35 ± 0.06	0.12 ± 0.04	0.47 ± 0.05	0.39 ± 0.06	0.17 ± 0.06	0.56 ± 0.06
	GK_a	0.57 ± 0.06	-	-	0.78 ± 0.04	-	-	0.66 ± 0.05	-	-	0.67 ± 0.06	-	-
	GK_d	-	0.91 ± 0.01	-	-	0.94 ± 0.01	-	-	0.92 ± 0.01	-	-	0.94 ± 0.01	-
	GK_{ad}	0.29 ± 0.07	0.42 ± 0.19	0.71 ± 0.09	0.56 ± 0.04	0.26 ± 0.15	0.82 ± 0.03	0.50 ± 0.06	0.25 ± 0.17	0.75 ± 0.05	0.24 ± 0.06	0.59 ± 0.23	0.83 ± 0.10
	CNV	0.36 ± 0.06	-	-	0.60 ± 0.07	-	-	0.50 ± 0.05	-	-	0.46 ± 0.06	-	-
IN	$NOIA_a$	0.34 ± 0.05	-	-	0.53 ± 0.06	-	-	0.46 ± 0.05	-	-	-	-	-
	$NOIA_d$	-	0.16 ± 0.02	-	-	0.12 ± 0.02	0.07 ± 0.02	-	0.11 ± 0.02	-	-	-	-
	$NOIA_{ad}$	0.36 ± 0.06	0.19 ± 0.06	0.56 ± 0.05	0.56 ± 0.06	0.12 ± 0.06	0.68 ± 0.05	0.46 ± 0.06	0.12 ± 0.05	0.58 ± 0.05	-	-	-
	GK_a	0.66 ± 0.05	-	-	0.84 ± 0.04	-	-	0.76 ± 0.04	-	-	-	-	-
	GK_d	-	0.93 ± 0.01	-	-	0.95 ± 0.00	-	-	0.94 ± 0.01	-	-	-	-
	GK_{ad}	0.25 ± 0.07	0.57 ± 0.19	0.82 ± 0.08	0.58 ± 0.04	0.29 ± 0.15	0.87 ± 0.03	0.62 ± 0.04	0.21 ± 0.14	0.83 ± 0.03	-	-	-
	CNV	0.46 ± 0.07	-	-	0.67 ± 0.06	-	-	0.65 ± 0.05	-	-	-	-	-

Narrow sense genomic heritability (h_a^2), the proportion of the total phenotypic variance explained by genomic dominance (d^2), and broad sense genomic heritability (h_{ad}^2).
Traits evaluated in low (LN) and ideal (IN) nitrogen are grain yield (GY, ton ha^{-1}), plant height (PH, cm), stay green (SG), and harmonic mean (HM, ton ha^{-1}).
Data are mean ± standard deviation (SD) estimated from fifty replications in independent validation.
Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV.

Table 4 Estimate of variance components from 452 hybrids (HELIX dataset)

Model	GY				PH				EH			
	σ_a^2	σ_d^2	σ_e^2	DIC	σ_a^2	σ_d^2	σ_e^2	DIC	σ_a^2	σ_d^2	σ_e^2	DIC
$NOIA_a$	0.28 ± 0.02	-	0.48 ± 0.02	764 ± 19.3	214 ± 17.5	-	76.6 ± 6.05	2508 ± 26.4	71.8 ± 6.54	-	24.6 ± 1.53	2125 ± 19.6
$NOIA_d$	-	1.06 ± 0.11	0.38 ± 0.04	771 ± 27.9	-	154 ± 25.9	198 ± 12.6	2817 ± 17.7	-	47.6 ± 8.79	59.6 ± 3.98	2411 ± 14.7
$NOIA_{ad}$	0.26 ± 0.02	0.70 ± 0.05	0.13 ± 0.00	489 ± 20.9	203 ± 12.9	90.1 ± 8.68	44.9 ± 5.18	2406 ± 34.1	67.7 ± 7.31	30.9 ± 2.89	13.4 ± 1.08	2005 ± 22.7
GK_a	0.77 ± 0.04	-	0.16 ± 0.01	544 ± 31.7	518 ± 22.2	-	40.8 ± 5.34	2388 ± 39.2	147 ± 11.0	-	12.9 ± 1.18	2004 ± 25.5
GK_d	-	4.41 ± 0.20	0.14 ± 0.01	509 ± 23.5	-	764 ± 27.1	36.9 ± 3.69	2392 ± 30.8	-	224 ± 10.5	11.5 ± 0.00	1998 ± 23.6
GK_{ad}	0.29 ± 0.03	2.78 ± 0.33	0.13 ± 0.01	494 ± 29.3	445 ± 23.7	94.3 ± 11.0	38.4 ± 4.39	2378 ± 35.0	124 ± 10.9	31.7 ± 4.88	12.0 ± 0.98	1990 ± 24.4
CNV	0.44 ± 0.04	-	0.46 ± 0.03	769 ± 25.2	270 ± 23.3	-	83.9 ± 0.02	2549 ± 20.6	77.5 ± 6.87	-	25.0 ± 1.59	2138 ± 20.2

Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_e^2), and deviance information criterion (*DIC*).
Traits are grain yield (GY, ton ha^{-1}), plant height (PH, cm), and ear height (PH, cm).
Data are mean ± standard deviation (SD) estimated from fifty replications in independent validation.
Models reported are additive, dominance, and additive-dominance NOIA/GK, and CNV.

4. CONTROLLING POPULATION STRUCTURE IN THE GENOMIC PREDICTION OF TROPICAL MAIZE HYBRIDS

ABSTRACT

Genomic prediction of single-crosses is a promising tool in maize breeding, reducing extensive field trial evaluation and increasing genetic gain per cycle. However, factors such as population structure (PS) can influence the accuracy of estimates of genomic breeding values (GEBV). In this study, we assessed PS in 452 hybrids; and applied the information into genomic prediction schemes, using (1) traditional GBLUP and four adjustment methods for PS, (2) a reparameterized Bayesian Whole-Genome Random Regression (WGRR) model, (3) within- and between-group hybrids prediction, and (4) within- (W-GBLUP) and multi-group (MG-GBLUP) analyses in stratified groups. Three groups were identified (K1, K2, and K3) in the hybrids, based on fineSTRUCTURE results. Adding four different sets of PS as covariates to GBLUP did not improve the prediction accuracy (r_{MP}) for grain yield (GY) and plant height (PH). However, using nonmetric multidimensional scaling dimensions and fineSTRUCTURE group clustering increased reliability for GY and PH, respectively. High r_{MP} for GY and PH were observed for within-group hybrids L1L1 (0.79; 0.78) and low r_{MP} for between-group hybrids L1L2 (0.43; 0.62) and L1L3 (0.66; 0.59). W-GBLUP analysis in the stratified groups resulted in low r_{MP}. On the other hand, MG-GBLUP showed high r_{MP} relative to W-GBLUP for both traits. Predicting by GBLUP with PS covariates is the best approach, increasing reliability and reducing bias. In addition, MG-GBLUP in stratified groups could be an efficient method, depending on the number of hybrids available in the breeding program.

Keywords: Stratified groups; MG-GBLUP; Linkage disequilibrium; Accuracy

4.1. INTRODUCTION

Tropical maize represents one of the most diverse source of germplasm used in several plant breeding programs (FAN *et al.* 2015; TEIXEIRA *et al.* 2015). Recently, high-density single-nucleotide polymorphisms (SNPs) have been used to characterize the heterotic pools via genetic diversity (OYEKUNLE *et al.* 2015) and population structure analysis (DA SILVA *et al.* 2015; NELSON *et al.* 2016). Moreover, the applicability of such diversity information extends to association studies (CHEN and LIPKA 2016), genomic prediction (MARULANDA *et al.* 2016), and germplasm architecture (BERNARDO and THOMPSON 2016).

Population structure (PS) in maize could arise from local adaptation or diversifying selection (OROZCO-RAMIREZ *et al.* 2016). For the temperate maize, several subpopulations/groups (flint, dent, stiff stalk, and non-stiff stalk) were described according to

morphological, genetic and environmental adaptability characteristics (SCHAEFER and BERNARDO 2013; RINCENT *et al.* 2014). However, the tropical materials are not well organized like the temperate, which can be due to the much stronger divergence of heterotic groups by long-term selection (WU *et al.* 2016). For example, in the International Maize and Wheat Improvement Center (CIMMYT), development of Lowland Tropical and Subtropical/Midaltitude subgroups began in the mid-1980s; nonetheless, temperate materials started around 100 years ago (UNTERSEER *et al.* 2016; WU *et al.* 2016). Detailed description of PS in maize lines of Brazil was reported by LANES *et al.* (2014). In this study, 81 microsatellite loci were screened for 90 maize parental inbreds of tropical hybrids to identify three heterotic pools (including tropical flint, semi-flint, and semi-dent), which generally agreed with what have been used by Brazilian maize seed companies.

Different ways to investigate PS can be classified into either non-model-based (or non-parametric) or model-based approaches. Non-parametric methods includes principal component analysis (PATTERSON *et al.* 2006; PRICE *et al.* 2006), discriminant analysis of principal components (JOMBART *et al.* 2010), and nonmetric multidimensional scaling (ZHU and YU 2009). For model-based clustering, the algorithm in ADMIXTURE v.1.23 (ALEXANDER *et al.* 2009), similar to STRUCTURE v.2.3.4, is a commonly used approach. Also, the recently developed ChromoPainter/fineSTRUCTURE v.2 (LAWSON *et al.* 2012) considers linkage disequilibrium (LD) patterns in the genome, aiming to make use of haplotype structure and extracting more information from the data. Furthermore, to identify the optimal number of clusters, methods such as k-means clustering (REIF *et al.* 2003; CROS *et al.* 2015; JAN *et al.* 2016), ADMIXTURE cross-validation (ALEXANDER *et al.* 2009), and ΔK Evanno criterion (EVANNO *et al.* 2005) are well adopted ones in practice.

Population structure variables have been proven very useful for many different applications, especially in association and prediction analyses. Generally, using PS as covariates could control potential confounding factors and improve statistical power by reducing residual variance (ASCHARD *et al.* 2015). In association studies, principal components (PCs) and admixture coefficients have been successfully used as fixed effects (covariates) in mixed-model equations (YU *et al.* 2006; PRICE *et al.* 2010; TUCKER *et al.* 2014). On the other hand, using PCs in genomic best linear unbiased prediction (GBLUP) model may result in an ill-posed model because the PCs enter both as fixed effects and implicitly through the random effect (DE LOS CAMPOS and SORENSEN 2014). Hence, Janss et al. (2012) proposed a reparameterized Bayesian Whole-Genome Random Regression (WGRR) model to handle this problem, drawing inferences based on all or some PCs, allowing a natural separation of across- and within-subpopulation genetic

variance. In plant breeding, GUO *et al.* (2014) applied this model in maize and rice populations to control PS, and found that the majority of genomic heritability was contributed by within-subpopulation genetic variance and, in all traits, the prediction accuracy was reduced when included PS information.

In genomic prediction, the presence of hidden or known structure, and family relatedness within a breeding population is critical when evaluating genomic estimated breeding values (GEBV), genomic heritability, and prediction accuracy, because it could lead to biased estimations (WINDHAUSEN *et al.* 2012; LEHERMEIER *et al.* 2014; UNTERSEER *et al.* 2014; ISIDRO *et al.* 2015; SPINDEL *et al.* 2015). Therefore, a common approach on prediction analysis is partitioning the genomic variability into within- and between-group components (TECHNOW *et al.* 2012). In animal breeding, within-group estimates of GEBV can be more accurate than between-group (SAATCHI *et al.* 2011; VENTURA *et al.* 2016), which can be due to non-persistent associations or inconsistent LD between SNPs and QTL across populations (HAYES *et al.* 2009; IHESHIULOR *et al.* 2016). However, in plant breeding, exploiting within-group analyses may not always improve prediction accuracy (SCHULZ-STREECK *et al.* 2012; CROS *et al.* 2015). It is proven that splitting the breeding population into subgroups could lead to a reduction of population size, loss of diversity, and besides that, it is assumed uncorrelated marker effects between subpopulations (RIEDELSHEIMER *et al.* 2013; ALBRECHT *et al.* 2014; HUANG *et al.* 2016). In order to overcome this last drawback, LEHERMEIER *et al.* (2015) proposed a multi-group (MG-GBLUP) analysis to control heterogeneity of marker effects between subpopulations, and found promising results depending on the genetic architecture of the trait.

