FEDERAL UNIVERSITY OF CEARÁ

AGRICULTURAL SCIENCES CENTER

DEPARTMENT OF PLANT SCIENCE

POSTGRADUATE COURSE IN AGRONOMY/PLANT SCIENCE

INGRID PINHEIRO MACHADO

ON THE USEFULNESS OF MOCK GENOMES TO DEFINE HETEROTIC POOLS, TESTERS, AND HYBRID PREDICTIONS IN ORPHAN CROPS

FORTALEZA

2022

INGRID PINHEIRO MACHADO

ON THE USEFULNESS OF MOCK GENOMES TO DEFINE HETEROTIC POOLS, TESTERS, AND HYBRID PREDICTIONS IN ORPHAN CROPS

Thesis presented to Postgraduate program in Agronomy/Plant Science of the Federal University of Ceará, as part of the requirements for obtaining the title of *Doctor Scientiae* in Agronomy/Plant Science. Concentration area: Plant Science. Research Line: Genetics and Plant Breeding.

Advisor: Prof. D.Sc. Júlio César do Vale Silva
Co-advisor: Prof. D.Sc Roberto Fritsche Neto

FORTALEZA
2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M131o Machado, Ingrid Pinheiro.
On the usefulness of mock genomes to define heterotic pools, testers, and hybrid predictions in orphan crops / Ingrid Pinheiro Machado. – 2022.
65 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências Agrárias, Programa de Pós-Graduação em Agronomia (Fitotecnia), Fortaleza, 2022.
Orientação: Prof. Dr. Júlio César do Vale Silva.
Coorientação: Prof. Dr. Roberto Fritsche Neto.

1. GBS. 2. SNP-array. 3. Minor crops. 4. Simulated genome. 5. Formation of heterotic groups. I. Título.
CDD 630

INGRID PINHEIRO MACHADO

ON THE USEFULNESS OF MOCK GENOMES TO DEFINE HETEROTIC POOLS, TESTERS, AND HYBRID PREDICTIONS IN ORPHAN CROPS

Thesis presented to Postgraduate program in Agronomy/Plant Science of the Federal University of Ceará, as part of the requirements for obtaining the title of *Doctor Scientiae* in Agronomy/Plant Science. Concentration area: Plant Science. Research Line: Genetics and Plant Breeding.

Approved in: 27/10/2022.

EXAMINATION BOARD

Prof. D. Sc. Júlio César do Vale Silva (Advisor)
Federal University of Ceará (UFC)

Prof. D. Sc Roberto Fritsche Neto (Co-advisor)
Louisiana State University (LSU)

D. Sc José Airton Rodrigues Nunes (Adviser)
Universidade Federal de Lavras (UFLA)

D. Sc Rafael Massahiro Yassue (Adviser)
GDM seeds

D. Sc Karina Lima Reis Borges (Adviser)
University of Florida (UF)

First, I thank God for giving me strength and courage throughout this journey. To my parents and my husband.

ACKNOWLEDGMENTS

To God, for always being present in my life, taking care and giving me strength to overcome all obstacles, even when I thought I couldn't. Thank you for being my source of love, wisdom, and protection.

To the Federal University of Ceará, especially the Postgraduate Program in Agronomy/Plant Science, for the opportunity to carry out this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, for financial support through the granting of the scholarship.

To the Allogamous Plant Breeding Laboratory (USP/Esalq) for welcoming me during the academic mobility period and for the opportunity to learn from such qualified professionals.

To Limagrain, for genotyping the study material.

To my parents, Jonas Machado Araújo and Antônia Ivonilde Pinheiro Pereira, who encouraged and supported me throughout this period.

To my grandmother, Maria Pinheiro Pereira, for her example of strength and resilience.

To my husband, Irisval Leite, for his love, collaboration, help, support, dedication, and patience. Without his support, it would have been infinitely more complex.

To Professor Júlio César do Vale Silva, for his excellent guidance since graduation, his contributions, the knowledge transmitted to me, his patience, and the professional example he is for me.

To the co-advisor, Professor Roberto Fritsche Neto, for having accepted and welcomed me in his laboratory, where I met many talented colleagues and had essential lessons, for the teachings passed on to me, not only scientific but also for the commitment and pleasure of education.

To Dr. Felipe Gonzaga Sabadin, for his significant contribution to the construction of this work, for his patience in passing on his knowledge, for his willingness to always help, and for having accepted to participate in the examination board.

To Drs. Rafael Massahiro Yassue and Karina Lima Reis Borges for their availability to participate in the examination board and contribute.

To my colleagues at the Allogamous Plant Breeding Laboratory for their welcome, especially Gabriela, Germano, Karina, Pedro, Rafael, Sabadin, Raísa, Giovanni, and Miguel.

To colleagues in the improvement group coordinated by Prof. Júlio César, Fernanda, Valnice, Lucas, and Anderson.

To my friends in the doctoral room, Johny, Arnaldo, Charles, Beatriz, and Rafael, for the coffees, conversations, and moments of relaxation. Especially to my friend, Fernanda Carla, who was with me in every moment of joy and despair, for always being willing to help when I needed it, for her advice and companionship.

To my friend Jessica, a sister I gained in postgraduate, thank you for always having a nice word for me and for your advice, care, and companionship.

To my friend Camila, who welcomed me when I needed it most, you were a gift from God in my life. For being this genuinely good person who cares, welcomes and is always willing to help. I learned a lot with you. Thank you so much.

To my friends, Cecília and Tamiris, who are my personal and professional references. For always being willing to advise and welcome me.

To Mr. Bezerra for all the care during the period of field experiments and for his willingness to always help. And to Vieira, for his friendship and respect since graduation, for always being willing to help.

I thank everyone who marked my path somehow; it contributed to the professional I became today.

ABSTRACT

The breeding programs cross-pollinated crops develop thousands of lines that, when combined, generate single-crosses that need to be evaluated for their performance in different sites, making this step the most expensive in released new cultivars. The molecular markers have proved to be a powerful tool in improving economically essential crops to accelerate this process. Currently, there are several genotyping platforms capable of providing thousands of SNP (Single Nucleotide Polymorphism) for performing genomic studies. However, the adoption of modern genomic enhancement for crops that do not yet have a reference genome is limited. Genotyping by sequencing (GBS) has emerged as an alternative to make such technologies viable for orphan crops. Once with these data, it is possible to build a simulated genome to perform the SNP calling where the discovery of polymorphisms will be intrinsic to the population under study without using an external genome. The term "orphan" is derived from the condition of neglect and helplessness of these crops by the scientific community, despite having great food and nutritional potential. Therefore, our goals were to verify whether the source of SNP can influence the assessment of the population structure of parental lines; ascertain if the SNP source can affect the determination of heterotic groups and the prediction of single-crosses performance, and to test if using GBS and the mock genome efficiently performs the SNP calling in orphan crops, the ones that don't have reference genome available. For this, maize was used as a model species, where 330 parental lines were genotyped by two standard genotyping platforms, SNP-array and GBS. GBS data were used for two purposes, to perform the SNP calling using the parental line B73 (GBS-B73) as a reference genome and to build a mock genome (GBS-Mock) to perform the SNP calling without needing an external genome, making three genotyping scenarios: SNP-array, GBS-B73, and GBS-Mock. These scenarios were used to conduct studies of population structure and genetic diversity among parental lines. After, we used phenotypic data of 751 single-crosses generated from the diallel of these parental lines. From there, genomic diallel analyses were performed to separate parental lines into heterotic groups and choose the best testers. Subsequently, an additive-dominant model was applied to predict the performance of single-crosses. The results showed that the GBS-Mock presented similar results to the standard population structure studies approach. The genotyping scenarios also did not differ in the division of heterotic groups and the definition of testers. In the genomic prediction study, GBS-Mock performed similarly to the SNP-array and GBS-B73. These results showed that a mock genome constructed from the population's intrinsic polymorphisms to perform the SNP calling is an excellent strategy to support plant breeders in studies of diversity, population structure, the definition of heterotic groups, choice of testers, and genomic prediction in species that still do not have a reference genome available. Because it is an alternative to the rapid advance of orphan crop breeding, in this context, genotyping via GBS associated with the mock genome is an effective alternative for performing genomic studies in orphan crops, especially those that do not have a reference genome.

Keywords: GBS, SNP-*array*; formation of heterotic groups; genomic prediction of single-crosses; minor crops; underused crops; simulated genome.

RESUMO

Os programas de melhoramento de culturas de polinização cruzada desenvolvem milhares de linhagens que, quando combinadas, geram cruzamentos-simples que precisam ser avaliados quanto ao seu desempenho em diferentes locais, tornando esta etapa a mais cara no lançamento de novas cultivares. Os marcadores moleculares têm se mostrado uma poderosa ferramenta no aprimoramento de culturas economicamente essenciais para acelerar esse processo. Atualmente, existem várias plataformas de genotipagem capazes de fornecer milhares de SNP (*Single Nucleotide Polymorphism*) para a realização de estudos genômicos. No entanto, a adoção de aprimoramento genômico moderno para culturas que ainda não possuem um genoma de referência é limitada. A genotipagem por sequenciamento (GBS) surgiu como uma alternativa para viabilizar tais tecnologias para culturas órfãs. Uma vez que, com esses dados, é possível construir um genoma simulado para realizar a chamada de SNP, onde a descoberta de polimorfismos será intrínseca à população de estudo sem utilizar um genoma externo. O termo "órfão" é derivado da condição de abandono e desamparo dessas culturas pela comunidade científica, apesar de possuírem grande potencial alimentar e nutricional. Portanto, nossos objetivos foram verificar se a fonte de SNP pode influenciar na avaliação da estrutura populacional de linhagens parentais; verificar se a fonte de SNP pode afetar a determinação de grupos heteróticos e a predição do desempenho de cruzamentos-simples, e testar se o uso de GBS e o genoma simulado realizam eficientemente a chamada de SNP em culturas órfãs, aquelas que não possuem genoma de referência disponível .Para isso, o milho foi utilizado como espécie modelo, onde 330 linhagens parentais foram genotipadas por duas plataformas de genotipagem padrão, SNP-*array* e GBS. Os dados do GBS foram usados para dois propósitos, para realizar a chamada de SNP usando a linha parental B73 (GBS-B73) como genoma de referência e para construir um genoma simulado (GBS-Mock) para realizar a chamada de SNP sem a necessidade de um genoma externo, compondo três cenários de genotipagem: SNP-*array*, GBS-B73 e GBS-*Mock*. Esses cenários foram usados para realizar estudos de estrutura populacional e diversidade genética entre linhagens parentais. Posteriormente, utilizamos dados fenotípicos de 751 cruzamentos-simples gerados a partir de um dialelo entre essas linhagens parentais. A partir daí, foram realizadas análises dialélicas genômicas para separar as linhagens parentais em grupos heteróticos e escolher os melhores testadores. Posteriormente, um modelo aditivo-dominante foi aplicado para prever o desempenho dos cruzamentos-simples. Os resultados mostraram que o GBS-*Mock* apresentou resultados de estudos de estrutura populacional semelhantes às abordagens padrão. Os cenários de genotipagem também não diferiram na divisão dos grupos heteróticos e na definição dos testadores. No estudo de predição genômica, GBS-*Mock* teve um desempenho semelhante a SNP-*array* e a GBS-B73. Esses resultados mostraram que um genoma simulado construído a partir de polimorfismos intrínsecos da população para realizar a chamada de SNP é uma excelente estratégia para apoiar melhoristas em estudos de diversidade, estrutura populacional, definição de grupos heteróticos, escolha de testadores e predição genômica em espécies que ainda não têm um genoma de referência disponível. Por ser uma alternativa para o rápido avanço do melhoramento de culturas órfãs, nesse contexto, GBS associada ao genoma simulado é uma alternativa eficaz para a realização de estudos genômicos em culturas órfãs, principalmente aquelas que não possuem genoma de referência.

Palavras-chave: GBS, SNP-*array*; formação de grupos heteróticos; predição genômica de cruzamentos-simples; culturas menores; culturas subutilizadas; genoma simulado.

LIST OF FIGURES

LIST OF SUPPLEMENTARY FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY TABLES

CONTENTS

1 INTRODUCTION

Breeding programs to obtain single-crosses are generally based on the development of inbred lines, followed by the progenies evaluation in heterotic pools. Thus, thousands of lines are developed and, when combined, generate single-crosses that needs to be evaluated for their performance in different locations making this step the most expensive in new cultivars development (BERNARDO, 1994; HALLAUER, 2010). For this reason, the scientific community has made significant investments in developing technologies and genomic resources to enable breeding programs to develop cultivars faster with cost-effectiveness (BATLEY; EDWARDS, 2016; BEVAN et al., 2017; BEVAN; UAUY, 2013).

In this context, molecular markers have been used to develop genomic tools to be employed in improving economically important crops (MAMMADOV et al., 2012; THOMSON, 2014). Currently, SNP (Single Nucleotide Polymorphism) markers are the most used in genomic studies (FRITSCHE-NETO et al., 2021), as they provide higher resolution due to their frequent occurrence and uniformity throughout the genome (GUPTA; RUSTGI; MIR, 2008). Rapid advances in next-generation sequencing (NGS) technologies, combined with high levels of diversity in SNP, have made it possible to develop high-throughput genotyping platforms (BACHLAVA et al., 2012).

The current genomic scenario in cultivated plants has faced a revolution due to NGS technologies, which have provided an infinity of sequencing information with remarkable improvements in coverage, time, and costs, making it possible to genotype thousands of samples with many markers (BEVAN; UAUY, 2013). Currently, there are several genotyping platforms for obtaining SNPs throughout the genome, with SNP-array and NGS platforms being the most appropriate for this purpose (RASHEED et al., 2017). There are many array-based genotyping platforms available in major crops such as maize (UNTERSEER et al., 2014), wheat (WINFIELD et al., 2016), rice (SINGH et al., 2015), and soybean (LEE et al., 2015) . These platforms have many advantages, such as fast scans with high call rates and density. However, they present an investigation bias when the set of individuals does not faithfully represent the genetic diversity explored in the study panel. Furthermore, it has a high cost, inaccessible to small breeding programs (FRASCAROLI; SCHRAG; MELCHINGER, 2013; MESSING; DOONER, 2006), especially those of unprofitable species.

Beyond crop-specific SNP arrays, NGS-based platforms are adaptable to various crops, regardless of prior knowledge of genomics, genome size, organization, or ploidy

(RASHEED et al., 2017). Genotyping-by-sequencing (GBS) appears as an alternative to overcome the verification bias since it is based on sequencing and, therefore, allows the discovery of alleles in the diversity panel analyzed, in addition to having a lower cost compared to SNP-array (HESLOT et al., 2013). GBS also appears as an option for genomic studies in orphan crops, especially when they do not have a reference genome (SABADIN et al., 2022). The term "orphan" is derived from the condition of neglect and helplessness of these crops by the scientific community, despite having great food and nutritional potential, leading to the designation of such species as underused, neglected, minor, or orphan crops (TADELE; ASSEFA, 2012). Adopting this technology in poorly studied crops would have a tangible impact on increasing genetic gains. Furthermore, with these data, it is possible to build a simulated genome to perform the SNP calling, where the discovery of polymorphisms will be intrinsic to the study population without using an external genome (MELO; BARTAULA; HALE, 2016). However, GBS generates many low-quality markers with a high rate of lost data (HESLOT et al., 2013).

The number of single-crosses is increasing each year quickly, while the adoption of non-hybrid cultivars is decreasing (SILVA DIAS, 2014). Genetic gains within a single-crosses breeding program can be accelerated with the help of diversity and population structure studies. These studies will support the intended use of germplasm and the identification and allocation of inbred lines in heterotic groups (WU et al., 2016). Making it thus, it's possible to optimize complementarity in addition to divergence. This is because it is known that forming these groups is the key to maximizing heterosis (BOEVEN; LONGIN; WÜRSCHUM, 2016). In crops that still do not have well-established groups, efforts have been made cluster them based on the genetic distance among lines.

Evaluating the performance of all single-crosses combinations of inbred lines that stand out is impractical in most cases. In this sense, genetic designs are crucial for identifying elite parental lines and assessing hybrid performance. Among them, diallel has been widely used to estimate the effects of general and specific combining ability and other genetic parameters (BEYENE; MUGO; KARAYA, 2011; FAN et al., 2014; HALLAUER, 2010). Additionally, genomic prediction models are useful tools for predicting the performance of untested single-crosses. However, the choice of genetic design can influence on the genomic prediction accuracy, with factorial and full diallel designs depicting higher accuracy (FRITSCHE-NETO; AKDEMIR; JANNINK, 2018). Another important point is that incorporate dominance effects are crucial to predict the performance of the population or single-crosses' expression of agronomic traits of more complex inheritance, such as productivity (ALVES et al., 2019; BERNARDO, 2010; FALCONER; MACKAY, 1996). Therefore, to accelerate genetic gain with limited resources, the prediction of single-crosses performance is highly important in modern breeding programs (BASNET et al., 2019).

Recent studies are comparing the performance of genotyping platforms and how this choice affects genomic studies (DARRIER et al., 2019; ELBASYONI et al., 2018). Chu et al. (2020) compared three genotyping platforms (microsatellites (SSR), GBS, and SNP-array) and observed an inquiry bias caused by the SNP-array, causing underestimates of diversity within the population. However, the choice of marker system did not significantly influence the prediction, except for SSR markers due to the low number of markers. Unlike what was observed by Chu et al. (2020), the study carried out by Negro et al. (2019), employing GBS and SNP-array platforms to assess genetic diversity in maize lines, revealed similar trends in the organization of population structure, suggesting that there is no strong ascertainment bias to decipher trends in the genetic structure of the panel.

Sabadin et al. (2022), using maize as a model species, carried out his study with two SNP-genotyping platforms: array and GBS, based on the B73 (tempered) reference genome. In addition, they built simulated reference genomes to perform the discovery of SNPs to capture the polymorphisms intrinsic to the study population without the need for an external genome. The authors aimed to build simulated references with GBS data and verify if using mock genomes to perform SNP calling is suitable for genomic prediction through additive and additive-dominant models. They concluded that the simulated genome based on the entire study population provides reliable estimatives and is a valid alternative for carrying out studies in species where the reference genome is unavailable.

There are some reports available. However, there is still no consensus on how the results provided by different platforms, including the use of a simulated genome, can generate in studies of population structure and genetic diversity, especially when the intention is to use this information for support the formation of heterotic groups, choice of testers and genomic prediction of single-crosses. In addition, the only published study on the mock genome to perform hybrid prediction uses a limited diversity panel (SABADIN et al., 2022) which may compromise the generated estimates. This information will be valuable to leverage genomic studies in minor crops without a reference genome. Therefore, our goals were to verify whether the source of SNP can influence the assessment of the population structure of parental lines; to ascertain if the source of SNP can affect the determination of heterotic groups and the prediction of single-crosses performance, and check if the GBS and the mock genome efficiently performs the SNP calling in orphan crops (without reference genome).

2 MATERIAL AND METHODS

2.1 Phenotypic data

The phenotypic dataset consists of 903 maize single-crosses (FRITSCHE-NETO et al., 2019) derived from a diallel cross between 49 parental lines to a tropical maize diversity, selected based on nitrogen use efficiency (MENDONÇA et al., 2017). Field trials were carried out in Anhembi (22º50'51"S, 48º01'06"W) and Piracicaba (22º42'23"S, 47º38'12"W), in the State of São Paulo, during the second growing season, from January to June 2016 and 2017. Single-crosses were evaluated in an augmented block design, where each block consisted of 16 single-crosses and two checks (commercial single-crosses). In both locations and years, the single-crosses were evaluated under two nitrogen (N) conditions, low N with 30 kg N ha^{-1} and ideal N with 100 kg N ha^{-1}. Each location x year x N level combination was defined as an environment.

Each plot consisted of 7 m rows spaced 0.50 m apart. Conventional fertilization and weed and pest control were carried out. The traits evaluated were grain yield (GY, mg ha^{-1}), plant height (PH, cm), and ear height (EH, cm). The plots were harvested manually, and the grain yield was corrected for 13% moisture. More details on the experimental design and cultivation practices for phenotypic dataset was previously reported by Fritsche-Neto et al. (2018) and Galli et al. (2020).

2.2 Genetic-statistical model for obtaining BLUEs

The joint analysis of each trait was performed to estimate the means of the single-crosses across the environments. Thus, an equation was adjusted to obtain the Best Linear Unbiased Estimator (BLUE) for each genotype and, later, the adjusted means of these across the environments evaluated by the following mixed model were estimated:

$$\boldsymbol{y} = \boldsymbol{Ql} + \boldsymbol{Sb} + \boldsymbol{Tc} + \boldsymbol{Ug} + \boldsymbol{Vi} + \boldsymbol{\mathcal{E}}$$

where $\boldsymbol{y}$ is the vector of phenotypic values of single-crosses and checks; $\boldsymbol{l}$ the vector of fixed effects of the environment (site x year x N level combination); $\boldsymbol{b}$ is the vector of random effect of block nested within environments, where $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_b)$; $\boldsymbol{c}$ is the vector of fixed effects of checks; $\boldsymbol{g}$ is the vector of fixed effects of single-crosses; $\boldsymbol{i}$ is the vector of fixed effects of the interaction checks x environments; $\boldsymbol{\mathcal{E}}$ is the vector of random residual effects, where $\boldsymbol{\mathcal{E}} \sim N(\boldsymbol{0}, \boldsymbol{De})$. An unstructured covariance matrix across environments was assumed for the residual term ($\boldsymbol{De}$). $\boldsymbol{Q}$, $\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are the incidence matrices for $\boldsymbol{l}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{g}$, and $\boldsymbol{i}$. The analysis was performed using the *ASReml-R* (BUTLER et al., 2018).

2.3 Genotypic data and analysis

The 360 lines of tropical maize belonging to the diversity mentioned above panel were genotyped using two high SNP density genotyping platforms: 1) Affymetrix® Axiom Maize Genotyping Array containing 18,413 SNPs (SNP-array) and 2) Genotyping-by-sequencing (GBS) method following the protocol described by (POLAND et al., 2012). In this last method, individual samples of genomic DNA were digested by two restriction enzymes, *PstI* and *MseI*, to reduce the genome complexity uniformly. Subsequently, the samples were included in a sequencing plate, performed on the Illumina NextSeq 500 platform (Illumina Inc., San Diego, CA, United States).

The raw GBS data were used for two purposes: the first was to perform the SNP calling using the B73 line of temperate germplasm as a reference genome. The second purpose was to build a simulated reference genome (mock genome), according to the GBS-SNP-CROP pipeline proposed by Melo et al. (2016), and use it to perform the SNP calling. More details about the mock reference can found in (SABADIN et al., 2022).

Further analysis were performed considering three SNP datasets: 1) SNP-array; 2) GBS with SNP call using B73 as the reference genome (GBS-B73) and 3) GBS with the simulated genome being used as the reference genome (GBS-Mock). For GBS datasets, SNPs were scored from raw data using the TASSEL 5.0 GBSv2 pipeline (GLAUBITZ et al., 2014) according to standard parameters. Using the BWA aligner (LI; DURBIN, 2009), the tags were aligned against the reference genome (GBS-B73 and GBS-Mock).

As two genotyping (SNP-array and GBS) were performed, the lines that showed a very contrasting genotypic profile between the two platforms were removed from the analysis to obtain a fair comparison. Thus, between sequencing errors and divergences in genotypic profiles between platforms, 330 lines remained, among which 45 parental lines make up the diallel, which generated 751 single-crosses. All sets of SNPs underwent quality control, in which low call rate (<90%) and non-biallelic markers were removed from the datasets. The remaining missing data were imputed by the Beagle 5.0 algorithm (BROWNING; ZHOU; BROWNING, 2018). Pairwise linkage disequilibrium was calculated as the correlation of allele frequencies squared (r^2), and values greater than 0.99 were removed from the datasets.

Subsequently, new quality control was performed, in which heterozygous loci in at least one individual were removed, and high-quality polymorphic SNPs from the parental lines were combined (*in silico*) to build an artificial single-crosses genomic matrix. In addition, duplicate markers between chromosomes were removed to avoid overparameterization caused by multicollinearity. Finally, markers with MAF (Minor Allele Frequency) < 0.05 were removed from the single-crosses genomic matrices. All quality control procedures were made using the *SNPRelate* package (ZHENG et al., 2012).

2.4 Analysis of population structure and genetic diversity

The three SNP datasets (SNP-array, GBS-B73, and GBS-Mock) from the 330 parental lines were used to assess the population structure of the panel. In these analyses, precisely, heterozygous loci and rare variants (MAF < 0.05) were considered to capture all diversity and variability to perform principal component analysis (PCA), and determine the relatedness between parental lines.

K-means clustering was applied, using the total Within-cluster Sum of Square (WSS) method to determine the optimal number of clusters so that the total intra-cluster variation is minimized (KASSAMBARA, 2017). For this, the *factoextra* package (KASSAMBARA; MUNDT, 2020) was used. Subsequently, the Kendall method determined the coincidence in forming clusters among the different datasets (KENDALL, 1938). Kendall's tau correlation coefficient was tested at a probability level of 0.01. PCA was performed, and biplots plots were constructed to assess population structure.

The genetic distances between the parental lines were calculated for each SNP dataset using the Rogers distance (ROGERS, 1972). Subsequently, to measure the correlation among the kinship matrices, the Mantel correlation test (MANTEL, 1967) was applied to detect significance. The Mantel correlation test is non-parametric and computes the significance of the correlation similarity measures using 1000 permutations of the rows and columns of one distance matrices. The heatmaps of the genetic distance matrices were obtained using the *superheat R* package (BARTER; YU, 2018). Correlations were obtained using the *vegan* package (OKSANEN et al., 2019), and each analysis was performed for each SNP dataset scenario.