In a typical maize hybrid breeding, inbred lines of different heterotic groups are crossed, and although assuming that two alleles share a common genetic background in hybrids, it is essential to find patterns of PS, and apply this information on genome-based prediction, in an attempt to identify high performing hybrids (ALBRECHT *et al.* 2014; LEHERMEIER *et al.* 2015). Therefore, our objectives were (i) to investigate PS in a set of tropical maize inbreds and the derived hybrids, and (ii) to control PS in genomic predictions of hybrids in four scenarios, using: (1) the traditional GBLUP and four adjustment methods of PS, (2) the reparameterized Bayesian WGRR model, (3) within- and between-group hybrids prediction, and (4) the within- (W-GBLUP) and multi-group (MG-GBLUP) analysis in stratified groups.

4.2. Materials and methods

4.2.1. Phenotypic data

We used 452 maize single-crosses (hybrid dataset) provided by Helix Sementes®, São Paulo, Brazil. The hybrids represent a partial diallel mating design between 128 tropical inbred lines (inbred dataset). No heterotic group information was available. The field design used was randomized complete block with two replications. Experimental trials were carried out in five sites in southern, southeastern, and west-central regions of Brazil in the first growing season of 2014/15. Details about the sites see SOUSA *et al.* (2017). The hybrids analyzed in each location varied, thus creating an unbalanced experiment. Two-row plots of 5 m spaced 0.70 m were used. Sowing density was about 63,000 kernels per hectare, under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, ton ha^{-1}) and plant height (PH, cm). Plots were mechanically harvested and converted to 13% moisture, and plant height measured from soil surface to the flag leaf collar on one representative plant within each plot (company criteria). We used a linear mixed model to calculate BLUPs for hybrids, including site as fixed effect, and hybrid and interaction as random effects. Heterogeneous residual variance structure was assumed across sites. Variance components and entry-mean based heritability were obtained for GY and PH, and the significance of the random effects of hybrids was assessed by the Likelihood Ratio Test (LRT) at 5% probability, using ASReml-R (BUTLER *et al.* 2009).

4.2.2. Genotypic data

The genotyping of the inbreds was performed by Affymetrix® platform, containing 614,000 SNPs (UNTERSEER *et al.* 2014). Markers with low call rate (<95%) and with at least one heterozygous combination were removed. Imputation was done based on homozygosity of an individual and marker frequency with missed points. Polymorphic SNP markers were used to build the hybrid genotype dataset, deduced by combining the genotypes from its two parents. Afterwards, minor allele frequency was conducted over hybrid markers considering the threshold of 0.05, resulting in a total of 52,700 high-quality SNPs distributed in the ten maize chromosomes as follows: (1) 7015, (2) 6020, (3) 6072, (4) 5953, (5) 6431, (6) 4736, (7) 5197, (8) 4436, (9) 3529, and (10) 3311.

Linkage disequilibrium (LD) among markers may lead to unstable estimates of PS (CAMPOY *et al.* 2016; GALINSKY *et al.* 2016). Therefore, we thinned both datasets using PLINK v.1.9 (PURCELL *et al.* 2007) by removing SNPs that were in LD, with a pairwise r^2 value greater than 0.7 within a 50-

SNP sliding window which was advanced by 10 SNPs each time. The final genomic data was 32,838 SNPs for the inbred dataset and 26,210 SNPs for the hybrid dataset, which was used as input to perform PS analysis.

4.2.3. Inference of population structure

Inbred dataset

We used four approaches to detect PS: (a) principal component analysis (PCA), (b) nonmetric multidimensional scaling (nMDS), (c) ADMIXTURE, and (d) ChromoPainter/fineSTRUCTURE. PCA was performed using SNPRelate-R (ZHENG *et al.* 2012) in the raw SNP data (32,838 SNPs), and the results were presented as two and three-dimensional principal component scores plots. For nMDS analysis, labdsv-R (ROBERTS 2016) was used in the roger's distance matrix, with three dimensions, and the first two dimensions were plotted.

ADMIXTURE was used to perform a maximum likelihood estimation of individual ancestries; and ChromoPainter and fineSTRUCTURE was used to find patterns of haplotype similarity. Firstly, we applied the ChromoPainter unlinked model on haplotypes, with 10 expectation maximization (EM) steps. Secondly, fineSTRUCTURE was used to perform Markov Chain Monte Carlo (MCMC) analysis with 100,000 burn-in iterations and sample iterations with a thinning interval of 1,000. Normalization parameter c was calculated following the unlinked case, $c=1/(N-1)$, where N is the number of individuals. Visualization of the posterior distribution of clusters was performed using the tree-building algorithm, and the number of clusters was inferred by, arbitrarily setting a cut-off in the tree.

To estimate the optimal number of clusters, the cross-validation errors from 1-12 K were analyzed in ADMIXTURE, and the Bayesian Information Criterion (BIC) values in k-means clustering, implemented in adegenet 2.0.1-R (JOMBART *et al.* 2015). Furthermore, to visualize the genetic differences between inbred lines, a neighbor-joining tree (NJT) was generated based on the Modified Rogers' distance. We also investigated LD structure within 70 kb of distance among all pairs of markers (32,838 SNPs), using PLINK v.1.9, and the values were reported as the average r^2 across 10 chromosomes.

Hybrid dataset

We used PCA, nMDS, and fineSTRUCTURE to detect PS following the same procedure of the inbred dataset. In addition, we built an artificial ADMIXTURE coefficient for the hybrids, following the equation: $ADM_{12} = \frac{ADM_{P1} + ADM_{P2}}{2}$, where ADM is the admixture coefficient of each parent, ranging from 0 to 1.

In order to visualize and describe related individuals, we used Discriminant Analysis of Principal Components (DAPC) (JOMBART *et al.* 2010), using the inferred groups of fineSTRUCTURE. The number of principal components to be retained in the discriminant analysis was set to 15 following alpha-score optimization, a method that finds a trade-off between discriminative power and model over-fitting. We also plotted the genomic relationship matrix (GRM) by a network graph, in which two hybrids were linked when their relationship coefficient was ≥0.6. The networks were visualized using the igraph-R with the Fruchterman Reingold layout.

4.2.4. Statistical models

Traditional GBLUP model

Additive-dominance GBLUP was used by fitting the following model:

$$y = Xb + Z_a a + Z_d d + e \quad (1)$$

where y is a vector of BLUP values of hybrids, b is a vector of fixed effects, a is a vector of additive genetic effects of the individuals, d is the vector of dominance effects, and e is a vector of random residuals. X, Z_a and Z_d are the incidence matrices for b, a, and d. The distributions assumed were $a \sim N(0, \sigma_a^2 G_a)$, $d \sim N(0, \sigma_d^2 G_d)$, and $e \sim N(0, \sigma_e^2 I_m)$. For genomic prediction, we used the non-pruned 52K SNP matrix. G_a and G_d are the additive and dominance genomic relationship matrix (GRM), following the equation: $G_a = \frac{W_A W_A'}{2\sum_{i=1}^{n} p_i(1 - p_i)}$ and $G_d = \frac{W_D W_D'}{4\sum_{i=1}^{n} (p_i(1 - p_i))^2}$, where p_i is frequency of one allele of the locus i and W is the matrix of incidence of markers (VANRADEN 2008; DA *et al.* 2014). The W_A matrix was coded as 0 for homozygote A_1A_1, 1 for heterozygote A_1A_2 and 2 for homozygote A_2A_2, for W_D was considered 0 for both homozygotes and 1 to heterozygote.

PS covariates

We applied the Q+K model (YU *et al.* 2006) on genomic prediction of hybrids for GY and PH, using four contrasting Q approaches. A model that (a) includes first three PCs (PC), (b) includes three dimensions of nonmetric multidimensional scaling (nMDS), (c) includes the artificial admixture coefficients (ADM), and (d) includes a matrix of zeros and ones based on fineSTRUCTURE group clustering (FINE). The PS related variables were used as fixed covariates in the GBLUP model. Furthermore, to select the top PCs (PATTERSON *et al.* 2006), we evaluated the number of statistically significant principal components, measured by the Tracy-Widom test using LEA-R (FRICHOT and FRANCOIS 2015), and added a varied number of PCs (3, 5, 10, 14) in GBLUP.

All variance components were determined using Bayesian generalized linear regression (BGLR) (PEREZ and DE LOS CAMPOS 2014) for the five mixed-models. We reported posterior mean estimates and standard deviations of the additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), and broad sense genomic heritability (h_g^2). We used a total of 30,000 MCMC iterations, 5,000 for burn-in, and 5 for thinning.

From GP models, we evaluated prediction accuracy (r_{MP}), correlation between BLUP values and predicted phenotypic values of the hybrids, from fifty replications, randomly sampling 75% of the hybrids to form the training set (TS) and the rest as validation set (VS). In addition, reliability (REL) (GORJANC *et al.* 2015) and deviance information criterion (DIC) (SHRINER and YI 2009) were used to compare the model performance. REL was calculated according to the formula: $REL = 1 - (PEV / \sigma_g^2)$, where *PEV* is the variance of prediction errors of the GEBV of the hybrid ($\hat{g}_i$). Note $PEV = SE(\hat{g}_i)^2 = var(g_i - \hat{g}_i)$, where *SE* is the standard error. The model with the highest *REL* value presented the best precision in an earlier study (HE *et al.* 2016). For DIC, the following equation was used: $DIC = D(\bar{\theta}) + 2pD$, where $D(\bar{\theta})$ is the deviance at the posterior mean of the model, and pD is the effective number of parameters. The model with the lowest DIC value presents the best data fit. The mean values of r_{MP}, *REL* and *AIC* estimated from fifty replications in independent were used in the overall model performance comparison.

Reparameterized Bayesian WGRR model

The reparameterized Bayesian WGRR model, proposed by Janss et al. (2012) was used to make prediction for across- and within-group based on PCA. Firstly, we followed the model:

$$y = \mu + U_a\alpha + U_d d + e$$

(2)

where U_a and U_d is an $n \times (n - 1)$ matrix of the eigenvectors (principal components) of G_a and G_d with n individuals, and a and d is an $(n - 1) \times 1$ vector of random effects. This model is equivalent to the additive-dominance GBLUP (JANSS *et al.* 2012; DE LOS CAMPOS and SORENSEN 2014). Secondly, we split the model for across-group:

$$y = \mu + \sum_{i=1}^{d} U_{a_i}\alpha_i + \sum_{i=1}^{d} U_{d_i} d_i + e \quad (3)$$

and within-group:

$$y = \mu + \sum_{i=d+1}^{n} U_{a_i}\alpha_i + \sum_{i=d+1}^{n} U_{d_i} d_i + e \quad (4)$$

where the model (3) represents the genomic variance explained by the first dominant d eigenvectors, and model (4) explains the remaining (n-d) eigenvectors. We used 3, 5, 10, 20, 40, 60, 80, and 100 as first d elements, to compare across- and within-group prediction accuracy. We used a total of 30,000 MCMC iterations and 5,000 for burn-in, using the BRR model implemented in BGLR-R (DE LOS CAMPOS and PÉREZ-RODRIGUEZ 2015). The prediction accuracy was assessed from fifty replications, randomly sampling 75% of the hybrids to form the training set (TS) and the rest as validation set (VS).

Hybrids group formation

We used two approaches to build groups of hybrids to assess prediction accuracy from fifty replications in independent validation. First, we used the inbred dataset (parents) to form hybrids between lines that belong to the same heterotic group (within-group hybrids) and hybrids between lines that belong to different groups (between-group hybrids). Second, we separated all 452 hybrids into stratified groups. In both cases, we used fineSTRUCTURE group clustering.