2.5 Full diallel genomic analysis

To find out how diversity and population structure can influence the formation of heterotic groups, it was necessary to construct *in silico* genome of the 751 single-crosses from parental lines. So, at this stage, we combined phenotypic and genotypic information from these individuals to estimate general (GCA) and specific combining abilities (SCA). For this, the following diallel model was adjusted:

$$\mathbf{y = Z_P g_P + Z_H h + \mathcal{E}}$$

where, $\boldsymbol{y}$ is the adjusted phenotypic data vector of the single-crosses for the trait; $\mathbf{g_P}$ is the random effect vector of the GCA captured by the markers of the parental lines, and **h** is the random effect vector of the SCA that denotes the interaction effects across the parental lines. $\mathbf{Z_P}$ e $\mathbf{Z_H}$ are incidence matrices that relate $\boldsymbol{y}$ to $\mathbf{g_P}$ and **h** to $g_p \sim N\,(\mathbf{0}, \sigma^2{}_p\mathbf{G_p})$ and $\mathbf{h} \sim N\,(\mathbf{0}, \sigma^2{}_H\mathbf{H})$, where $\sigma^2{}_P$ and $\sigma^2{}_H$ are variance components associated with GCA and SCA, respectively. And $\mathbf{G_P}$ and **H** are relationship matrices for the parental lines and single-crosses, respectively. Finally, $\mathbf{\mathcal{E}} \sim N\,(\mathbf{0}, \sigma^2{}_\varepsilon\,\mathbf{I})$, where $\sigma^2{}_\varepsilon$ is the variance associated with the residuals.

The $\mathbf{G_P}$ relationship matrix was calculated using the SNP markers according to (VANRADEN, 2008), where $\mathbf{W_P}$ is the matrix of centered and patterned markers. So, $\mathbf{G_P} = \frac{\mathbf{W_p W_p'}}{p}$ (LOPEZ-CRUZ et al., 2015; TECHNOW et al., 2014), where p is the number of markers. This resulted in an average diagonal G_p value of ~ 1; therefore, $\sigma^2{}_p$ was defined on the same scale as $\sigma^2{}_\varepsilon$.

The elements of the **H** matrix were obtained directly from the $\mathbf{G_P}$ (BERNARDO, 2002; TECHNOW et al., 2014). The matrix **H** for all possible crosses was obtained with the Kronecker product between $\mathbf{G_P}$**'s**, $\mathbf{H} = \mathbf{G_p} \otimes \mathbf{G_p}$ (COVARRUBIAS-PAZARAN, 2016).

A model was built with their respective kernels for each SNP marker source. Analyzes were performed using the *ASReml-R* package (BUTLER et al., 2018).

2.6 Heterotic groups and testers

The determination of heterotic groups was performed based on SCA estimates for each trait. These estimates corresponded to a matrix of genetic distances. According to (FALCONER; MACKAY, 1996), the genetic distance between parents positively affects heterosis. This association depends on dominance effects or differences in the frequency of the alleles that control the trait considered (FALCONER, 1960). Burstin et al. (1994) also found that SCA variance is an indicator for predicting hybrid performance by genetic distance between parents. According to this information, it was assumed that the higher the SCA estimates, the greater the distance between the parents and the more significant the heterosis. From this, the 45 lines were divided into heterotic groups.

The SCA estimates were submitted to a clustering algorithm, K-means, which grouped them according to the SCA estimates. To estimate the correlation between the heterotic groups formed for the different genotyping methods, Pearson's correlation was applied and tested at a probability level of 0.01 by Student's t-test. Subsequently, the identification of the best tester in each group was performed according to the GCA estimates. The best tester of a given group was the line that showed the highest GCA with the other group. Based on this, the coincidence of testers between the scenarios was evaluated. Pearson's correlations were tested at a probability level of 0.01 by Student's t-test.

2.7 Obtaining single-crosses combinations and genomic prediction

After the parental lines were divided into heterotic groups, only the single-crosses corresponding to interpopulation crosses via North Carolina II (NCII) design were considered for the following analyses. The number of single-crosses changed according to the configuration of heterotic groups for each trait in the three SNP scenarios.

For the genomic prediction of the single-crosses, an additive-dominant GBLUP (Genomic Best Linear Unbiased Prediction) model was used, as described below:

$$\mathbf{y = Za + Zd + \boldsymbol{\varepsilon}}$$

where, **y** is the adjusted means vector of the single-crosses for the trait; **a** is the vector of additive genetic effects of individuals, where $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G_a}\sigma^2_a)$; **d** is the vector of dominance effects, where $\mathbf{d} \sim N(\mathbf{0}, \mathbf{G_d}\sigma^2_d)$; and $\boldsymbol{\varepsilon}$ is the random effects vector of the residuals, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\,\sigma^2_{\varepsilon})$. **Z** is the incidence matrix for **a** and **d**. σ^2_a is the additive genomic variance, σ^2_d is the dominance genomic variance, and σ^2_{ε} is the residual variance. $\mathbf{G_a}$ and $\mathbf{G_d}$ are the additive and dominance genomic relationship matrices of the single-crosses, where $\boldsymbol{G_a} = \frac{\boldsymbol{W_A W'_A}}{2\sum_{i=1}^{n} p_i(1-p_i)}$ and $\boldsymbol{G_d} = \frac{\boldsymbol{W_D W'_D}}{4\sum_{i=1}^{n} (p_i(1-p_i))^2}$, where p_i is the frequency of an allele at locus i and **W** is the matrix incidence of markers (VANRADEN, 2008). The $\mathbf{W_A}$ matrix was encoded as 0 for A_1A_1, 1 for A_1A_2 heterozygote, and 2 for A_2A_2 homozygote. For $\mathbf{W_D}$, genotypes were coded as 0 for both homozygotes and 1 for the heterozygote. The genomic relationship matrices were built using the *snpReady* package (GRANATO et al., 2018). And the genomic prediction models were performed using the *sommer* package (COVARRUBIAS-PAZARAN, 2016). It is worth noting that all three sets of markers were used to build the kernels. The Mantel correlation test (MANTEL, 1967) was applied to detect significance between the additive and dominance genomic relationship matrices.

To evaluate the model, an alpha-based cross-validation scheme (CV-α) was used (YASSUE et al., 2021), which is an extension of the methodology proposed by Shao (1993).

This methodology is based on assigning treatments to folds based on alpha-lattice principles. CV-α plans to create scenarios with two, three, or four replications, regardless of the number of treatments. Each repetition was divided into folds, and the number of folds determined the percentage of training and validation sets. Each fold between repetitions was based on α (0.1) to reduce the simultaneity of any two treatments in the same fold (block) between repetitions (PATTERSON; WILLIAMS, 1976). We used five folds with four repetitions to estimate the predictive capabilities.

The predictive ability was estimated by Pearson's correlation between predicted genotypic and observed values. Correspondence between phenotypic and genotypic selection was calculated for each set of markers through the percentage of common genotypes selected by their adjusted means from the phenotypic analysis and their Genomic Estimated Breeding Values (GEBV) from the genomic prediction model concerning different intensities of selection (1%, 10%, 20%, 30%, and 40%). The heritability in the broad-sense (H^2) and the narrow-sense (h^2) was also estimated by the equations below:

$$\mathrm{H}^2 = \frac{\hat{\sigma}_{\mathrm{a}}^2+\hat{\sigma}_{\mathrm{d}}{}^2}{(\hat{\sigma}_{\mathrm{a}}^2+\hat{\sigma}_{\mathrm{d}}^2+\hat{\sigma}_{\varepsilon}^2)} \text{ and } \mathrm{h}^2 = \frac{\hat{\sigma}_{\mathrm{a}}^2}{(\hat{\sigma}_{\mathrm{a}}^2+\hat{\sigma}_{\mathrm{d}}^2+\hat{\sigma}_{\varepsilon}^2)}$$

where $\hat{\sigma}_a^2$ is the additive genetic variance, $\hat{\sigma}_d^2$ is the dominance genetic variance, and $\hat{\sigma}_{\varepsilon}^2$ is the residual variance.

3 RESULTS

3.1 Quality control

The first quality control, carried out to prepare the parental lines subset, generated a more accentuated reduction of markers. GBS-B73 was the most affected dataset, with a decrease of 91.5%, followed by GBS-Mock with 89.5% and SNP-array with 31%. The second quality control, which prepared the set of markers for the analysis of single-crosses, reduced the markers more moderately (Table 1).

3.2 Genetic diversity and population structure

According to the WSS method, for all datasets, the optimal number of clusters among the 330 lines that minimized within-group variance and maximized between-group variance was six (Fig. 1). Subsequently, the K-means clustering method showed the six subpopulations formed between the parental lines for all datasets (Fig. 2). There is a remarkable similarity in the arrangement of clusters among datasets. This similarity is confirmed by the coincidence values in the clustering (Table 2), with correlation coefficients above 0.95.

Concerning principal component analysis (PCA), the SNP datasets showed similar performances regarding the variance explained by the principal components. The first principal components hold the highest percentage of explained variance (Fig. 3a). When considering the first ten main components, SNP-array showed the highest value of cumulative explained variance (27.3%). At the same time, GBS-B73 and GBS-Mock presented discounts of 24.1% and 16.8%, respectively (Fig. 3b).

In general, PCA revealed that the first eigenvectors exhibited similar patterns of variance in all combinations between datasets, supported by the coefficient of determination (R^2). However, the other eigenvectors showed less similarity between the captured variance patterns (Fig. 4). The first four eigenvectors of SNP-array and GBS-B73 showed high values of R^2 (Fig. 4a). In contrast, for SNP-array and GBS-Mock, the highest values of R^2 were concentrated in the first three eigenvectors (Fig. 4b). For GBS-B73 and GBS-Mock all eigenvectors showed high magnitude R^2, the former being slightly higher than the others (Fig. 4c).

Bi-plots were constructed to visualize the spatial distribution of lines in all SNP datasets (Fig. 5, Fig. S1, Fig. S2). For this, the first three PCs were used, together with the information obtained by the K-means clustering method (Fig. 2). All datasets showed the same pattern of dispersion among the lines, in agreement with the cluster analysis, which suggests that the SNP datasets capture similar patterns of variance (Fig. 5).

Rogers distance matrices (GD) from all SNP datasets sampled similar groups and subgroups, with slight differences between them (Fig. 6). Regarding the Mantel correlations between the GDs, high magnitude correlations (>0.83) were observed involving different scenarios (Table 3).

3.3 Variance components, genomic heritability, and genomic relationship matrices (GRMs)

Broad and narrow heritabilities were higher for EH, followed by PH and GY (Table S1). For all SNP datasets, GY showed broad-sense heritability, on average, 36% higher than narrow-sense heritability. This difference is significantly smaller for the other characteristics, 15% and 6%, for PH and EH, respectively. For GY, the narrow-sense heritability for all SNP datasets was practically the same. As for PH, there was a slight difference in SNP-array, and GBS-Mock presented heritability slightly higher than GBS-B73. For EH traits, narrow-sense heritability varied little among SNP datasets, with GBS-Mock and SNP-array showing the highest heritabilities. The heritabilities in the broad-sense (H^2) followed the same tendency.

Regarding the additive genomic relationship matrices (*Ga*) across the single-crosses, SNP-array, GBS-B73, and GBS-Mock showed high Mantel correlations (Table 4, Fig. 7a, b, c). On the other hand, the genomic dominance relationship matrices (*Gd*) showed lower correlations than *Ga*. The correlations between the dominant relationship matrices (*Gd*) were lower but still from medium to high. GBS-Mock stands out with a correlation of 0.72 with GBS-B73.

3.4 Heterotic groups and testers

The 45 parental lines were divided into heterotic groups based on SCA estimates as the genetic distance between them for the evaluated traits, GY, PH, and EH. Accordingly, two heterotic groups were formed for all SNP datasets (Fig. 8). The formation of heterotic groups among the SNP datasets was quite similar, with high correlations, higher than 0.94 for GY and 0.87 for PH and EH (Table 5). There was, at most, a change in the allocation of two parental lines between heterotic groups in different SNP datasets. Likewise, the SCA correlations of the parental lines among the SNP datasets were higher than 0.96 (Table S2).

GCA estimates from each parental line, trait, and SNP dataset were used to choose the best tester in each group (Table 6). Thus, the testers matched among SNP datasets in the respective heterotic groups for each trait. The tester chosen, based on GY, for heterotic group one (HP_1) was L023, and for heterotic group two (HP_2), it was L006. As for PH and EH, L001 was elected as HP1 tester and L003 as HP2 tester. The correlations between the GCAs confirm this result, with maximum correlations (Table S3).

3.5 Genomic prediction

The predictive ability estimated by the additive-dominant model for all traits did not vary significantly among SNP datasets (Fig. 9). The mean values of PA were 0.58 for GY, 0.64 for PH, and 0.83 for EH. The coincidence between selected individuals based on the adjusted means of the phenotypic analysis and the GEBVs of the genomic prediction model was generally satisfactory. It increased with rising selection intensity (Fig. 10). Although GY is considered the most complex, the selection coincidence levels of this one was similar to the other traits. SNP-array showed slightly higher coincidence values for almost all selection intensities. However, the different datasets showed approximate coincident values.

4 DISCUSSION

Recent advances in crop genetics and genomics have gained remarkable attention and offered genotyping technologies (CHAKRADHAR; HINDU; REDDY, 2017). Various genotyping platforms are available to meet the most diverse needs regarding costs per sample and different marker densities (THOMSON, 2014). Genotyping-by-sequencing (GBS), in particular, has emerged as a cost-effective strategy for genome-wide SNP discovery and population genotyping due to the simple library preparation and the robust approach to genome reduction (ELSHIRE et al., 2011).

All this progress, including well-characterized genes and vast collections of genetic and genomic resources, focus on a small group of crops (TESTER; LANGRIDGE, 2010), into the detriment of smaller agricultural species, considered as orphans, historically poorly researched (MAYES et al., 2012), in that the large majority do not have a reference genome. Sabadin et al. (2022) showed that using mock genomes could be a worthy strategy that permit to use SNP markers for genomic selection in orphan crops. However, orphan crops breeding program focused on hybrids development also need to determine heterotic groups to maximize the heterosis. Our study aims to go forward and verify the usefulness of mock genomes as a method to permit a reliable heterotic groups clustering.