Within- and multi-group analysis

We used the stratified groups (subgroups) to make inferences of hybrid prediction using two approaches, proposed and detailed in Lehermeier et al. (2015). The first is a stratified within-group analysis (W-GBLUP), estimating marker effects and variance components within each K separately, with a specific GRM. The second method is a multivariate approach that uses multi-group data and accounts for heterogeneity (MG-GBLUP), with population-specific marker

effects that can be correlated between subpopulations. We used the additive (G_a) and dominance (G_d) GRM, and reported posterior mean estimates and standard deviations of the additive variance ($\sigma^2_{a_k}$), dominance variance ($\sigma^2_{a_k}$), error variance ($\sigma^2_{\varepsilon_k}$), and broad sense genomic heritability ($h^2_{g_k}$) for each K. We reported the posterior mean estimates and standard deviations of genomic correlations derived from $\sum_g$ (genomic variance-covariance matrix among groups) for both traits. We used a total of 30,000 MCMC iterations and 5,000 for burn-in to estimate these parameters, using the MTM-R package.

The prediction accuracy of W-GBLUP and MG-GBLUP were assessed with fifty replications from independent validation, randomly sampling 75% of the hybrids to form the training set (TS) and the rest as validation set (VS). According to Lehermeier et al. (2015), MG-GBLUP estimates different genomic values for each individual inside K, where the estimated values $\boldsymbol{\sigma}^2_{g_k}$ are specific for each K and is used as estimated values for all individuals belonging to that particular group.

To each group K, marker effect based on the adjusted entry means for GY and PH were estimated, using rrBLUP-R (ENDELMAN 2015). Besides that, LD structure was investigated within 70 kb of distance among all pairs of markers, and the values were reported as the average r^2 across 10 chromosomes.

4.3. RESULTS

Inbred PS

In ADMIXTURE analysis, the optimal number of clusters was L=7 with the smallest cross-validation error (Additional file 1: Figure S1A, Figure S2). The k-means clustering identified L=3 with the smallest BIC value (Additional file 1: Figure S1B). The fineSTRUCTURE result is a coancestry heatmap, which shows the amount of shared genetic chunks between the inbred lines. We defined a cutoff on the maximum *a posteriori* tree with three groups (L), each containing 100 (L1), 13 (L2), and 15 (L3) inbred lines (Additional file 3: Figure S3). Moreover, PCA, nMDS, and cluster (NJ tree) analysis also revealed levels of PS identified in both model-based clustering (Additional file 4: Figure S4). The first two PCs explained 5.36% and 4.24% of the total variance, clearly splitting the groups along the axis. However, nMDS analysis revealed that L1 and L3 were clustered together, but separated from L2. The relationship between LD and physical distance was plotted (Additional file 4: Figure S4D), and the LD decayed faster with the r^2 dropping to half its maximum value within 1.3 kb.

Hybrid PS

The unlinked coancestry heatmap of fineSTRUCTURE clustered hybrids into three groups (K), containing 113 (K1), 121 (K2), and 218 (K3) hybrids (Fig. 1A). Three subgroups within-group K1 were also clearly shown. In the artificial admixture coefficients (Fig. 1B), we found a mixture of groups in the hybrids. PCA and nMDS dots were color-coded based on the fineSTRUCTURE group clustering. The first two PCs explained 7.40% and 6.05% of the total variance (Fig. 1C). Furthermore, the 3-D PCA score plot (Additional file 5: Figure S5A) revealed a clear separation of K1 from K2, wherein PC1, PC2, and PC3 together explained 18.3% of data variation. The within-group individuals of K1 were spread along the axis (blue density plot), confirming the subgroups identified in fineSTRUCTURE (Fig. 1A; S5A). In addition, a pattern also was detected for nMDS analysis (Fig. 1D). Network graph revealed that individuals from K2 and K3 are more related according to the GRM (Fig. 1E).

The DAPC plot (Additional file 5: Figure S5B) using two discriminant functions indicated that K1 were highly discriminated from K2, with strong separation along the principle component axes. The plot did not reveal high discrimination between K2 and K3, since overlapping existed between groups.

Hybrid prediction

From the phenotypic analysis, it was found significant differences in the hybrids by the likelihood ratio test ($P<0.05$), for GY and PH. Entry-mean based heritability was 0.77 for GY and 0.86, reflecting good accuracy of phenotypic evaluation. The BLUP mean for GY varied from 3.39 to 9.37 ton ha^{-1}, and for PH from 185 to 277 cm.

Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), broad sense genomic heritability (h_g^2), reliability (*REL*), and deviance information criterion (*DIC*), for each trait and model were presented in Table 1. Estimates of variance components and genomic heritability varied slightly among models. The r_{MP} did not vary among all tested models for GY and PH (Fig. 2), showing low SD for PC and nMDS. However, the highest *REL* and lowest *DIC* was observed in nMDS and FINE for GY. For PH, FINE and ADM was the best models for *REL* and *DIC*.

The proportion of variance explained by the first two eigenvectors was relatively small (Fig. 3A). Based on Tracy-Widom test, the significant axes of variation to account for genetic structure were 14 (Fig. 3A). For both traits, the r_{MP} slightly decreased when added more than three PCs in GBLUP model (Fig. 3B), showing that three PCs in the model could be efficient to account population structure.

The r_{MP} of Bayesian WGRR model (Equation 2) performed the same as the additive-dominance GBLUP, for GY (0.74) and PH (0.80). However, the r_{MP} for GY in across-group with d=3 and d=5 was 0.30 and 0.35, and for within-group was 0.62 and 0.62, respectively (Fig. 4). For PH, across-group with d=3, was 0.22 and for within-subpopulation 0.77.

Within- and between-group hybrids prediction

We used within- (L1L1) and between-group (L1L2, L1L3) hybrids to assess prediction accuracy for GY and PH (Fig. 5). Since L2L2, L2L3, and L3L3 groups were small, no prediction was done. We observed high r_{MP} and h_g for L1L1 and low for L1L2 and L1L3 for GY and PH, respectively.

Within- and multi-group analysis

We used within-group (K1, K2, and K3) hybrids to investigate genetic parameters for GY and PH (Table 2). For both traits, lower estimates of $h^2_{g_k}$ were observed from W-GBLUP compared with MG-GBLUP. Prediction accuracy varied between groups and traits, in both models. However, MG-GBLUP presented higher values of r_{MP} relative to W-GBLUP. Posterior mean estimates and posterior standard deviations of genomic correlations from MG-GBLUP for GY varied among the three groups 0.48±0.15 (K1-K2), 0.75±0.09 (K1-K3), and 0.34±0.14 (K2-K3). For PH, the values was 0.49±0.14 (K1-K2), 0.74±0.10 (K1-K3), and 0.34±0.16 (K2-K3).

The relationship between LD and physical distance (kb) was plotted for K (452), K1 (113), K2 (121), and K3 (218) (Additional file 6: Figure S6A). LD (r^2) rapidly decayed in accordance with the highest number of individuals inside the group. For K, K1, K2, and K3 the LD decayed with the r^2 dropping to half their maximum value within 5.5, 6.5, 10, and 11.5 kb, respectively. Additive marker effects distribution estimated across the groups were different for GY, but was similar for PH, in all ten chromosomes (Additional file 6: Figure S6A; Figure S6B). Pearson correlation (r) between group SNP effect for GY was 0.27 (K1-K2), 0.46 (K1-K3), and 0.14 (K2-K3). For PH, the r was 0.33 (K1-K2), 0.38 (K1-K3), and 0.46 (K2-K3).

4.4. DISCUSSION

The most common source of tropical germplasm found in the breeding programs are Tusón, Tuxpeño, Antigua Composite, Suwan-1, and Cuban Flint (also called Cateto in Brazil) (HALLAUER and CARENA 2014), and as observed in previous studies, the number of sub-groups inside tropical and sub-tropical still diverge (REIF *et al.* 2003; MOLIN *et al.* 2013; WU *et al.* 2016).

In the present study, 128 tropical inbred lines were characterized using k-means clustering and two model-based approaches to identify groups/clusters. Based on k-means we classified three groups, which were consistent according to fineSTRUCTURE (Additional file 3: Figure S3) and PCA (Additional file 4: Figure S4A). In within-group L1, five distinct subgroups were revealed, explaining the seven groups identified in ADMIXTURE results (Additional file 2: Figure S2A).

Another way to visualize the structure of populations is by the extent of linkage disequilibrium, which has influence on resolution of the genome-wide analysis (YANG *et al.* 2011). In our study, the LD decayed within 1.3 kb (Additional file 4: Figure S4D), which was consistent to the findings of UNTERSEER *et al.* (2014). These authors worked with 285 temperate and tropical maize lines genotyped with 600K SNPs, and found L=7 in ADMIXTURE, and observed fastest LD decay in (sub)tropical lines (70 kb) explained by the high heterogeneity inside group. CHIA *et al.* (2012) and YAN *et al.* (2009) also found fastest LD decay within distances between 5 and 10 kb, respectively, in highly diverse tropical maize lines.

In a maize breeding program, it is common verifying PS among inbred lines to explore heterosis in divergent parental crossing (FERNANDES *et al.* 2015; MUNDIM *et al.* 2015). However, hybrids generated from several divergent heterotic parents should present high levels of structuring, confirming our results identified from fineSTRUCTURE, PC, and DAPC results (Fig. 1 and Fig. S5). For example, within-group K1 (Fig. 1A) revealed three distinct sub-grouping, which can be identified in 2-D (Fig. 1C) and 3-D (Additional file 5: Figure S5) PCA graph.

In this work, both traits showed high levels of r_{MP} and h_g^2 for GY (0.74; 0.79) and PH (0.80; 0.86) from traditional GBLUP (Fig. 2A-B). Similar findings were observed by MAENHOUT *et al.* (2010), MASSMAN *et al.* (2013) and DOS SANTOS *et al.* (2016). Moreover, the methods GBLUP, PC, nMDS, ADM, and FINE were compared in terms of r_{MP}, *REL*, and *AIC* (Table 1). However, the prediction accuracy remained the same for the models. Clearly, the GRM implicitly captured genetic variation from PS and admixture of the hybrids. According to ISIDRO *et al.* (2015) traits are largely impacted by PS. Thus, prediction accuracies depend on the interaction of trait architecture and levels of PS. On contrary, including PS covariates reported herein showed better performance in terms of reliability and DIC, which could substantially reduce the standard error of the genetic variant association and depending on the causal correlation between the covariate and effect, increases power.

Several studies have been successfully conducted including PC as covariates in GWAS analysis (WANG *et al.* 2011; SUKUMARAN *et al.* 2015; ZHANG *et al.* 2016). However, in genomic prediction studies, adding PC eigenvectors in the model have shown low r_{MP} (DAETWYLER *et al.* 2012; NEWELL and JANNINK 2014). As already reported by JANSS *et al.* (2012), the PCs added as

fixed effects in the GBLUP enter twice in the model, causing misleading interpretations. It can be seen in this study, where even being observed structuring in PC plot (Fig. 1C), including the first three PCs did not change r_{MP} scenario (Fig. 2), and reduced REL for GY. Furthermore, we used Tracy-Widom test to select the top principal components, but the r_{MP} remained the same when added the first 5, 10, and 14 significant PCs in GBLUP model for both traits (Fig. 3B).

The prediction including three nMDS dimensions performed better than the others methods for GY in terms of reliability. In a GWAS analysis, ZHU and YU (2009) compared nMDS and PC, and found an increase in power and a decrease in false positive rate using nMDS associated with genomic kinship. In addition, SUKUMARAN *et al.* (2012) worked with PS of 300 wheat lines for ten grain quality traits, and tested three mixed models including admixture coefficients, nMDS, and PCA as fixed covariates in GWAS analysis. The authors found nMDS as the best approach for phosphorus (P) trait. On the other hand, in our results, ADM was the lowest ranked method so far according to *REL* for GY. In contrast, for PH showed better performance compared to GBLUP. In animal prediction, THOMASEN *et al.* (2013) studied US and Danish Jersey cattle by including admixture coefficients estimated in STRUCTURE in the genomic prediction model, and did not find any improvement of prediction reliabilities.

To predict across- and within-group, we followed the approach proposed by JANSS *et al.* (2012), but including dominance information. Within-group could derive genomic relationships that do not include the contribution to genetic similarity of the 1st *d* principal components of G_a and G_d (DE LOS CAMPOS *et al.* 2013). In our study, r_{MP} of traditional GBLUP (Equation 1) and reparameterized Bayesian WGRR model (Equation 2) were similar for both traits, but for within-group with d=3 (PC4 to PC100) for PH (0.77) was almost similar to traditional GBLUP (0.80). Similar application of the model was described by GUO *et al.* (2014) who worked with rice and maize panels, and used different numbers of principal components, d=4 and d=2, in the genomic model, respectively. The authors found that controlling for subpopulation structure significantly decreased r_{MP} in both panels, and concluded that within-group variation is a major resource of genetic variance.

We observed high values of BLUP mean in L1L3 (between-group hybrid) for GY (7.65) and PH (240). However, the r_{MP} was low for both traits (Fig. 5). In this case, the mean values could be related to high heterosis, and the r_{MP} to low genomic relationship. One reason could be the fact that between-group hybrids from genetically distant parents have a low degree of relationship. TECHNOW *et al.* (2014) also observed lower prediction accuracies in single-cross maize hybrids when the parents are distant relatives. Therefore, applying markers to predict hybrids when considering diallels that involve lines from the same heterotic group may yield

different results than when considering factorial designs that involve two separate sets of lines that belong to different heterotic groups (MELCHINGER *et al.* 1992; CHARCOSSET and ESSIOUX 1994).

Within- and between-group hybrids prediction showed a clear understanding of the importance of relatedness among hybrids and, certainly, new methods are required to increase estimates of prediction between-group hybrids. This could be seen in L1L2 group which presented lower number of individuals compared with L1L3, and the r_{MP} was higher (Fig. 5), confirming our results that unrelated individuals in the training population (TP) affected more the prediction. Similar results was observed by LORENZ and SMITH (2015) who observed that adding increasingly unrelated individuals to the TP reduced prediction compared with smaller TP consisting of highly related individuals. In another detailed study, RIEDELSHEIMER *et al.* (2013) compared prediction accuracies for within-population prediction using full sibs of 635 doubled haploid (DH) lines genotyped with 16K SNPs, and found higher values increasing the sample size. ALBRECHT *et al.* (2014) also worked with genome-based prediction within-, across- and whole-group of DH lines, genotyped with 56K SNPs, and found higher whole-group predictive ability compared with within- and across-group for grain dry matter content.

From our findings, r_{MP} was higher in MG-GBLUP when compared with W-GBLUP for both traits (Table 2). According to Lehermeier et al. (2015) MG-GBLUP allows subpopulation-specific marker effects, borrowing the information between subpopulations. In contrast, within-group prediction (W-GBLUP) reduces training size, nevertheless, increases the relationship between genotypes (RIEDELSHEIMER *et al.* 2012; LEHERMEIER *et al.* 2015; HUANG *et al.* 2016; IHESHIULOR *et al.* 2016; MENDES and DE SOUZA 2016). Higher estimates of r_{MP} were observed from MG-GBLUP in K3, for GY (0.77) and PH (0.84) relative to the whole-group from traditional GBLUP, showing the efficiency of the method. However, K1 and K2 presented lower r_{MP}, due to reduced sample size. For W-GBLUP, the group with the largest sample size (K3=218) achieved similar values of r_{MP} relative to the whole-population (K=452) GBLUP for GY (0.73; 0.73) and PH (0.78; 0.80), respectively (Table 1 and 2). The estimated genomic correlations between subpopulations K1-K3 was high for both traits, which is agreement with the GRM (Fig. 1E). In addition, genomic heritability ($h^2_{g_k}$) tended to be higher in MG-GBLUP relative to W-GBLUP. According to DE LOS CAMPOS *et al.* (2015), prediction and genomic heritability are two different problems, and a model that may yield higher values of $h^2_{g_k}$ may have a relatively poor prediction performance, and vice versa. In our case, we observed high broad sense genomic heritability is associated with higher prediction accuracies.

Recent studies showed contrasting results about whole- or within-group prediction. SCHULZ-STREECK *et al.* (2012) found better prediction accuracy joining all populations derived from five biparental populations of maize. RIEDELSHEIMER *et al.* (2012) also studied PS splitting the whole population in within-group of related lines, and showed that population structuring reduced prediction accuracy in 3.6% for SNPs relative to whole-population. On the other hand, MENDES and DE SOUZA (2016) working with 250 tropical maize single-crosses genotyped with 614 AFLP marker, studied PS on the r_{MP} within- and across-groups, and found high accuracy estimates for within-group prediction.

Additive marker effects estimated across all three groups were not consistent for GY (Additional file 6: Figure S6B), and we observed that high variance of marker effects lead to higher r_{MP}. TECHNOW *et al.* (2014) also found differences in additive marker effects for GY in two maize germplasms. Besides that, LD pattern varied across all groups, showing greater LD estimates when sample size is smaller. YAN *et al.* (2009) showed the same tendency working with a diverse global maize collection.

4.5. CONCLUSION

Adding four different sets of PS as covariates to GBLUP increased reliability and reduced bias of prediction. In addition, depending on the number of hybrids available in the breeding program, apply MG-GBLUP in stratified groups could be an efficient method, maintaining the high relatedness and a considerable training population size. Further studies is required to increase accuracy in between-group hybrids, which is the main goal of maize breeding.

REFERENCES

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

Alexander, D. H., J. Novembre and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19**:** 1655-1664.

Aschard, H., B. J. Vilhjalmsson, A. D. Joshi, A. L. Price and P. Kraft, 2015 Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. American Journal of Human Genetics 96**:** 329-339.

Bernardo, R., and A. M. Thompson, 2016 Germplasm architecture revealed through chromosomal effects for quantitative traits in maize. The Plant Genome 9**:** 1-11.

Butler, D. G., B. R. Cullis, A. R. Gilmour and B. J. Gogel, 2009 ASReml-R reference manual. Department of Primary Industries, Queensland, Australia.

Campoy, J. A., E. Lerigoleur-Balsemin, H. Christmann, R. Beauvieux, N. Girollet *et al.*, 2016 Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of Prunus avium L. landraces and bred cultivars. BMC Plant Biol 16**:** 49.

Charcosset, A., and L. Essioux, 1994 The effect of population-structure on the relationship between heterosis and heterozygosity at marker loci. Theoretical and Applied Genetics 89**:** 336-343.

Chen, A. H., and A. E. Lipka, 2016 The use of targeted marker subsets to account for population structure and relatedness in genome-wide association studies of maize (*Zea mays* L.). G3: Genes|Genomes|Genetics 6**:** 2365-2374.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics 44**:** 803-U238.

Cros, D., M. Denis, L. Sanchez, B. Cochard, A. Flori *et al.*, 2015 Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 128**:** 397-410.

da Silva, T. A., L. B. Cantagalli, J. Saavedra, A. D. Lopes, C. A. Mangolin *et al.*, 2015 Population structure and genetic diversity of Brazilian popcorn germplasm inferred by microsatellite markers. Electronic Journal of Biotechnology 18**:** 181-187.

Da, Y., C. K. Wang, S. W. Wang and G. Hu, 2014 Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. Plos One 9**:** e87666.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf and B. J. Hayes, 2012 Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science 90**:** 3375-3384.

de los Campos, G., and P. Pérez-Rodriguez, 2015 Bayesian Generalized Linear Regression. R Package Version 1.0.4.

de los Campos, G., and D. Sorensen, 2014 On the genomic analysis of data from structured populations. J Anim Breed Genet 131**:** 163-164.

de los Campos, G., D. Sorensen and D. Gianola, 2015 Genomic heritability: what is it? Plos Genetics 11**:** e1005048.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. Plos Genetics 9.

Dos Santos, J. P., R. C. Vasconcellos, L. P. Pires, M. Balestre and R. G. Von Pinho, 2016 Inclusion of Dominance Effects in the Multivariate GBLUP Model. PLoS One 11**:** e0152045.

Endelman, J. B., 2015 Ridge regression and other kernels for genomic selection. rrBLUP-package Version: 4.4.

Evanno, G., S. Regnaut and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14**:** 2611-2620.

Fan, X. M., Y. Q. Bi, Y. D. Zhang, D. P. Jeffers, W. H. Yao *et al.*, 2015 Use of the suwan1 heterotic group in maize breeding programs in southwestern China. Agronomy Journal 107**:** 2353-2362.

Fernandes, E. H., I. Schuster, C. A. Scapim, E. S. N. Vieira and M. M. D. Coan, 2015 Genetic diversity in elite inbred lines of maize and its association with heterosis. Genetics and Molecular Research 14**:** 6509-6517.

Frichot, E., and O. Francois, 2015 LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6**:** 925-929.

Galinsky, K. J., G. Bhatia, P. R. Loh, S. Georgiev, S. Mukherjee *et al.*, 2016 Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am J Hum Genet 98**:** 456-472.

Gorjanc, G., P. Bijma and J. M. Hickey, 2015 Reliability of pedigree-based and genomic evaluations in selected populations. Genet Sel Evol 47**:** 65.

Guo, Z., D. M. Tucker, C. J. Basten, H. Gandhi, E. Ersoz *et al.*, 2014 The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127**:** 749-762.

Hallauer, A. R., and M. J. Carena, 2014 Adaptation of tropical maize germplasm to temperate environments. Euphytica 196**:** 1-11.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41.

He, S., A. W. Schulthess, V. Mirdita, Y. Zhao, V. Korzun *et al.*, 2016 Genomic selection in a commercial winter wheat population. Theor Appl Genet 129**:** 641-651.

Huang, M., A. Cabrera, A. Hoffstetter, C. Griffey, D. Van Sanford *et al.*, 2016 Genomic selection for wheat traits and trait stability. Theor Appl Genet 129**:** 1697-1710.

Iheshiulor, O. O. M., J. A. Woolliams, X. J. Yu, R. Wellmann and T. H. E. Meuwissen, 2016 Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. Genetics Selection Evolution 48.

Isidro, J., J. L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. Theor Appl Genet 128**:** 145-158.

Jan, H. U., A. Abbadi, S. Lucke, R. A. Nichols and R. J. Snowdon, 2016 Genomic prediction of testcross performance in canola (*Brassica napus*). PLoS One 11**:** e0147769.

Janss, L., G. de Los Campos, N. Sheehan and D. Sorensen, 2012 Inferences from genomic models in stratified populations. Genetics 192**:** 693-704.

Jombart, T., C. Collins, Z. N. Kamvar, R. Lustrik, P. Solymos *et al.*, 2015 adegenet: exploratory analysis of genetic and genomic data. R Package Version 2.0.1.

Jombart, T., S. Devillard and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11.

Lanes, E. C. M., J. M. S. Viana, G. P. Paes, M. F. B. Paula, C. Maia *et al.*, 2014 Population structure and genetic diversity of maize inbreds derived from tropical hybrids. Genetics and Molecular Research 13**:** 7365-7376.

Lawson, D. J., G. Hellenthal, S. Myers and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genet 8**:** e1002453.

Lehermeier, C., N. Kramer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of multiparental populations of maize *(Zea mays* L.) for genome-based prediction. Genetics 198**:** 3-16.

Lehermeier, C., C. C. Schon and G. de los Campos, 2015 Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics 201**:** 323–337.

Lorenz, A. J., and K. P. Smith, 2015 Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. Crop Science 55**:** 2657-2667.

Maenhout, S., B. De Baets and G. Haesaert, 2010 Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theoretical and Applied Genetics 120**:** 415-427.

Marulanda, J. J., X. Mi, A. E. Melchinger, J. L. Xu, T. Würschum *et al.*, 2016 Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. Theoretical and Applied Genetics**:** 1-13.

Massman, J. M., A. Gordillo, R. E. Lorenzana and R. Bernardo, 2013 Genomewide predictions from maize single-cross data. Theoretical and Applied Genetics 126**:** 13-22.

Melchinger, A. E., J. Boppenmaier, B. S. Dhillon, W. G. Pollmer and R. G. Herrmann, 1992 Genetic diversity for RFLPs in European Maize Inbreds. 2. Relation to performance of hybrids within versus between heterotic groups for forage traits. Theoretical and Applied Genetics 84**:** 672-681.

Mendes, M. P., and C. L. de Souza, 2016 Genomewide prediction of tropical maize single-crosses. Euphytica 209**:** 651-663.

Molin, D., C. J. Coelho, D. S. Maximo, F. S. Ferreira, J. R. Gardingo *et al.*, 2013 Genetic diversity in the germplasm of tropical maize landraces determined using molecular markers. Genetics and Molecular Research 12**:** 99-114.