4.1 Influence of genotyping methods on population structure and diversity

The study of the characterization of genetic diversity, population structure, and genetic relationships among elite parents of germplasm, based on the use of molecular markers, can accelerate genetic gains in breeding programs (ADU et al., 2019; ROMAY et al., 2013). This study helps understand how the germplasm is organized in selecting parents that present effective contributions and in the designation of heterotic groups (WU et al., 2016). In this sense, genomic data not only allows the estimation of genetic diversity but also combines them with phenotypic information to find new functional genes and build prediction models (MILNER et al., 2019). However, in this topic, the focus is on whether, with the simulated reference genome, there is the discovery of the same polymorphisms and how it reflects on the population structure of the lines.

The WSS method indicated the optimal number of clusters by locating a curve on the plot, generally considered an indicator of the optimal number of clusters (KASSAMBARA, 2017). With this information and the results of the K-means clustering, the parental lines were partitioned into subpopulations, where the SNP datasets showed similar behavior (Fig.1, Fig. 2, Table 2), in agreement with the spatial distributions obtained in the bi-plot graphs (Fig. 5), in which all SNP datasets showed the same dispersion pattern between lines. This suggests that the SNP datasets capture similar patterns of variance, despite the difference in the number of markers between them, where GBS-Mock has a lower number (Table 1) and the difference in the genotyping platform itself (array and GBS). Thus, SNP-array, GBS-B73, and GBS-Mock revealed similar performances concerning genetic diversity and the population structure of parental lines. Darrier et al. (2019), when comparing the performance of two genotyping platforms, SNP-array and GBS, to investigate the extent and pattern of genetic variance within a collection of barley genotypes, observed that the two methodologies selectively access the informative polymorphism in different portions of the barley genome. But despite this, their results showed that in the comparison between similarity matrices, there was a positive correlation between both approaches, supporting the validity of the use of both.

PCA shows that these variance patterns captured by the SNP datasets are more similar concerning the first eigenvectors (Fig. 4). However, the captured variance is more consistent when comparing GBS-B73 and GBS-Mock (Fig. 4c). This can be explained by the verification bias existing in the SNP-array since this bias arises when the markers are not obtained from a random sample of the polymorphisms of the population of interest, since the matrix is constructed using temperate maize lines (FRASCAROLI; SCHRAG; MELCHINGER, 2013; HESLOT et al., 2013; UNTERSEER et al., 2014), and the lines in the study are from tropical germplasm.

The matrices of genetic distances among the parental lines revealed similar patterns, showing the formation of subpopulations between the lines (Fig. 6). When using wheat as a model species to test for the presence of verification bias and investigate its impact on genetic diversity estimates, Chu et al. (2020) observed a tendency for SNP-array, leading to an underestimation of molecular diversity within the population. These results agree with a previous study on wheat lines (ELBASYONI et al., 2018) and maize lines (FRASCAROLI; SCHRAG; MELCHINGER, 2013). Despite the verification bias mentioned above and the difference between the reference genome used, the temperate B73 genome, or the Mock genome, the population structure between the lines did not show a significant difference, as the correlations between the matrices of genetic distances were of high magnitude. Even though GBS-Mock uses a different reference genome from SNP-array and GBS-B73, the correlation between them was high (Table 3). Elbasyoni et al. (2018), investigating the influence of SNPs from different genotyping platforms on genomic prediction, observed a high correlation (r = 0.77) between SNP-array and GBS genetic distance matrices. These high magnitude correlations suggest that the broad sampling of diversity is well represented by the approaches used in the study. This is supported by the GWAS study by Darrier et al. (2019). They indicated that methods using SNP-array and GBS could detect markers closely associated with genes that control key phenotypic traits.

4.2 Influence of genotyping methods in the determination of heterotic groups and choice of testers

Heterosis is a fundamental phenomenon in obtaining superior hybrids. Establishing heterotic groups to exploit them effectively throughout the breeding cycles is necessary. These, in turn, are made up of genetically related parental lines, which generate little or no heterosis when crossed with each other. Crossing with lines from another heterotic group tends to result in vigorous hybrids (LEE, 1995). Therefore, genetic diversity among heterotic groups tends to increase the level of heterosis detected in hybrid combinations (FALCONER; MACKAY, 1996; FU et al., 2014). Badu-Apraku et al. (2011) reported in their diallel study between maize lines that their genetic diversity was small and, because of this, distinct heterotic groups could not be identified. Significant genetic diversity was found in a similar study with other maize lines, and two clear heterotic groups were identified. The type of predominant gene action in the parents under investigation is another factor that affects heterotic clustering. When additive and non-additive effects are significant, and there is a predominance of additive gene action over non-additive gene action, heterotic groups are easily identified (BADU-APRAKU et al., 2015, 2016a, 2016b).

The PH and EH traits showed higher proportions of additive variance captured by the Ga matrices than GY (Table S1). Although these traits have polygenic inheritance, GY is the most complex trait and most influenced by dominance deviations (Fischer et al., 2008; Hallauer, 2010). According to Hallauer (2010), most of the loci involved with GY in maize are due to the occurrence of dominance. This is reflected in a greater difference between H^2 and h^2 for GY than for the other traits, confirming the greater influence of dominance deviations on this trait. The additive genomic relationship matrices of the single-crosses (*Ga*) showed high correlations among SNP-array, GBS-B73, and GBS-Mock, indicating that these approaches capture similar additive variance patterns. GBS-Mock captures additive relationships in single-crosses similar to standard procedures, SNP-array, and GBS-B73 (Table 4, Fig. 7a, b, c). On the other hand, the correlations between the dominant relationship matrices (*Gd*) were lower but still from medium to high. In both *Ga* and *Gd*, the correlations between SNP-array and

GBS-Mock were lower, which can be explained by the fact that these SNP datasets use different reference genomes to perform SNP calling.

SCA reflects the action of non-additive gene effects, indicating intra-allelic interactions, is one of the most important parameters in identifying superior hybrids, and is an indicator of genetic distance between parents (CARVALHO, 1993; SPRAGUE; TATUM, 1942). Thus, using the SCA estimates as the genetic distance between the lines to identify the panel structure, two heterotic groups were formed, in which the distance between them is maximized. The correlations between the SCA estimates were almost perfect (Table S2). In other words, SNP-array, GBS-B73, and GBS-Mock presented equivalent SCA estimates. Thus, the composition of heterotic groups practically did not change from one SNP dataset to another. Therefore, the determination of heterotic groups was similar regardless of the platform used (Fig. 8, Table 5).

In addition to presenting distinct heterotic groups, a well-established breeding program also offers good testers. When crossed with parental lines, these provide information about the genetic value of the lines when evaluating the ability to combine between them since it is associated with the additive effects of alleles and additive-type epistatic actions (ALBRECHT et al., 2014; CRUZ; VENCOVSKY, 1989). The correct choice of a tester can have great significance in the expectation of a successful selection process (MIRANDA FILHO, 2018). According to Hallauer and Martinson (1975), a good tester presents simplicity in use, information that correctly classifies the relative merit of the lines, and potential for maximizing genetic gain. Thus, based on the GCA estimates between the lines, testers were elected for each heterotic group based on the evaluated traits and the SNP datasets. As expected, there were no differences in tester choice between SNP datasets, as the correlations between GCA estimates across rows were perfect (Table 6, Table S3).

Once previous results regarding the study of population structure of parental lines, the genotyping approaches produced very similar results but not the same, it was expected that this would somehow influence the formation of heterotic groups and the choice of testers. However, given the results, the genotyping platform, and, more specifically, the approach that uses the simulated genome as a strategy, the GBS-Mock, produces similar results to the standard procedures.

4.3 Influence of genotyping methods on genomic prediction of single-crosses

Assessing the performance of all single-crosses combinations of parental lines that excel in a breeding program is impractical in most cases, given that the number of combinations grows exponentially as the number of elite parents increases. Thus, obtaining estimates of the genetic values of single-crosses not evaluated became viable with the increased availability of molecular markers and genomic prediction models (HALLAUER, 2010). Therefore, to accelerate genetic gain with limited resources, the prediction of single-crosses performance is highly important in modern breeding programs (BASNET et al., 2019).

However, few works still address how genotyping platforms influence single-crosses' prediction and, more specifically, regarding the mock genome as a tool for more sophisticated studies, such as genomic prediction. Only one recent study shows the mock genome's efficiency in predicting maize single-crosses, which may be an alternative for crops that do not yet have a reference genome (SABADIN et al., 2022). However, our study is more complete and more representative because getting approaches from the population structure phase is crucial for the intended use of germplasm through the division of heterotic groups, the definition of testers, and, finally, the genomic prediction of single-crosses.

GY showed the lowest predictive abilities in all SNP datasets, and EH was the highest by the additive-dominant GBLUP prediction model (Fig. 9, Table S1). Combs and Bernardo (2013) suggested that genomic predictions are more accurate for traits with higher heritability. In the results of Hayes et al. (2010), complex traits controlled by many small effect loci, such as GY, showed lower predictive abilities than less complex traits. Although GBS-Mock has a lower number of markers, this approach presented a similar performance to the other SNP datasets for all characteristics, corroborating the hypothesis that it is possible to substantially reduce the number of markers and maintain a high predictive ability (MA et al., 2016; SOUSA et al., 2019; TAYEH et al., 2015), with the caveat that over the generations, the accuracy decreases significantly. Higher markers densities are recommended for better long-term selection responses (DOVALE et al., 2022). These results were expected since the prediction model used was the GBLUP, which uses a genomic relationship matrix between individuals to perform the predictions. The genetic distance estimates between the SNP datasets were very similar (Fig. 6).

Selection intensity must be chosen thoughtfully, as genetic variability can be drastically reduced with high selection pressure. The choice of appropriate selection intensities depends on the size of the population and the duration of the breeding program, whether short-term or long-term. In general, selection intensities ranging from 10 to 40% are used in plant breeding, the highest being applied at the beginning of a breeding program (HALLAUER, 2010). For the coincidence of individuals by phenotypic selection and genomic selection, the SNP datasets showed similar behavior as the selection intensity was increased, being more pronounced from 1 to 10% of selection intensity. From then on, observing the coincidence of selection gains smaller increments (Fig. 10). Our results for predictive ability and coincidence of selection agree with the results of Sabadin et al. (2022). It is valid to consider that different intensities modify the response rates. Thus, this coincidence between phenotypic and genomic selections is expected to reach a plateau and subsequently decrease.

Despite the apparent differences between SNP datasets, the general message is that these approaches perform comparably in the types of analyses performed in this study, even accessing different types of genomic sequences. While SNP-array is derived from exome capture and therefore focused on coding sequence variation, the GBS data represent a wider diversity survey in genomic regions associated with low levels of DNA methylation, which may also include many genes and gene regulatory regions (DARRIER et al., 2019; NEGRO et al., 2019). On the other hand, the physical distribution of markers reveals higher frequencies of SNPs at the gene-rich telomeric ends of each of the chromosomes for both approaches, with this frequency being more pronounced in SNP-array (BAYER et al., 2017). The platforms probably capture nearby markers in linkage disequilibrium with QTLs (Quantitative Trait Loci). In this sense, using different platforms can be advantageous, as it allows the identification of different QTLs.

4.4 Possible applications of the Mock genome in plant breeding

The advances in genomics in recent years have increased the accuracy and efficiency of breeding programs for many crops, especially those that dominate global food production. However, the adoption of genomic enhancement for several other staple crops essential in developing countries is still limited, especially for traits under complex genetic control, which are crucial to crop performance (VARSHNEY et al., 2012). Until recently, only the main commercial crops benefited from state-of-the-art technologies. However, the development of the GBS platform emerged as an alternative for using such technologies to be viable for orphan crops (DAVEY et al., 2011; VARSHNEY et al., 2009; VARSHNEY; MAY, 2012). Approaches like this have the potential to convert orphan crops into crops rich in genomic resources (VARSHNEY et al., 2012).

With the development of new technologies accessible to these neglected crops, the breeding process can be substantially reduced. Previously, this process was much slower than nowadays. Rice, for example, took almost 20 years to stop being an orphan crop and become a basic model for cereals (VARSHNEY et al., 2009). Introducing these crops into the genomic era also accelerates the identification of genes underlying important agronomic traits and improves our understanding of the evolution of these species (YE; FAN, 2021). However, many more minor crops are becoming rich in genetic resources as a result of investments from various public and private initiatives, such as the African Orphan Crops Consortium (AOCC) (HENDRE et al., 2019), which is a global partnership that is generating resources genomics for 101 African orphans. One of the objectives of this Consortium is to create reference genomes for these cultures. According to Armstead et al. (2009), species without a reference genome are harmed by being left out of these technologies that can improve breeding schemes and accelerate the development of new cultivars. Although some efforts are being made to pay greater attention to these cultures (CHIURUGWI et al., 2019; GREGORY et al., 2019; JAMNADASS et al., 2020; PADULOSI, 2017), the ideal is still far from being achieved with a view to several species relevant to local diets around the world that are understudied.