Mundim, G. B., J. M. S. Viana, C. Maia, G. P. Paes, R. O. DeLima *et al.*, 2015 Inferring tropical popcorn gene pools based on molecular and phenotypic data. Euphytica 202**:** 55-68.

Nelson, P. T., M. D. Krakowsky, N. D. Coles, J. B. Holland, D. M. Bubeck *et al.*, 2016 Genetic characterization of the North Carolina State University maize lines. Crop Science 56**:** 259-275.

Newell, M. A., and J. L. Jannink, 2014 Genomic selection in plant breeding. Crop Breeding: Methods and Protocols**:** 117-130.

Orozco-Ramirez, Q., J. Ross-Ibarra, A. Santacruz-Varela and S. Brush, 2016 Maize diversity associated with social origin and environmental variation in Southern Mexico. Heredity 116**:** 477-484.

Oyekunle, M., B. Badu-Apraku, S. Hearne and J. Franco, 2015 Genetic diversity of tropical early-maturing maize inbreds and their performance in hybrid combinations under drought and optimum growing conditions. Field Crops Research 170**:** 55-65.

Patterson, N., A. L. Price and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet 2**:** e190.

Perez, P., and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics 198**:** 483-U463.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38**:** 904-909.

Price, A. L., N. A. Zaitlen, D. Reich and N. Patterson, 2010 New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11**:** 459-463.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81**:** 559-575.

Reif, J. C., A. E. Melchinger, X. C. Xia, M. L. Warburton, D. A. Hoisington *et al.*, 2003 Use of SSRs for establishing heterotic groups in subtropical maize. Theoretical and Applied Genetics 107**:** 947-957.

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

Riedelsheimer, C., J. B. Endelman, M. Stange, M. E. Sorrells, J. L. Jannink *et al.*, 2013 Genomic predictability of interconnected biparental maize populations. Genetics 194**:** 493-503.

Rincent, R., S. Nicolas, S. Bouchet, T. Altmann, D. Brunel *et al.*, 2014 Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production. Theor Appl Genet 127**:** 2313-2331.

Roberts, D. W., 2016 labdsv: Ordination and Multivariate Analysis for Ecology. R Package Version 1.6-1.

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim *et al.*, 2011 Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43**:** 1-16.

Schaefer, C. M., and R. Bernardo, 2013 Population structure and single nucleotide polymorphism diversity of historical Minnesota maize inbreds. Crop Science 53**:** 1529.

Schulz-Streeck, T., J. O. Ogutu, Z. Karaman, C. Knaak and H. P. Piepho, 2012 Genomic selection using multiple populations. Crop Science 52**:** 2453-2461.

Shriner, D., and N. J. Yi, 2009 Deviance information criterion (DIC) in Bayesian multiple QTL mapping. Computational Statistics & Data Analysis 53**:** 1850-1860.

Sousa, M. B. E., J. Cuevas, E. G. D. Couto, P. Perez-Rodriguez, D. Jarquin *et al.*, 2017 Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype x Environment Interaction. G3-Genes Genomes Genetics 7**:** 1995-2014.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard *et al.*, 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 Genetics 11.

Sukumaran, S., S. Dreisigacker, M. Lopes, P. Chavez and M. P. Reynolds, 2015 Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. Theoretical and Applied Genetics 128**:** 353-363.

Sukumaran, S., W. Xiang, S. R. Bean, J. F. Pedersen, S. Kresovich *et al.*, 2012 Association mapping for grain quality in a diverse sorghum collection. The Plant Genome Journal 5**:** 126-135.

Technow, F., C. Riedelsheimer, T. A. Schrag and A. E. Melchinger, 2012 Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125**:** 1181-1194.

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197**:** 1343-1355.

Teixeira, J. E. C., T. Weldekidan, N. de Leon, S. Flint-Garcia, J. B. Holland *et al.*, 2015 Hallauer's Tuson: a decade of selection for tropical-to-temperate phenological adaptation in maize. Heredity 114**:** 229-240.

Thomasen, J. R., A. C. Sorensen, G. Su, P. Madsen, M. S. Lund *et al.*, 2013 The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. Journal of Animal Science 91**:** 3105-3112.

Tucker, G., A. L. Price and B. Berger, 2014 Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. Genetics 197**:** 1045-1049.

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak *et al.*, 2014 A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics. 15**:** 1-15.

Unterseer, S., S. D. Pophaly, R. Peis, P. Westermeier, M. Mayer *et al.*, 2016 A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. Genome Biology 17.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J Dairy Sci 91**:** 4414-4423.

Ventura, R., S. Larmer, F. S. Schenkel, S. P. Miller and P. Sullivan, 2016 Genomic clustering helps to improve prediction in a multibreed population. Journal of Animal Science 94**:** 1844-1856.

Wang, M. L., S. Sukumaran, N. A. Barkley, Z. B. Chen, C. Y. Chen *et al.*, 2011 Population structure and marker-trait association analysis of the US peanut (*Arachis hypogaea* L.) mini-core collection. Theoretical and Applied Genetics 123**:** 1307-1317.

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J. L. Jannink *et al.*, 2012 Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3-Genes Genomes Genetics 2**:** 1427-1436.

Wu, Y. S., F. S. Vicente, K. J. Huang, T. Dhliwayo, D. E. Costich *et al.*, 2016 Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. Theoretical and Applied Genetics 129**:** 753-765.

Yan, J. B., T. Shah, M. L. Warburton, E. S. Buckler, M. D. McMullen *et al.*, 2009 Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. Plos One 4.

Yang, X. H., S. B. Gao, S. T. Xu, Z. X. Zhang, B. M. Prasanna *et al.*, 2011 Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Molecular Breeding 28**:** 511-526.

Yu, J. M., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38**:** 203-208.

Zhang, J., Q. Song, P. B. Cregan and G. L. Jiang, 2016 Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). Theor Appl Genet 129**:** 117-130.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie *et al.*, 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28**:** 3326-3328.

Zhu, C., and J. Yu, 2009 Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics 182**:** 875-888.

FIGURES

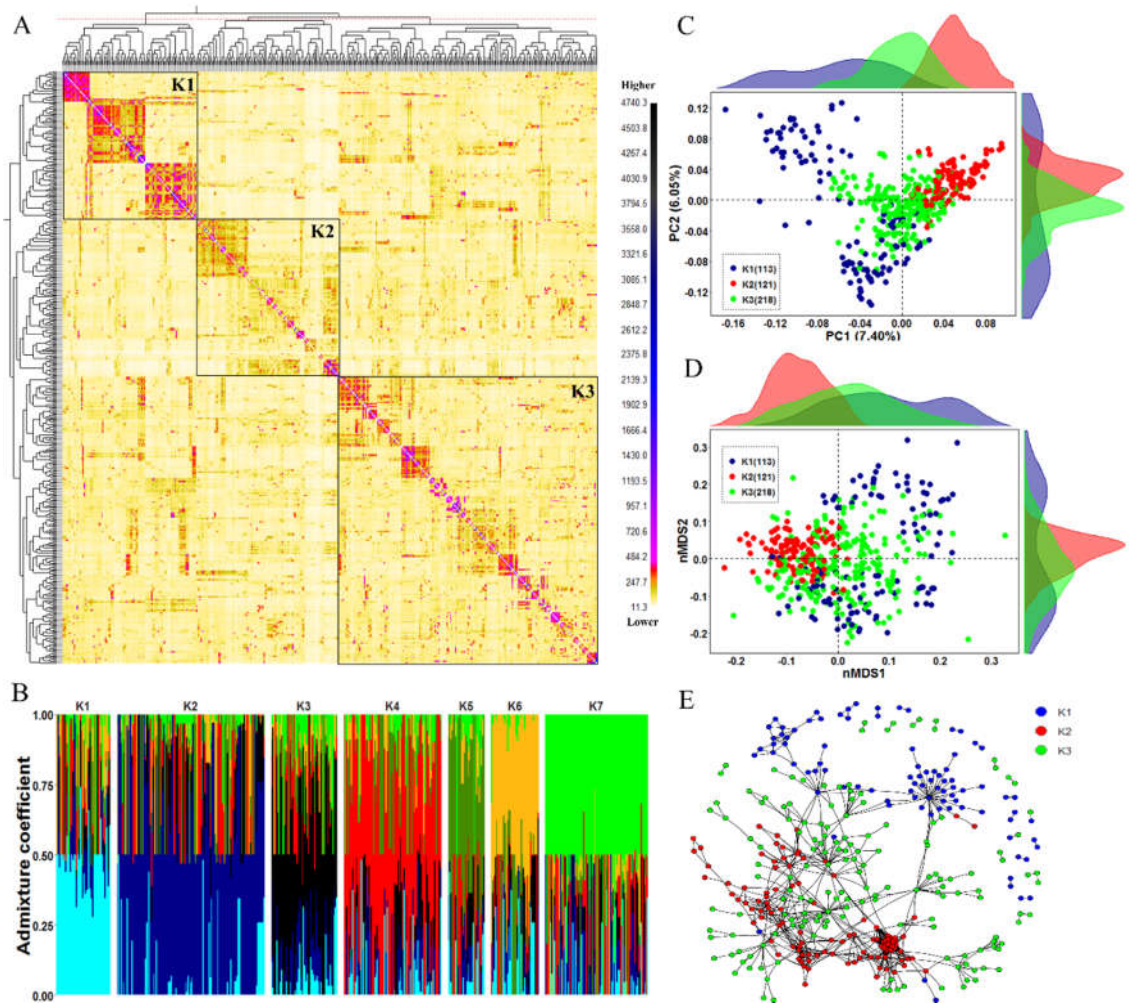


Figure 1. Population structure analysis in 452 tropical maize single-cross hybrids. (A) Coancestry heatmap of fineSTRUCTURE unlinked model. Scale shows lower (white) to higher (black) amount of shared genetic chunks between the individuals. On the left and top is the maximum a posteriori (MAP) tree. Dashed red line is the cutoff threshold splitting K1, K2, and K3 groups. B) Artificial admixture coefficients, where each colour represents a group (K1-K7). C) First two principal components, applied to raw SNP data (32,838 SNPs). The percentages in parentheses in the axis titles represent the variance explained by each of the two principal components. D) First two nonmetric multidimensional scaling (nMDS) dimensions, applied to Roger's distance matrix. E) Network representation of the GRM, where individuals were linked when their relationship coefficient was ≥0.6 (not all hybrids are shown). Colors in B, C, and E indicate three groups clustered from fineSTRUCTURE results. Number of hybrids per group is indicated in parenthesis. Density plot shows the distribution of individuals in each group.

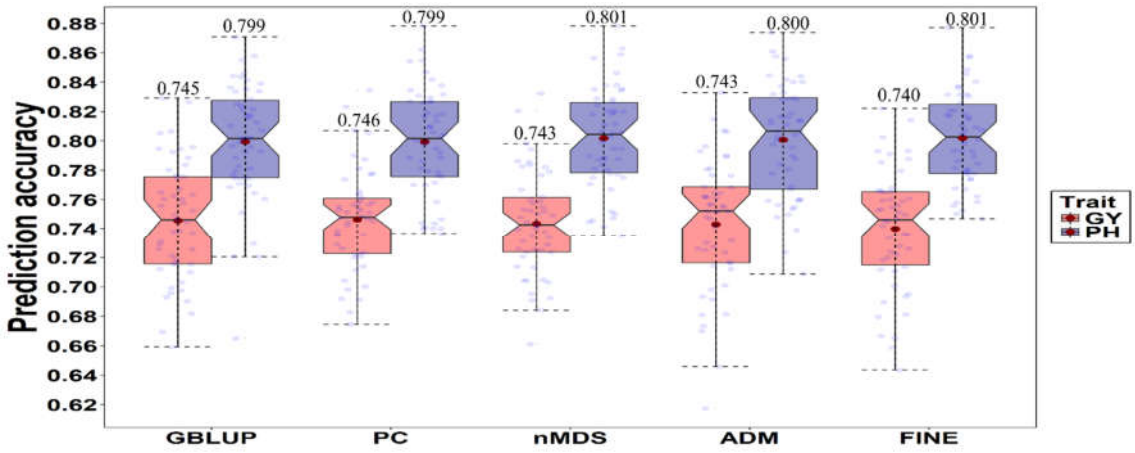


Figure 2. Comparison of prediction accuracy. Boxplots of prediction accuracy for grain yield (GY) and plant height (PH), from GBLUP and four fixed covariates: principal components (PC), nonmetric multidimensional scaling dimensions (nMDS), admixture coefficients (ADM), and fineSTRUCTURE group clustering (FINE). Red dots and numbers above the dashed line are representing the mean for each distribution.

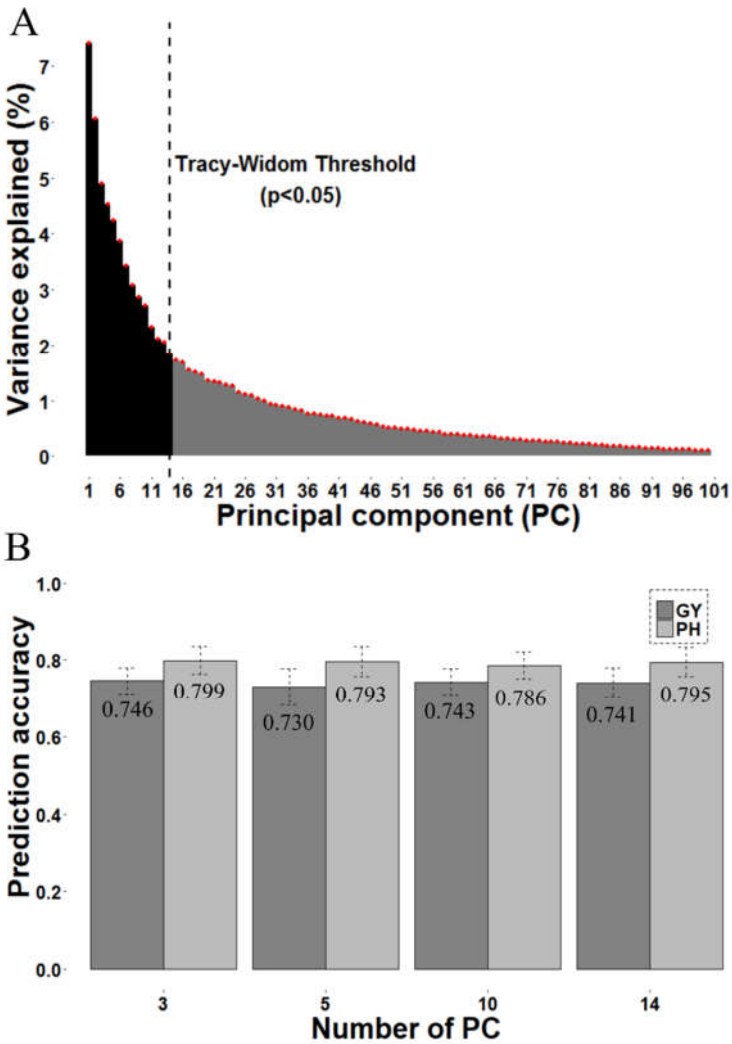


Figure 3. Top principal components and comparison of prediction accuracy. (A) Percentage of variance explained by the principal components. The number of statistically significant ($p<0.05$) principal components, measured by the Tracy-Widom statistic, is shown in the black region. (B) Barplot (mean±SD) of prediction accuracy from GBLUP with 3, 5, 10, and 14 PCs as fixed covariates.

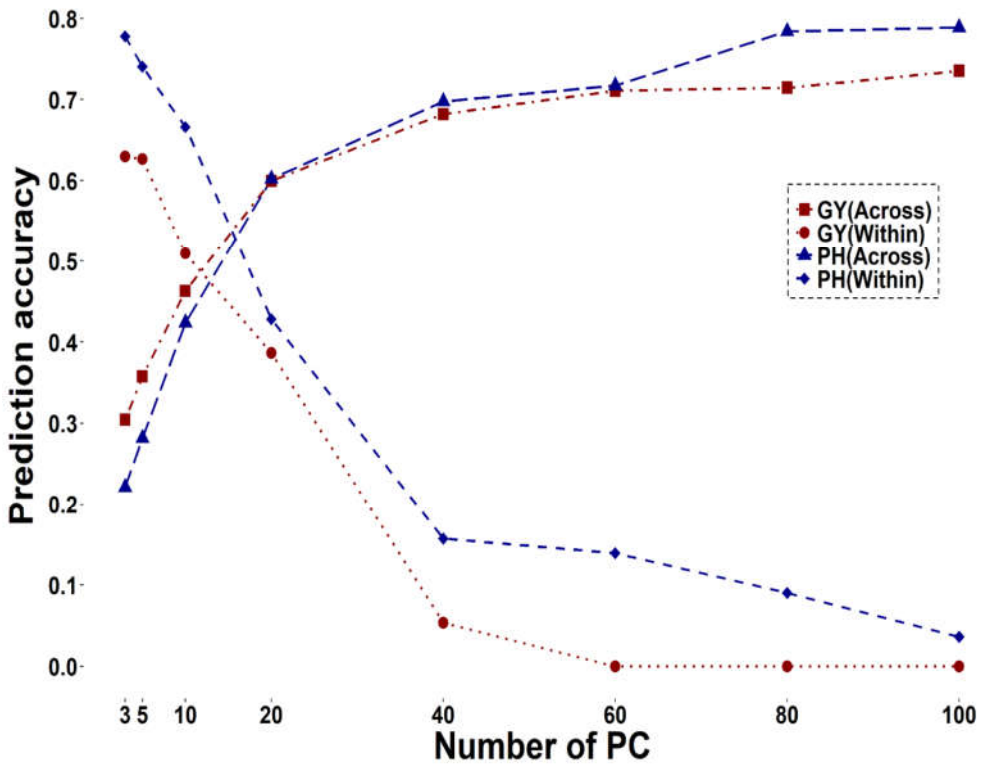


Figure 4. Relationship between estimates of prediction accuracy and number of principal components (PC). Posterior mean of prediction accuracy across- and within-group after accounting for the proportion of variance due to the *d* eigenvectors.

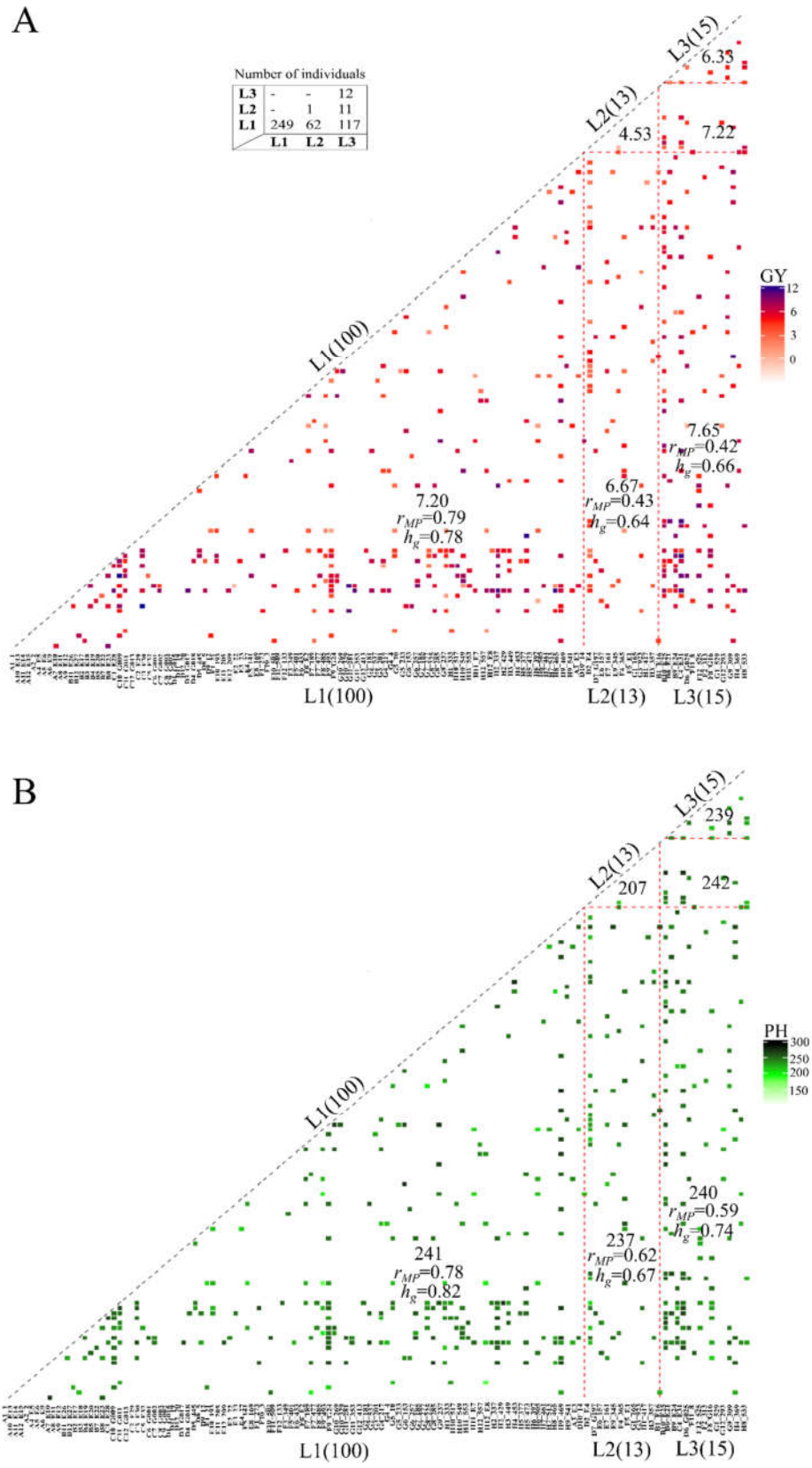
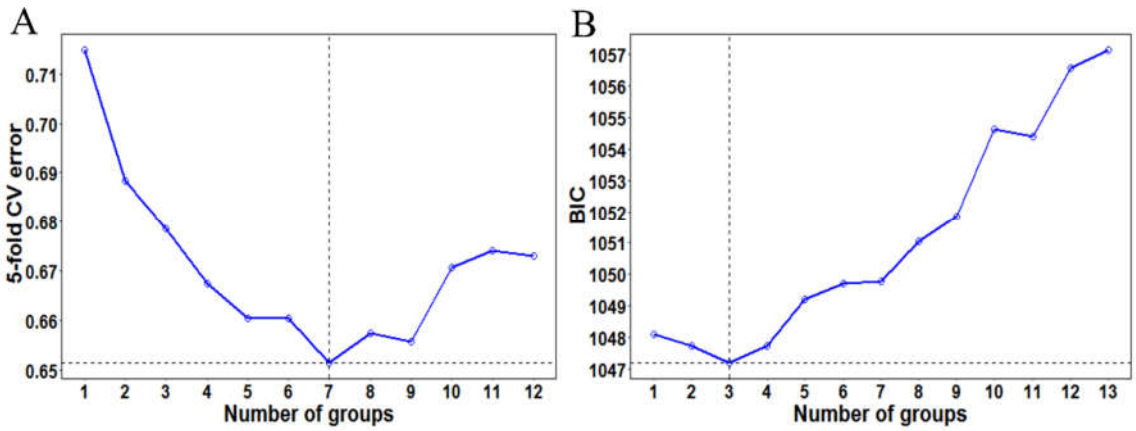
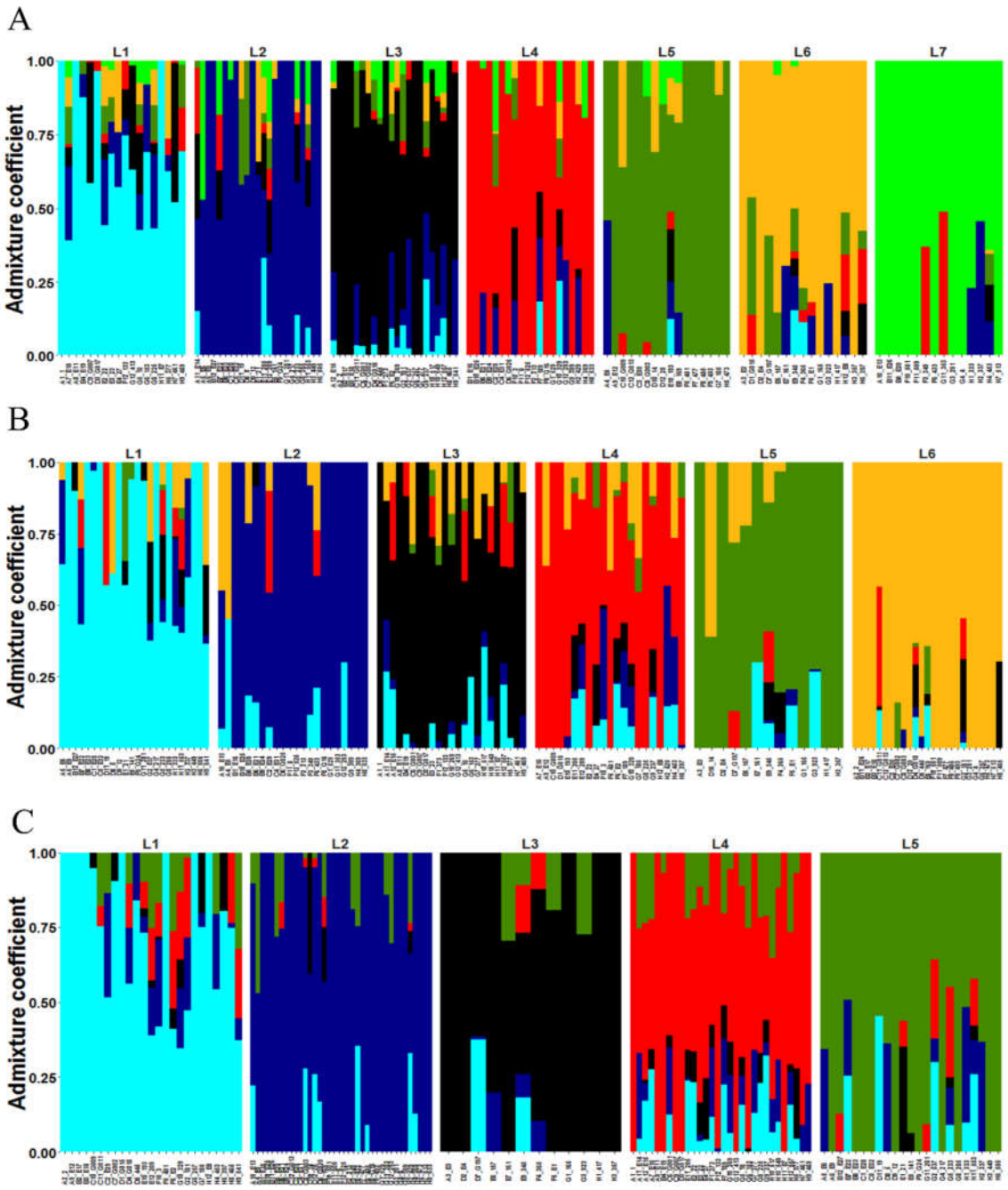