This process takes time despite initiatives and investments to sequence and assembles reference genomes for orphan crops. Not all crops will benefit, so they will not be able to take advantage of modern breeding tools. While these advances are being consolidated, mock genomes can an alternative, where the absence of a reference genome presented a barrier to the efficient use of GBS data (HALE; MELO; GUSTAFSON, 2018; MELO; GUTHRIE; HALE, 2017). In the meantime, the present study has shown that using a population-adapted mock reference to perform SNP discovery is a valid alternative, particularly for species where the reference genome is unavailable. With this approach, it was possible to carry out studies to outline a breeding program as a whole, from studies of diversity and population structure(ARREDONDO; MARCHINI; CRUZAN, 2018; BARTAULA et al., 2018; SUNSERI et al., 2018), to genomic prediction studies Sabadin et al. (2022). However, it is important to emphasize that a population with maximum representativeness must be considered when building a mock reference aiming to capture all polymorphism into the population (SABADIN et al., 2022).

Despite these significant advantages, using a mock genome in genomic studies must consider some caveats. For example, cross-pollinated crops or orphan polyploid crops have genomes too complex to be sequenced. Another challenge lies in the SNP calling due to the limitations of GBS, which can lead to incorrect identification of homozygotes and heterozygotes because of the low coverage of NGS reads, in addition to a large number of lost and low-quality data (HESLOT et al., 2013). According to Sabadin et al. (2022), the mock genomes do not present the physical position of the markers in a constant reference, which hinders studies such as GWAS and candidate gene discovery. The study by Negro et al., (2019) states that SNP-array and GBS are complementary to detect QTLs tagging different haplotypes in association studies. In this sense, using other platforms can be advantageous, as it allows the identification of additional QTLs. However, no studies still demonstrate the performance of using mock genome for these purposes. When looking for these larger effect marks, the results will probably differ from those obtained with SNP-array due to changes in coverage between platforms.

Given what has been shown, it is possible to infer and recommend that a mock genome constructed from the population's polymorphisms to perform the SNP calling is an excellent strategy to support plant breeders in studies of diversity, population structure, the definition of heterotic groups, choice of testers and genomic prediction in species that still do not have a reference genome available, which is an alternative for the rapid advancement of orphan crop improvement. This approach will play a key role in improving the genetic potential of orphan crops and helping develop sustainable food systems.

5 CONCLUSIONS

Different SNP sources showed similar results regarding the population structure of the lines, determining heterotic groups, and in the genomic prediction of single-crosses, in which GBS-Mock gives results comparable to the standard approaches, i.e., SNP-array and GBS with the B73 line genome as the reference genome. In this context, GBS associated with the mock genome is an effective alternative for performing genomic studies in orphan crops, especially for species that do not have a genome reference.

REFERENCES

ADU, G. B. et al. Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. **PLoS ONE**, [*s. l.*] v. 14, n. 4, 1 abr. 2019.

ALBRECHT, J. et al. Correlated loss of ecosystem services in coupled mutualistic networks. **Nature Communications**, v. 5, 8 maio, 2014.

ALVES, F. C. et al. Bayesian analysis and prediction of hybrid performance. **Plant Methods**, v. 15, n. 1, 7 fev. 2019.

ARMSTEAD, I. et al. Bioinformatics in the orphan crops. **Briefings in Bioinformatics**, 2009.

ARREDONDO, T. M.; MARCHINI, G. L.; CRUZAN, M. B. Evidence for human-mediated range expansion and gene flow in an invasive grass. Proceedings of the Royal Society B: Biological Sciences. **Anais**...Royal Society Publishing, 11 jul. 2018.

BACHLAVA, E. et al. Snp discovery and development of a high-density genotyping array for sunflower. **PLoS ONE**, v. 7, n. 1, 4 jan. 2012.

BADU-APRAKU, B. et al. Biplot analysis of diallel crosses of early maturing tropical yellow maize inbreds in stress and nonstress environments. **Crop Science**, v. 51, n. 1, p. 173–188, jan. 2011.

BADU-APRAKU, B. et al. Heterotic responses among crosses of IITA and CIMMYT early white maize inbred lines under multiple stress environments. **Euphytica**, v. 206, n. 1, p. 245–262, 1 nov. 2015.

BADU-APRAKU, B. et al. Gene action and heterotic groups of early white quality protein maize inbreds under multiple stress environments. **Crop Science**, v. 56, n. 1, p. 183–199, 1 jan. 2016a.

BADU-APRAKU, B. et al. Heterotic patterns of IITA and CIMMYT early-maturing yellow maize inbreds under contrasting environments. **Agronomy Journal**, v. 108, n. 4, p. 1321–1336, 1 jul. 2016b.

BARTAULA, R. et al. An interspecific barberry hybrid enables genetic dissection of non-host resistance to the stem rust pathogen Puccinia graminis. **Journal of Experimental Botany**, v. 69, n. 10, p. 2483–2493, 27 abr. 2018.

BARTER, R. L.; YU, B. Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. **Journal of Computational and Graphical Statistics**, v. 27, n. 4, p. 910–922, 2 out. 2018.

BASNET, B. R. et al. Hybrid Wheat Prediction Using Genomic, Pedigree, and Environmental Covariables Interaction Models. **The Plant Genome**, v. 12, n. 1, p. 180051, mar. 2019.

BATLEY, J.; EDWARDS, D. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. **Current Opinion in Plant Biology** Elsevier Ltd, 1 abr. 2016.

BAYER, M. M. et al. Development and evaluation of a barley 50k iSelect SNP array. **Frontiers in Plant Science**, v. 8, 17 out. 2017.

BERNARDO, R. CROP BREEDING, GENETICS & CYTOLOGY Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. [*s.l*: s.n.].

BERNARDO, R. **Breeding for quantitative traits in plants**. Woodbury: Stemma Press, 2002.

BERNARDO, R. Genome-wide selection with minimal crossing in self-pollinated crops. **Crop Science**, mar. 2010.

BEVAN, M. W. et al. **Genomic innovation for crop improvement**. **Nature**Nature Publishing Group, 15 mar. 2017.

BEVAN, M. W.; UAUY, C. Genomics reveals new landscapes for crop improvement R E V I E W. **Genome Biology**. [*s.l*: s.n.].

BEYENE, Y.; MUGO, S. N.; KARAYA, H. Genotype by environment interactions and yield stability of stem borer resistant maize hybrids in Kenya Impact of Genetically modified organisms View project Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG) View project Tadele Tefera International Water Management InstituteArticle in **AFRICAN JOURNAL OF BIOTECHNOLOGY**. [s.l: s.n.].

BOEVEN, P. H. G.; LONGIN, C. F. H.; WÜRSCHUM, T. A unified framework for hybrid breeding and the establishment of heterotic groups in wheat. **Theoretical and Applied Genetics**, v. 129, n. 6, p. 1231–1245, 1 jun. 2016.

BROWNING, B. L.; ZHOU, Y.; BROWNING, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. **American Journal of Human Genetics**, v. 103, n. 3, p. 338–348, 6 set. 2018.

BURSTIN, J. et al. **Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines**. [s.l: s.n.].

BUTLER, D. G. et al. ASReml-R Reference Manual Version 4 ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML). [s.l: s.n.].

CARVALHO, G. R. Evolutionary aspects of fish distribution: genetic variability and adaptation. **Journal of Fish Biology**, v. 43, n. sa, p. 53–73, dez. 1993.

CHAKRADHAR, T.; HINDU, V.; REDDY, P. S. **Genomic-based-breeding tools for tropical maize improvement**. **Genetica**Springer International Publishing, 1 dez. 2017.

CHIURUGWI, T. et al. Speed breeding orphan crops. **Theoretical and Applied Genetics**Springer Verlag, 1 mar. 2019.

CHU, J. et al. Suitability of Single-Nucleotide Polymorphism Arrays Versus Genotyping-By-Sequencing for Genebank Genomics in Wheat. **Frontiers in Plant Science**, v. 11, 14 fev. 2020a.

CHU, J. et al. Suitability of Single-Nucleotide Polymorphism Arrays Versus Genotyping-By-Sequencing for Genebank Genomics in Wheat. **Frontiers in Plant Science**, v. 11, 14 fev. 2020b.

COMBS, E.; BERNARDO, R. Accuracy of Genome-wide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. **The Plant Genome**, v. 6, n. 1, mar. 2013.

COVARRUBIAS-PAZARAN, G. Genome-Assisted prediction of quantitative traits using the r package sommer. **PLoS ONE**, v. 11, n. 6, 1 jun. 2016.

CRUZ, C. D.; VENCOVSKY, R. Comparação de alguns metodos de análise dialélica. **Revista Brasileira de Genética = Brazilian Journal of Genetics**, v. 12, n. 2, p. 425–38, 1989.

DARRIER, B. et al. A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. **Frontiers in Plant Science**, v. 10, 16 abr. 2019.

DAVEY, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature Reviews Genetics**, jun. 2011.

DOVALE, J. C. et al. Genotyping marker density and prediction models effects in long-term breeding schemes of cross-pollinated crops. **Theoretical and Applied Genetics**, 20 out. 2022.

ELBASYONI, I. S. et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. **Plant Science**, v. 270, p. 123–130, 1 maio 2018.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, v. 6, n. 5, 2011.

FALCONER, D. S. **Introduction to Quantitative Genetics**. Edinburgh: Oliver and Boyd Ltd., 1960.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. 4th. ed. Harlow: Addison Wesley Longman, 1996.

FAN, X. M. et al. Reciprocal diallel crosses impact combining ability, variance estimation, and heterotic group classification. **Crop Science**, v. 54, n. 1, p. 89–97, jan. 2014.

FISCHER, S. et al. Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. **Plant Breeding**, v. 127, n. 5, p. 446–451, out. 2008.

FRASCAROLI, E.; SCHRAG, T. A.; MELCHINGER, A. E. Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. **Theoretical and Applied Genetics**, v. 126, n. 1, p. 133–141, jan. 2013.

FRITSCHE-NETO, R. et al. USP tropical maize hybrid panel. **Mendeley Data**, v. 3, 2019.

FRITSCHE-NETO, R. et al. Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review. **Frontiers in Plant Science**Frontiers Media S.A., 1 jul. 2021.

FRITSCHE-NETO, R.; AKDEMIR, D.; JANNINK, J. L. Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. **Theoretical and Applied Genetics**, v. 131, n. 5, p. 1153–1162, 1 maio 2018.

FU, D. et al. Utilization of crop heterosis: A review. **Euphytica**Kluwer Academic Publishers, 1 maio 2014.

GALLI, G. et al. On the usefulness of parental lines GWAS for predicting low heritability traits in tropical maize hybrids. **PLoS ONE**, v. 15, n. 2, 1 fev. 2020.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high-capacity genotyping by sequencing analysis pipeline. **PLoS ONE**, v. 9, n. 2, 28 fev. 2014.

GRANATO, I. S. C. et al. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, v. 38, n. 8, 1 ago. 2018.

GREGORY, P. J. et al. Crops For the Future (CFF): an overview of research efforts in the adoption of underutilised species. **Planta**Springer Verlag, 1 set. 2019.

GUPTA, P. K.; RUSTGI, S.; MIR, R. R. Array-based high-throughput DNA markers for crop improvement. **Heredity**, jul. 2008.

HALE, I.; MELO, A. T. O.; GUSTAFSON, H. Sex-linked molecular markers for two cold-hardy kiwifruit species, Actinidia arguta and A. Kolomikta. **European Journal of Horticultural Science**, v. 83, n. 4, p. 236–246, 2018.

HALLAUER, A. R.; C. M. J.; M. F. J. B. **Quantitative Genetics in Maize Breeding**. [*s.l*: s.n.]. v. 6

HALLAUER, A. R.; MARTINSON, C. A. Maternal Effects in Maize Hybrids Infected with Bipolaris maydis (Nisikado) Shoemaker, Race T 1. **Crop Science**, v. 15, n. 5, p. 686–689, set. 1975.

HAYES, B. J. et al. Genetic architecture of complex traits and accuracy of genomic Prediction: Coat colour, Milk-fat percentage, and type in holstein cattle as contrasting model traits. **PLoS Genetics**, v. 6, n. 9, set. 2010.

HENDRE, P. S. et al. African Orphan Crops Consortium (AOCC): status of developing genomic resources for African orphan crops. **PlantaSpringer** Verlag, , 1 set. 2019.

HESLOT, N. et al. Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. **PLoS ONE**, v. 8, n. 9, 5 set. 2013.

JAMNADASS, R. et al. Enhancing African orphan crops with genomics. **Nature Genetics**Nature Research, 1 abr. 2020.

KASSAMBARA, A. Multivariate Analysis I Practical Guide To Cluster Analysis in R Unsupervised Machine Learning. [*s.l: s.n.*].

KASSAMBARA, A.; MUNDT, F. Extract and Visualize the Results of Multivariate Data Analyses. Package “factoextra”.

KENDALL, M. G. A New Measure of Rank Correlation. **Biometrika**, v. 30, n. 1/2, p. 81–93, jun. 1938.

LEE, M. DNA MARKERS AND PLANT BREEDING PROGRAMS. v. 55, p. 265–344, 1995.

LEE, Y. G. et al. Development, validation and genetic analysis of a large soybean SNP genotyping array. **Plant Journal**, v. 81, n. 4, p. 625–636, 1 fev. 2015.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, jul. 2009.

LOPEZ-CRUZ, M. et al. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. **G3: Genes, Genomes, Genetics**, v. 5, n. 4, p. 569–582, 2015.

MA, Y. et al. Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). **Molecular Breeding**, v. 36, n. 8, 1 ago. 2016.