A
Number of individuals
L3 - - 12
L2 - 1 11
L1 249 62 117
L1 L2 L3
L3(15)
L2(13)
L1(100)
6.33
4.53
7.22
GY
12
6
3
0
7.65
r_{MP}=0.42
h_g=0.66
6.67
r_{MP}=0.43
h_g=0.64
7.20
r_{MP}=0.79
h_g=0.78
L1(100)
L2(13)
L3(15)
B
L3(15)
L2(13)
L1(100)
239
207
242
PH
300
250
200
150
240
r_{MP}=0.59
h_g=0.74
237
r_{MP}=0.62
h_g=0.67
241
r_{MP}=0.78
h_g=0.82
L1(100)
L2(13)
L3(15)

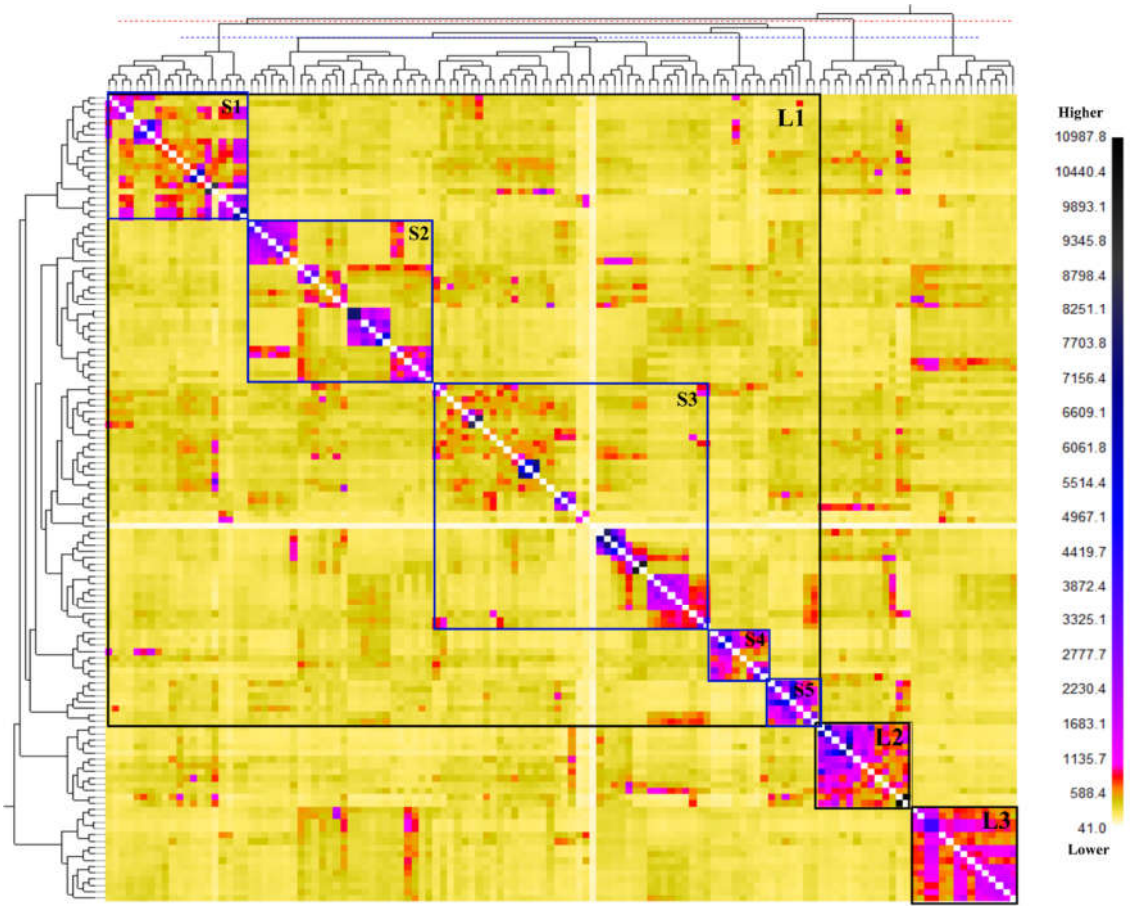
Figure 5. Within- and between-group hybrids from a partial diallel cross mating scheme of 128 inbred lines. The lines are grouped in L1, L2, and L3 based on fineSTRUCTURE results. Inside the plot is shown the trait mean, prediction accuracy (r_{MP}), and broad sense genomic heritability (h_g) estimated in traditional GBLUP for GY (A) and PH (B). Top left diagram is the number of individuals in each group.



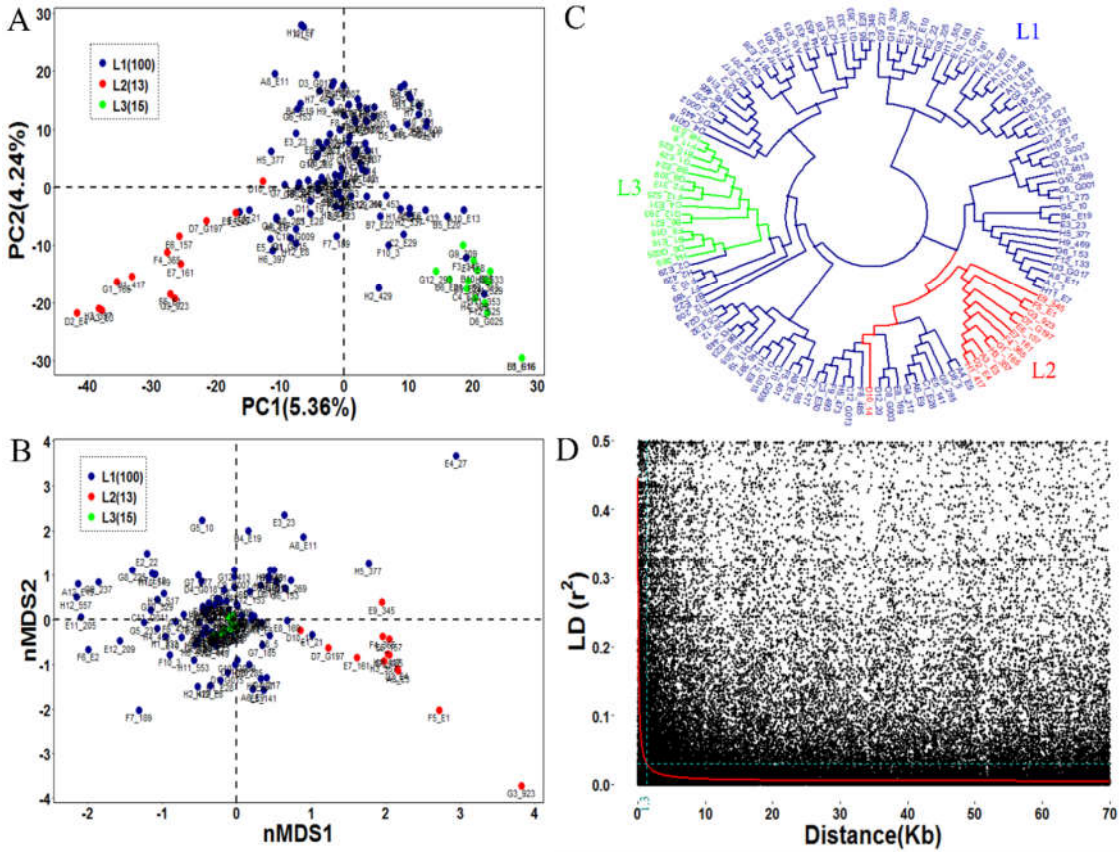
Additional file 1: Figure S1. Inference of group number of 128 tropical maize inbred lines. (A) 5-fold cross-validation error of ADMIXTURE, and (B) BIC values of k-means clustering. The dashed black line shows the number of groups inferred in each method.