MAMMADOV, J. et al. SNP markers and their impact on plant breeding. **International Journal of Plant Genomics**, 2012.

MANTEL, N. The detection of disease clustering and a generalized regression approach. **Cancer Research**, v. 27, p. 209–220, 1967.

MAYES, S. et al. The potential for underutilized crops to improve security of food production. **Journal of Experimental Botany**, v. 63, n. 3, p. 1075–1079, fev. 2012.

MELO, A. T. O.; BARTAULA, R.; HALE, I. GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. **BMC Bioinformatics**, v. 17, n. 1, 12 jan. 2016.

MELO, A. T. O.; GUTHRIE, R. S.; HALE, I. GBS-based deconvolution of the surviving North American collection of cold-hardy kiwifruit (Actinidia spp.) germplasm. **PLoS ONE**, v. 12, n. 1, 1 jan. 2017.

MENDONÇA, L. DE F. et al. Accuracy and simultaneous selection gains for N-stress tolerance and N-use efficiency in maize tropical lines. **Scientia Agricola**, v. 74, n. 6, p. 481–488, 2017.

MESSING, J.; DOONER, H. K. Organization and variability of the maize genome. **Current Opinion in Plant Biology**, abr. 2006.

MILNER, S. G. et al. Genebank genomics highlights the diversity of a global barley collection. **Nature Genetics**, v. 51, n. 2, p. 319–326, 1 fev. 2019.

MIRANDA FILHO, J. B. Testadores e dialelo. Em: DELIMA, R.; BORÉM, A. (Eds.). **Melhoramento de Milho**. 1. ed. Viçosa, MG: [s.n.]. p. 130–158.

NEGRO, S. S. et al. Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. **BMC Plant Biology**, v. 19, n. 1, 16 jul. 2019.

OKSANEN, J. et al. Vegan: Community Ecology Package.

PADULOSI, S. Bring NUS back to the table! p. 21–22, 2017.

PATTERSON, H. D.; WILLIAMS, E. R. A New Class of Resolvable Incomplete Block Designs. **Biometrika**, v. 63, n. 1, p. 83, abr. 1976.

POLAND, J. A. et al. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **PLoS ONE**, v. 7, n. 2, 28 fev. 2012.

RASHEED, A. et al. Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. **Molecular PlantCell** Press, 7 ago. 2017.

ROGERS, J. S. Measures of Genetic Similarity and Genetic Distance. **Studies in Genetics VII, University of Texas Publication 7213**, p. 145–153, 1972.

ROMAY, M. C. et al. Comprehensive genotyping of the USA national maize inbred seed bank. **Genome Biology**, v. 14, n. 6, 11 jun. 2013.

SABADIN, F. et al. Population-tailored mock genome enables genomic studies in species without a reference genome. **Molecular Genetics and Genomics**, v. 297, n. 1, p. 33–46, 1 jan. 2022.

SHAO, J. Linear Model Selection by Cross-Validation. **Journal of the American Statistical Association**, v. 88, n. 422, p. 486, jun. 1993.

SILVA DIAS, J. C. DA. Guiding strategies for breeding vegetable cultivars. **Agricultural Sciences**, v. 05, n. 01, p. 9–32, 2014.

SINGH, N. et al. Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. **Scientific Reports**, v. 5, 26 jun. 2015.

SOUSA, M. B. et al. Increasing accuracy and reducing costs of genomic prediction by marker selection. **Euphytica**Springer Netherlands, 1 fev. 2019.

SPRAGUE, G. F.; TATUM, L. A. General vs Combining Ability in Single Crosses of Corn. **Agronomy Journal**, v. 34, n. 10, p. 923–932, out. 1942.

SUNSERI, F. et al. Single nucleotide polymorphism profiles reveal an admixture genetic structure of grapevine germplasm from Calabria, Italy, uncovering its key role for the diversification of cultivars in the Mediterranean Basin. **Australian Journal of Grape and Wine Research**, v. 24, n. 3, p. 345–359, 1 jul. 2018.

TADELE, Z.; ASSEFA, K. Increasing food production in africa by boosting the productivity of understudied crops. **Agronomy**MDPI AG, 1 dez. 2012.

TAYEH, N. et al. Genomic prediction in pea: Effect of marker density and training population size and composition on prediction accuracy. **Frontiers in Plant Science**, v. 6, n. NOVEMBER, 17 nov. 2015.

TECHNOW, F. et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. **Genetics**, v. 197, n. 4, p. 1343–1355, 2014.

TESTER, M.; LANGRIDGE, P. Breeding Technologies to Increase Crop Production in a Changing World. **Science**, v. 327, n. 5967, p. 818–822, 12 fev. 2010.

THOMSON, M. J. High-Throughput SNP Genotyping to Accelerate Crop Improvement. **Plant Breeding and Biotechnology**, v. 2, n. 3, p. 195–212, 30 set. 2014.

UNTERSEER, S. et al. A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. **BMC Genomics**, v. 15, n. 1, 29 set. 2014.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VARSHNEY, R. K. et al. Orphan legume crops enter the genomics era! **Current Opinion in Plant Biology**, abr. 2009.

VARSHNEY, R. K. et al. Can genomics boost productivity of orphan crops? **Nature Biotechnology**, v. 30, n. 12, p. 1172–1176, 7 dez. 2012.

VARSHNEY, R. K.; MAY, G. D. Next-generation sequencing technologies: Opportunities and obligations in plant genomics. **Briefings in Functional Genomics**, jan. 2012.

WINFIELD, M. O. et al. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. **Plant Biotechnology Journal**, v. 14, n. 5, p. 1195–1206, 1 maio 2016.

WU, Y. et al. Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. **Theoretical and Applied Genetics**, v. 129, n. 4, p. 753–765, 1 abr. 2016.

YASSUE, R. M. et al. CV-α: designing validations sets to increase the precision and enable multiple comparison tests in genomic prediction. **Euphytica**, v. 217, n. 6, 1 jun. 2021.

YE, C. Y.; FAN, L. Orphan Crops and their Wild Relatives in the Genomic Era. **Molecular Plant**Cell Press, , 4 jan. 2021.

ZHENG, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. **Bioinformatics**, v. 28, n. 24, p. 3326–3328, dez. 2012.

APPENDIX A - LIST OF TABLES AND FIGURES

Table 1 Number of markers scored (raw data) and the final number of markers used to assess 330 tropical parental lines and 751 maize single-crosses after quality control for all SNPs datasets

	SNP datasets[a]		
	SNP-array	**GBS-B73**	**GBS-Mock**
Raw data	18,413	131,350	46,926
Lines[b]	12,704	11,153	4,935
Single-crosses[c]	11,884	10,361	4,801

[a] SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines;

[b] number of markers used to evaluate parental lines (population structure)

[c] number of markers used to assess single-crosses (diallel, heterotic groups, genomic prediction).

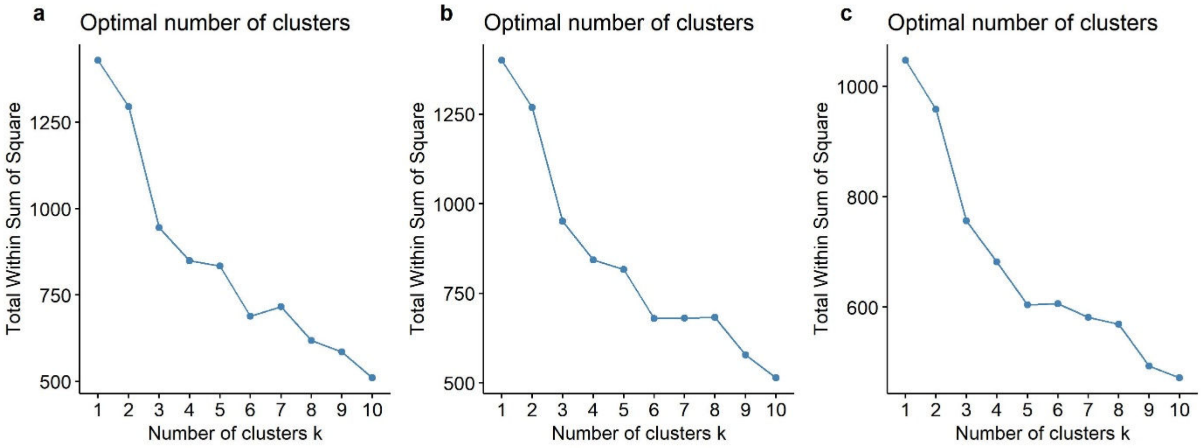

Figure 1 Optimal number of clusters formed among the 330 parental lines for all SNP datasets by Total Within Sum of Square (WSS) method. **a** SNP-array; **b** GBS-B73; **c** GBS-Mock.

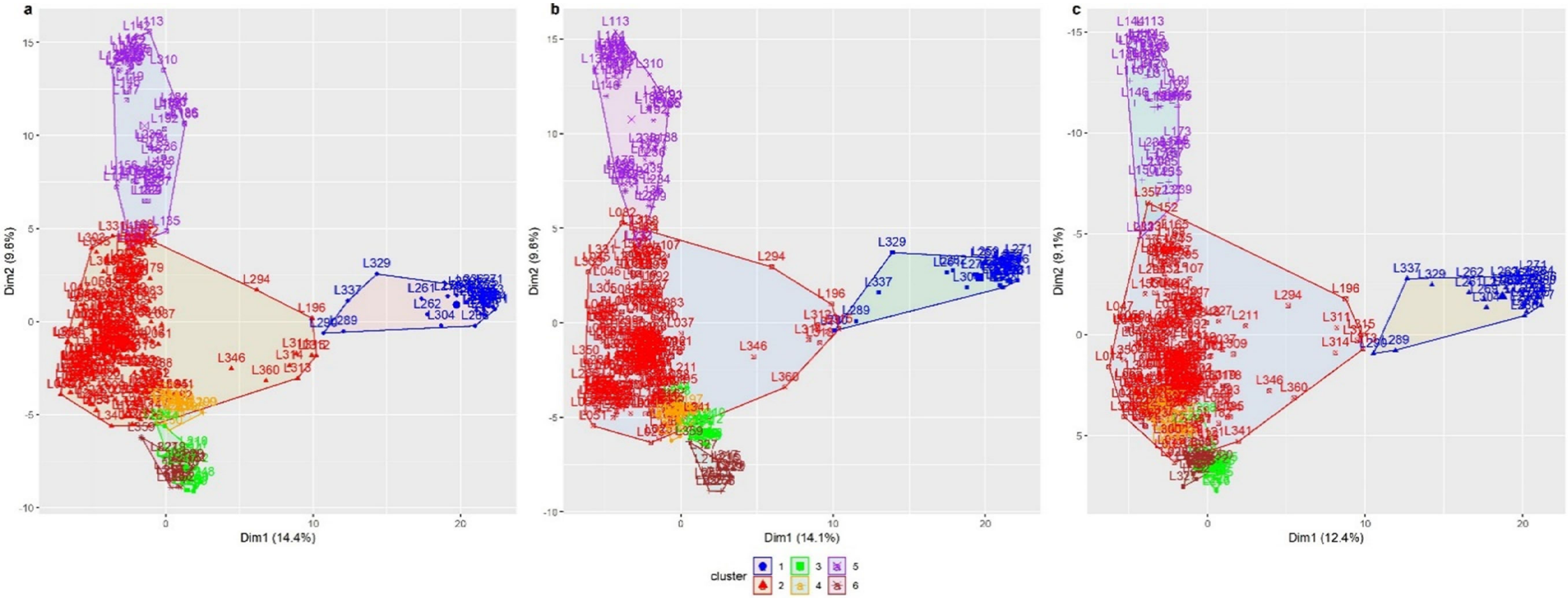

Figure 2 K-means clustering method among the 330 parental lines for all SNP datasets a SNP-array; b GBS-B73; c GBS-Mock.

Table 2 Kendall's correlation in the clustering of datasets

	GBS-B73	**GBS-Mock**
SNP-array	0.99**	0.96**
GBS-B73	-	0.97**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Empirical significance level from permutations.

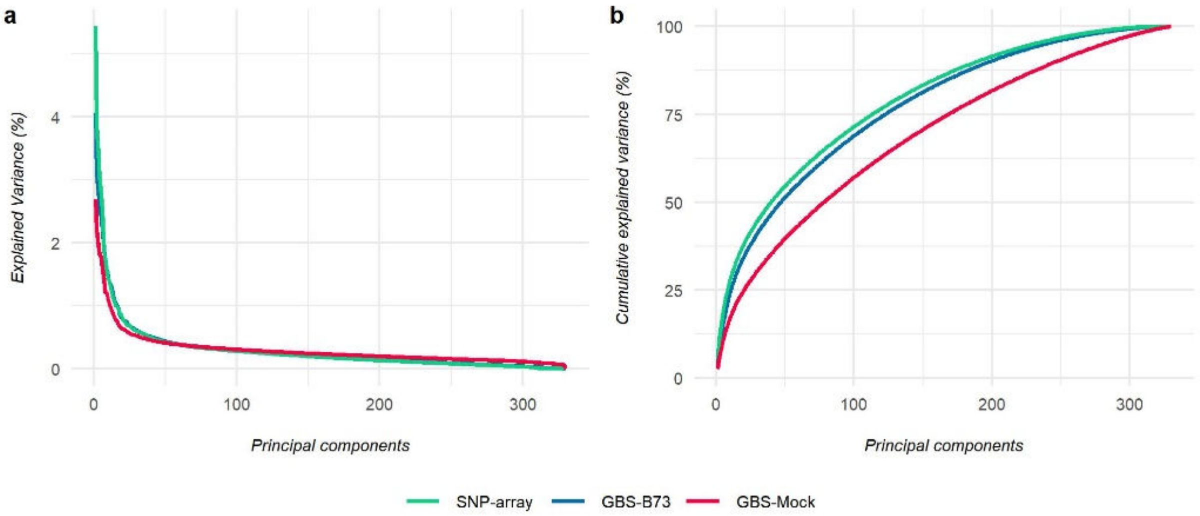

Figure 3 a Variance explained by the principal components (PCA) from SNP-array, GBS-B73 e GBS-Mock SNP datasets for 330 tropical parental lines; **b** Cumulative explained variance estimated by principal components from SNP-array, GBS-B73 e GBS-Mock SNP datasets for 360 tropical parental lines.

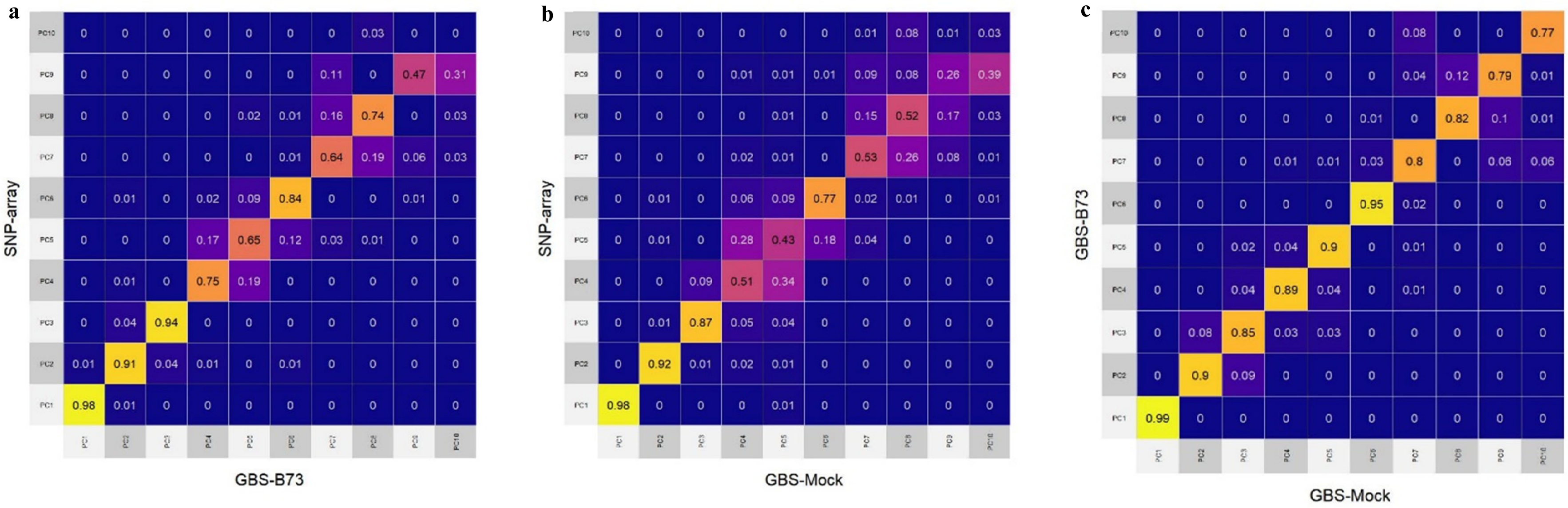

Figure 4 Heatmap of the coefficient of determination (R^2) of the ten first eigenvectors among **a** SNP-array and GBS-B73; **b** SNP-array and GBS-Mock; e **c** GBS-B73 and GBS-Mock.

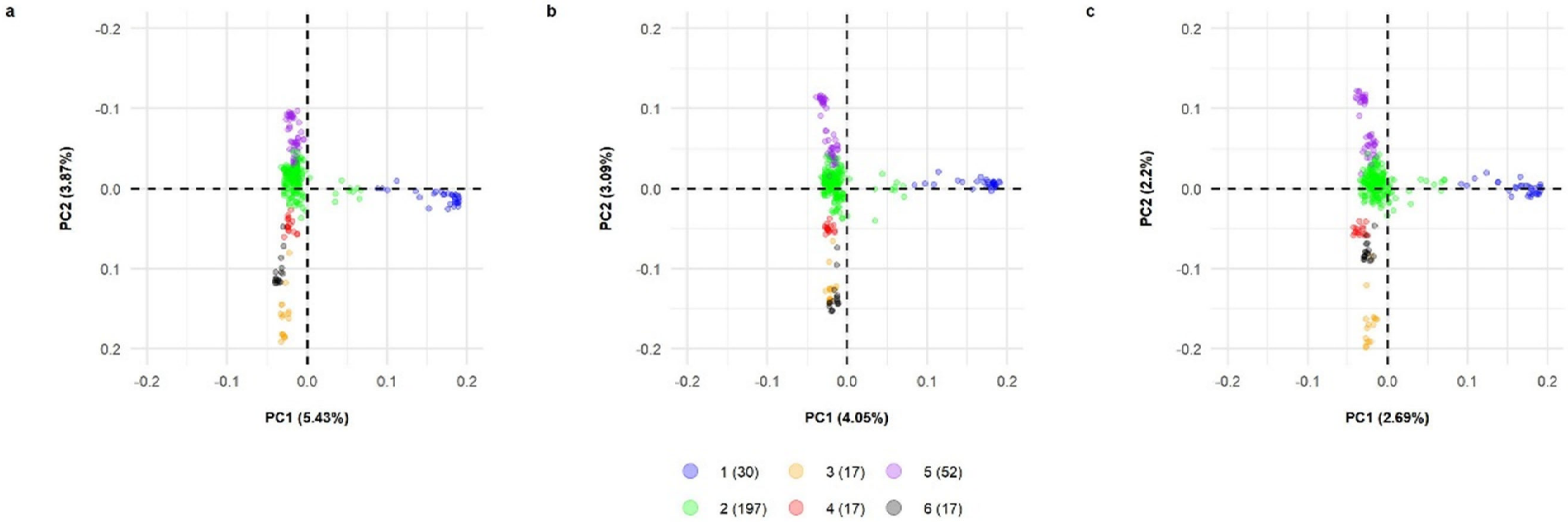

Figure 5 Bi-plot among two first principal components using all datasets for 330 tropical parental lines **a** SNP-array; **b** GBS-B73 and **c** GBS-Mock. Explained variance percentages of each principal component are in parentheses. Clusters were used to color-coded parental lines.

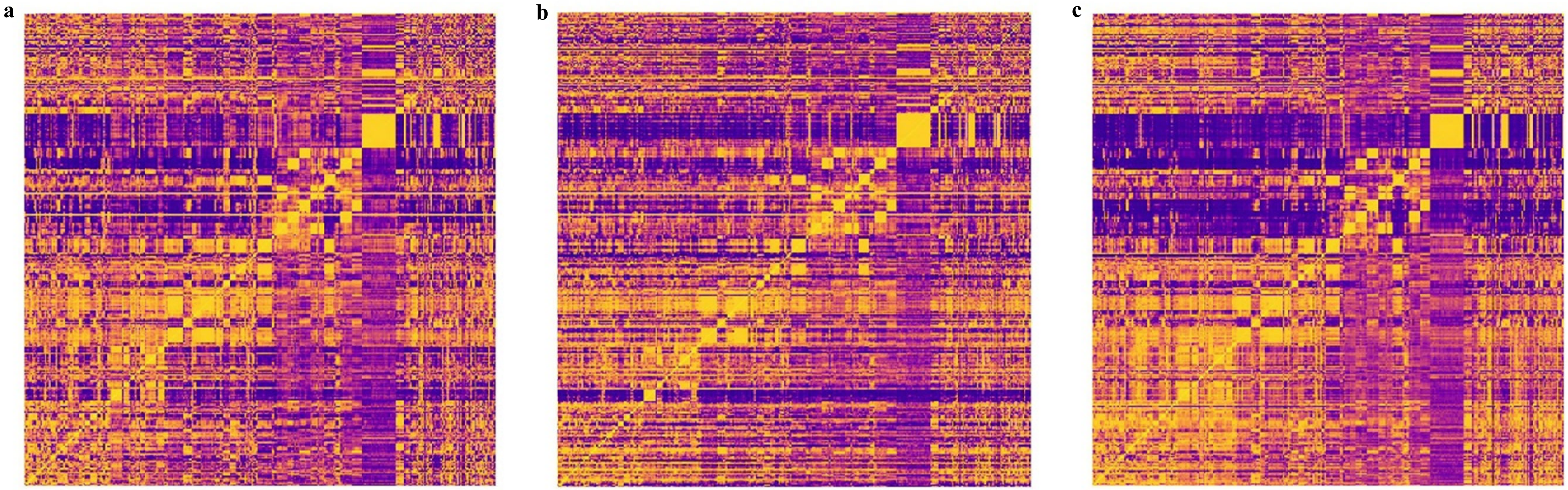

Figure 6. Heatmaps of the Rogers genetic distance matrices estimated from a SNP-array, b GBS-B73, and c GBS-Mock-All SNP datasets for 330 tropical parental lines. Lines and columns of each plot were clustered according to the Euclidian distance performed in the Roger genetic distance matrix from the SNP-array dataset.

Table 3. Mantel correlation of Rogers genetic distance (**GD**) matrices estimated from SNP-array, GBS-B73, and GBS-Mock markers.

	GBS-B73	**GBS-Mock**
SNP-array	0.91**	0.83**
GBS-B73	-	0.91**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Empirical significance level from permutations.

Roger's genetic distance (**GD**) matrices were computed with markers from 330 parental lines data.

Table 4. Mantel correlation of additive genomic relationship ($\boldsymbol{Ga}$) and dominance genomic relationship ($\boldsymbol{Gd}$) matrices from SNP-array, GBS-B73, and GBS-Mock markers

		GBS-B73	**GBS-Mock**
$\boldsymbol{G_a}$	**SNP-array**	0.97**	0.96**
	GBS-B73	-	0.99**
$\boldsymbol{G_d}$	**SNP-array**	0.78**	0.58**
	GBS-B73	-	0.72**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Empirical significance level from permutations.

$\boldsymbol{Ga}$ and $\boldsymbol{Gd}$ matrices were computed with markers from 751 maize singles-crosses.

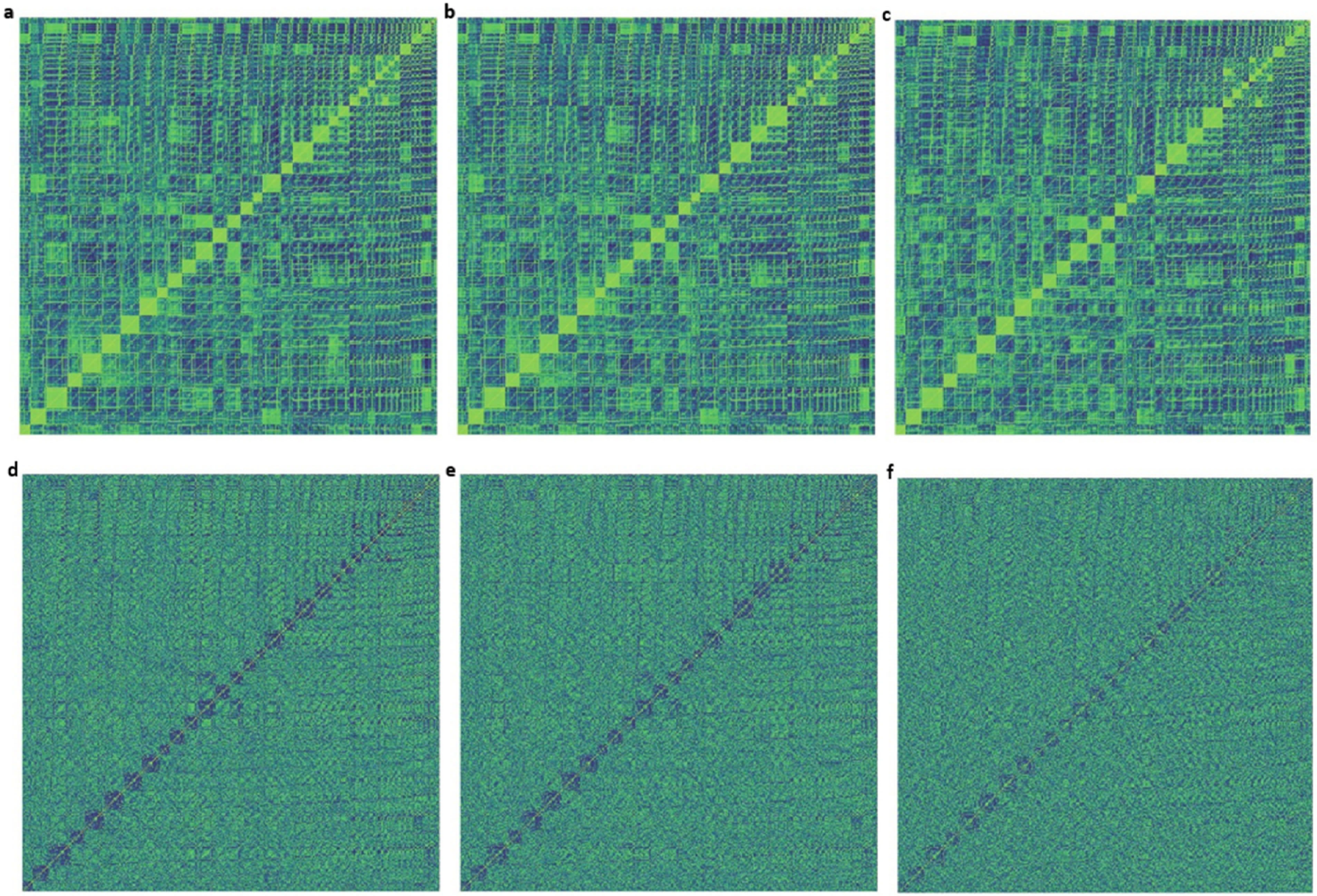

Figure 7 Heatmaps of the (a, b and c) additive genomic relationship ($\boldsymbol{Ga}$), and (d, e and f) dominance genomic relationship ($\boldsymbol{Gd}$) matrices estimated from (a and d) SNP-array, (b and e) GBS-B73, and (c and f) GBS-Mock SNP datasets for 751 tropical maize single crosses. Lines and columns of each plot were clustered according to the Euclidian distance performed in the genomic relationship matrices from the SNP-array dataset.