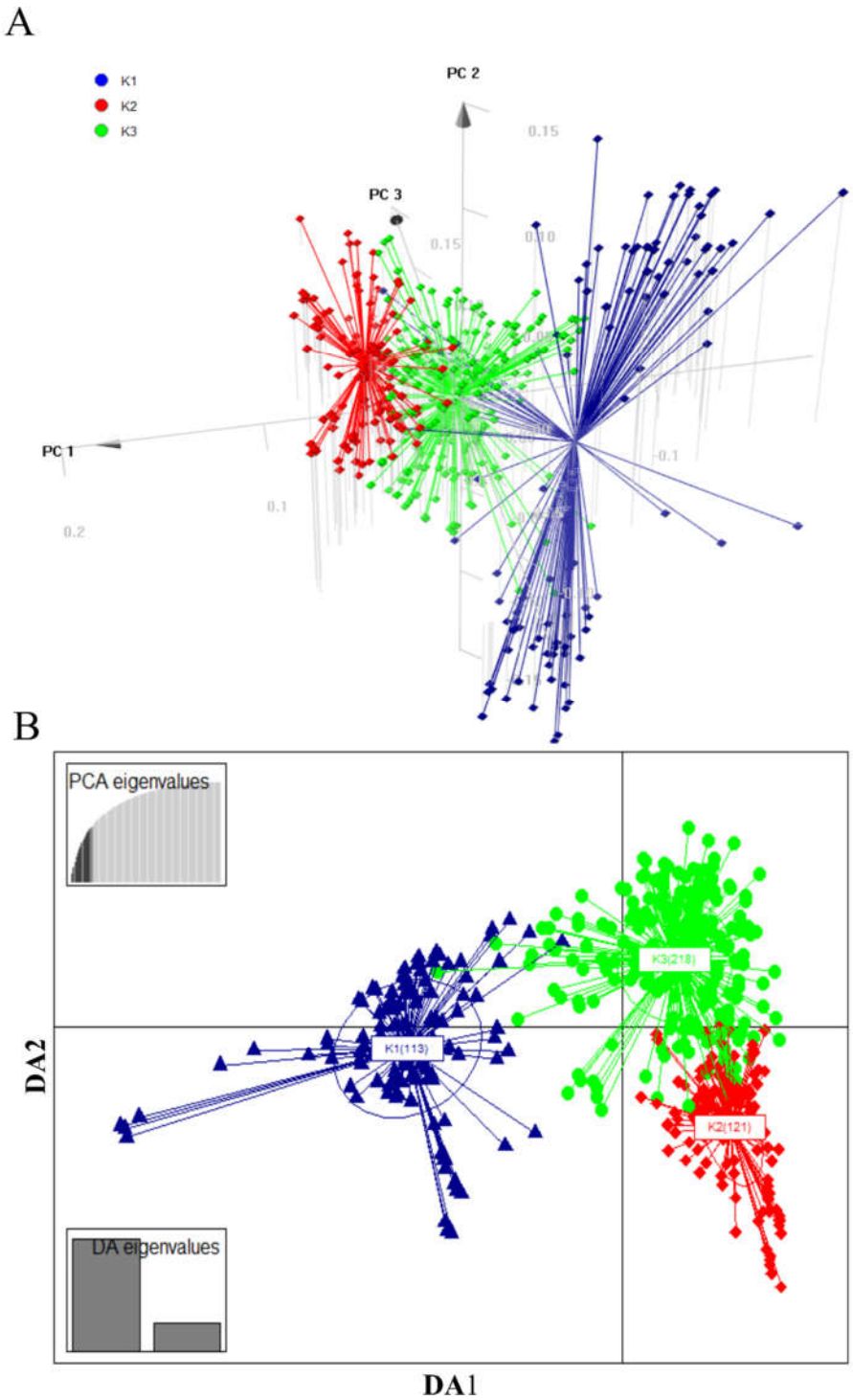
Additional file 2: Figure S2. ADMIXTURE results of 128 tropical maize inbred lines. Clustering assignments inferred in L7 (A), L6 (B), and L5 (C) groups. Each individual is represented by a single vertical line divided into L colored segments. White color separates the groups (L).



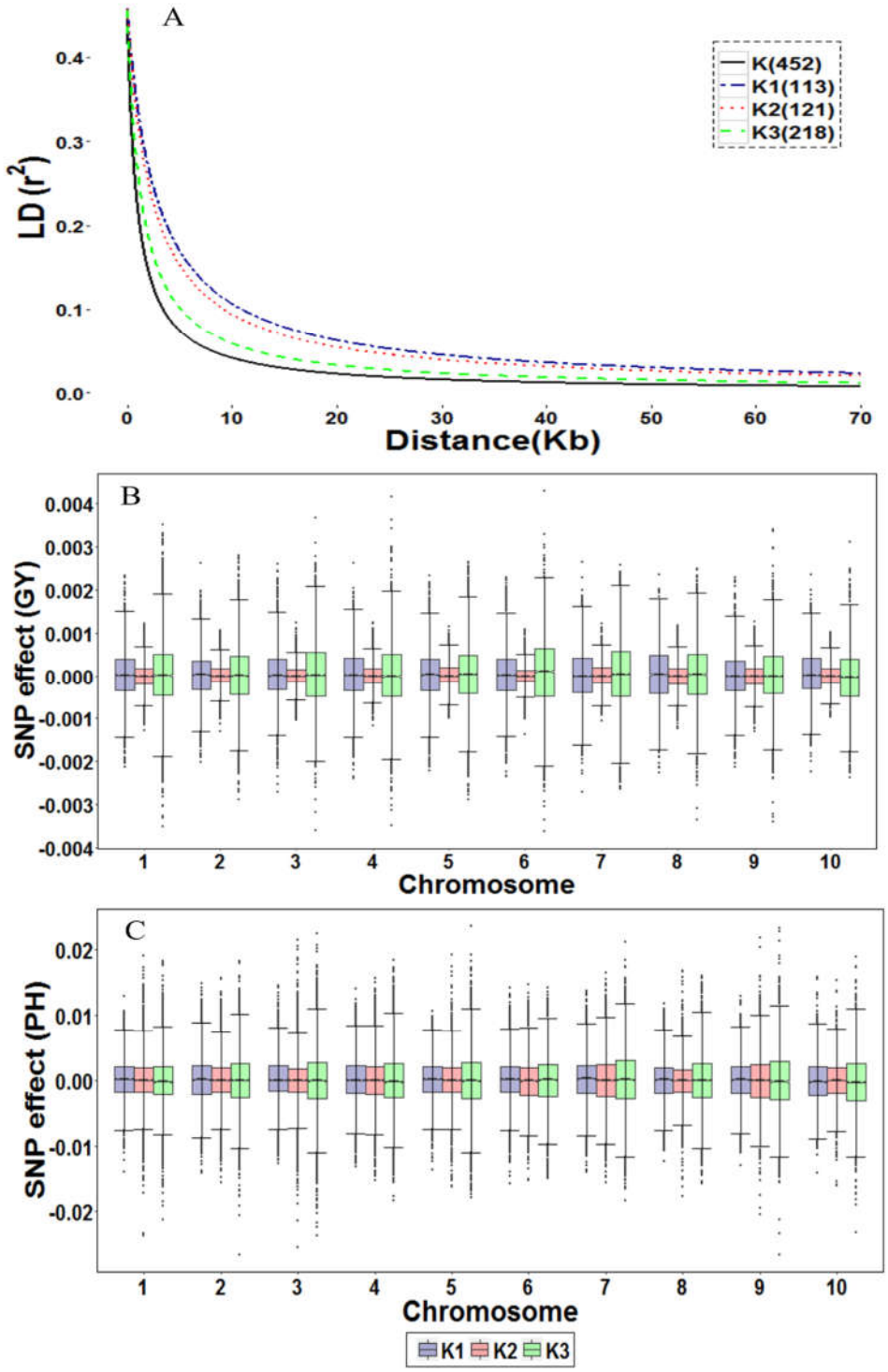
Additional file 3: Figure S3. Coancestry heatmap of fineSTRUCTURE unlinked model. Scale shows lower (white) to higher (black) amount of shared genetic chunks between the inbred lines. On the left and top is the maximum *a posteriori* (MAP) tree. Dashed red line is the cutoff threshold splitting L1, L2, and L3 groups. Dashed blue line clustered the subgroup S1, S2, S3, S4, and S5.



Additional file 4: Figure S4. A) First two (A) principal components, (B) dimensions of nMDS, (C) Neighbor-Joining Tree based on Modified Rogers' distance, and (D) pattern of linkage disequilibrium (LD) within 70 kb of distance among all pairs of marker (32K) for 128 tropical maize inbreds, coloured by fineSTRUCTURE group-clustering.



Additional file 5: Figure S5. A) 3-D PCA score plot for the first three principal components. B) First two principal components of the Discriminant Analysis of Principal Components (DAPC).



Additional file 6: Figure S6. (A) Pattern of linkage disequilibrium (LD) within 70 kb of distance among all pairs of marker (26K SNPs) for 452(K), 113(K1), 121(K2), and 218(K3) individuals. Values reported are the average r^2 across 10 chromosomes. Boxplots of additive marker effect estimates for (B) GY and (C) PH, obtained from rrBLUP-R, for K1, K2, and K3 groups.

TABLES

Table 1. Genetic parameter for grain yield (GY) and plant height (PH).

Model	Grain yield						Plant height					
	σ_a^2	σ_d^2	σ_ε^2	h_g^2	REL	DIC	σ_a^2	σ_d^2	σ_ε^2	h_g^2	REL	DIC
GBLUP	0.28±0.02[a]	0.24±0.01	0.13±0.00	0.79±0.04	0.54±0.03	474±22	212±14.7	38.3±8.82	40.8±3.93	0.86±0.04	0.83±0.01	2392±30
PC[b]	0.26±0.02	0.24±0.01	0.13±0.01	0.77±0.02	0.51±0.04	485±23	214±14.2	35.1±2.94	41.7±5.72	0.84±0.02	0.86±0.00	2392±42
nMDS[b]	0.29±0.02	0.23±0.01	0.13±0.01	0.78±0.04	0.59±0.03	484±26	211±13.2	33.8±3.00	42.2±5.50	0.84±0.03	0.83±0.03	2395±41
ADM[b]	0.25±0.01	0.24±0.01	0.13±0.01	0.77±0.05	0.45±0.03	485±25	221±16.2	34.8±3.14	41.3±4.88	0.85±0.04	0.87±0.00	2390±36
FINE[b]	0.27±0.02	0.23±0.01	0.13±0.00	0.79±0.02	0.54±0.04	473±20	210±14.7	34.2±1.87	42.1±4.68	0.85±0.04	0.89±0.03	2395±33

Additive variance (σ_a^2), dominance variance (σ_d^2), error variance (σ_ε^2), broad sense genomic heritability (h_g^2), reliability (REL), and deviance information criterion (DIC).

[a]Data are posterior mean ± SD estimated from fifty replications in independent validation.

[b]Variables used as fixed covariates in GBLUP model.

Table 2 Within- (W-GBLUP) and multi-group (MG-GBLUP) analysis for K1, K2, and K3 groups.

Model	Pop	Grain yield					Plant height				
		$\sigma^2_{a_k}$	$\sigma^2_{d_k}$	$\sigma^2_{\varepsilon_k}$	$h^2_{g_k}$	r_{MP}	$\sigma^2_{a_k}$	$\sigma^2_{d_k}$	$\sigma^2_{\varepsilon_k}$	$h^2_{g_k}$	r_{MP}
W-GBLUP	K_1	0.30±0.05[a]	0.17±0.02	0.21±0.02	0.68±0.10	0.58±0.11	131±17.1	55.6±5.41	70.12±6.43	0.71±0.09	0.64±0.09
	K_2	0.20±0.01	0.31±0.04	0.18±0.01	0.75±0.04	0.60±0.14	155±14.1	45.5±5.13	47.1±4.03	0.82±0.04	0.71±0.05
	K_3	0.38±0.03	0.20±0.01	0.14±0.00	0.79±0.05	0.73±0.05	165±11.2	37.8±4.31	51.0±7.17	0.80±0.06	0.78±0.05
MG-GBLUP	K_1	0.45±0.03	0.24±0.01	0.21±0.01	0.76±0.01	0.63±0.12	215±35.3	50.6±17.2	31.6±15.9	0.88±0.06	0.72±0.06
	K_2	0.36±0.02	0.43±0.03	0.21±0.01	0.78±0.00	0.68±0.09	274±36.3	62.7±8.07	8.43±1.29	0.97±0.00	0.78±0.07
	K_3	0.44±0.03	0.24±0.01	0.16±0.00	0.80±0.00	0.77±0.05	221±20.4	27.1±7.27	42.3±11.6	0.85±0.03	0.84±0.04

Additive variance ($\sigma^2_{a_k}$), dominance variance ($\sigma^2_{d_k}$), error variance ($\sigma^2_{\varepsilon_k}$), broad sense genomic heritability ($h^2_{g_k}$), and prediction accuracy (r_{MP}) for grain yield (GY) and plant height (PH).

[a]Data are posterior mean ± SD estimated from fifty replications in independent validation.

5. GENERAL CONCLUSION

Multi-trait genomic prediction using the combination of selection indices is an excellent strategy to increase accuracy of selection under abiotic stress conditions.

Including dominance deviation and copy number variation effects in single-crosses prediction for complex traits seems to be suitable, due to the increase of accuracies and reduction of model bias.

Controlling population structure in genomic prediction models may increase reliability and precision of estimation of genomic breeding values in maize hybrids.