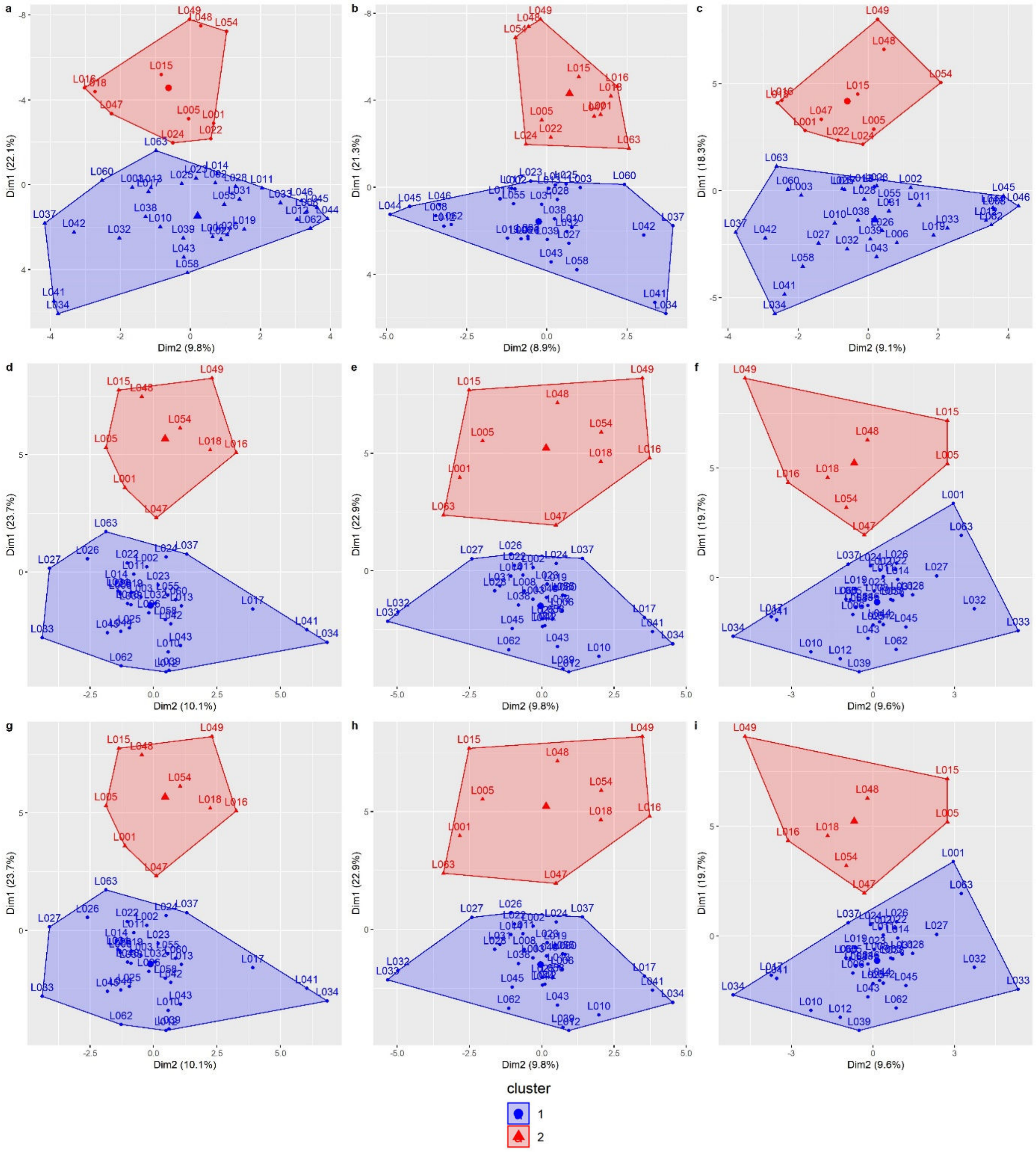

Figure 8 Heterotic groups among the 45 tropical parental lines for all traits **a** SNP-array (GY), **b** GBS-B73 (GY), **c** GBS-Mock (GY), **d** SNP-array (PH), **e** GBS-B73 (PH), **f** GBS-Mock (PH), **g** SNP-array (EH), **h** GBS-B73 (EH), and **i** GBS-Mock (EH). GY: grain yield; PH: plant height: EH: ear.

Table 5 Coincidence in dividing the parental lines into heterotic groups among datasets.

		GBS-B73	GBS-Mock
GY	**SNP-array**	0.94**	1.00**
	GBS-B73	-	0.94**
PH	**SNP-array**	0.94**	0.93**
	GBS-B73	-	0.87**
EH	**SNP-array**	0.94**	0.93**
	GBS-B73	-	0.87**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Significant at the 0.01 probability level by the t-test.

Table 6 The best tester per trait, dataset, and heterotic group

		Testers	
		HP 1	**HP 2**
GY	**SNP-array**	L023	L006
	GBS-B73	L023	L006
	GBS-Mock	L023	L006
PH	**SNP-array**	L001	L003
	GBS-B73	L001	L003
	GBS-Mock	L001	L003
EH	**SNP-array**	L001	L003
	GBS-B73	L001	L003
	GBS-Mock	L001	L003

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

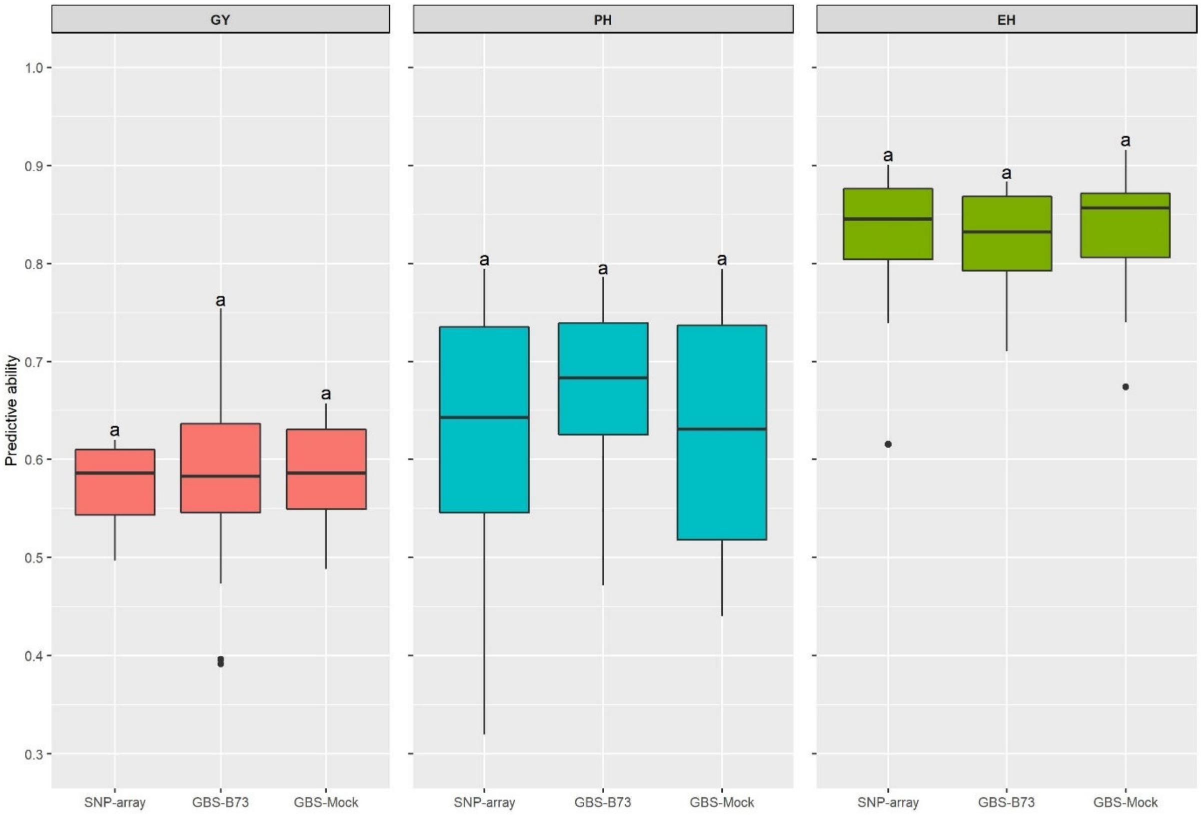

Figure 9 Predictive ability via additive-dominant GBLUP model from SNP datasets (SNP-array, GBS-B73, and GBS-Mock).

Different letters indicate significant group differences (post hoc nonparametric Tukey's test, P < 0.05).
GY: grain yield; PH: plant height: EH: ear.

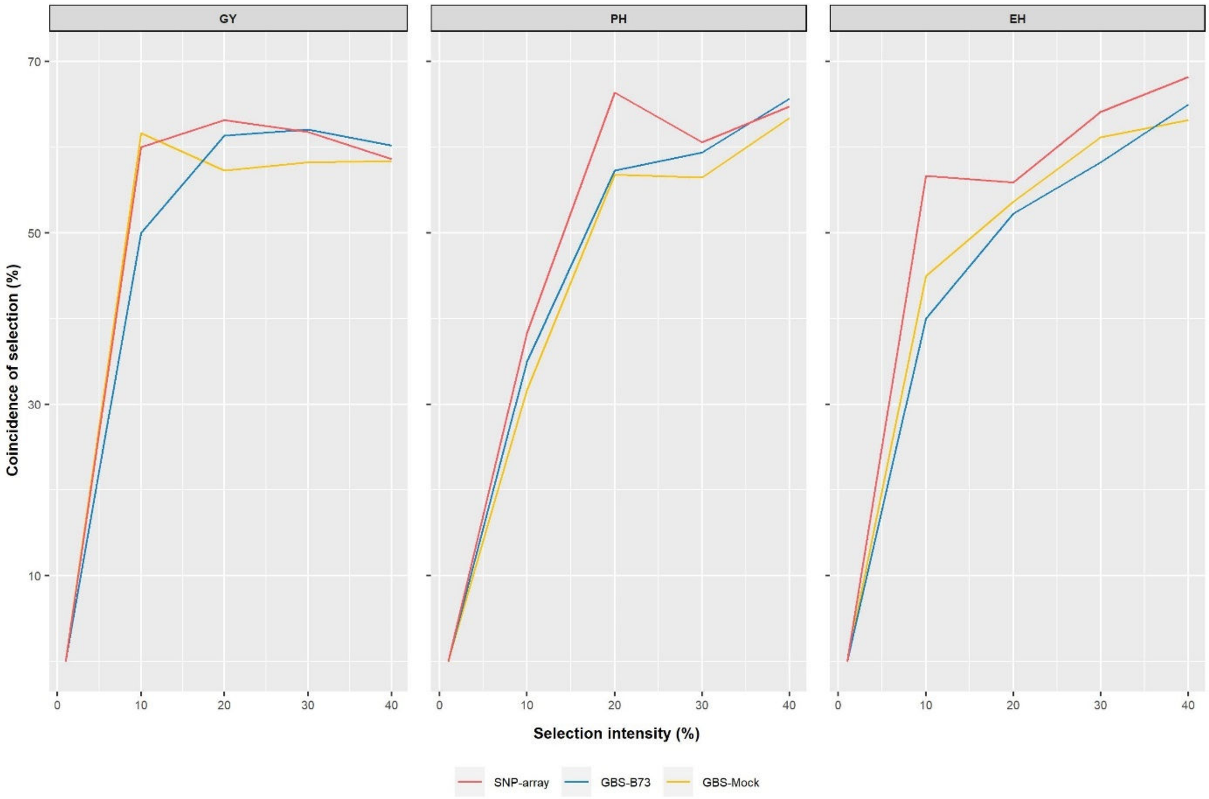

Figure 10 Coincidence of selection according to the genomic prediction model. Coincidence of selection percentage (y-axis) over a series of continuous selection intensities (1-40%) (x-axis).

Color lines represent the SNP datasets (SNP-array, GBS-B73, and GBS-Mock).

GY: grain yield; PH: plant height: EH: ear.

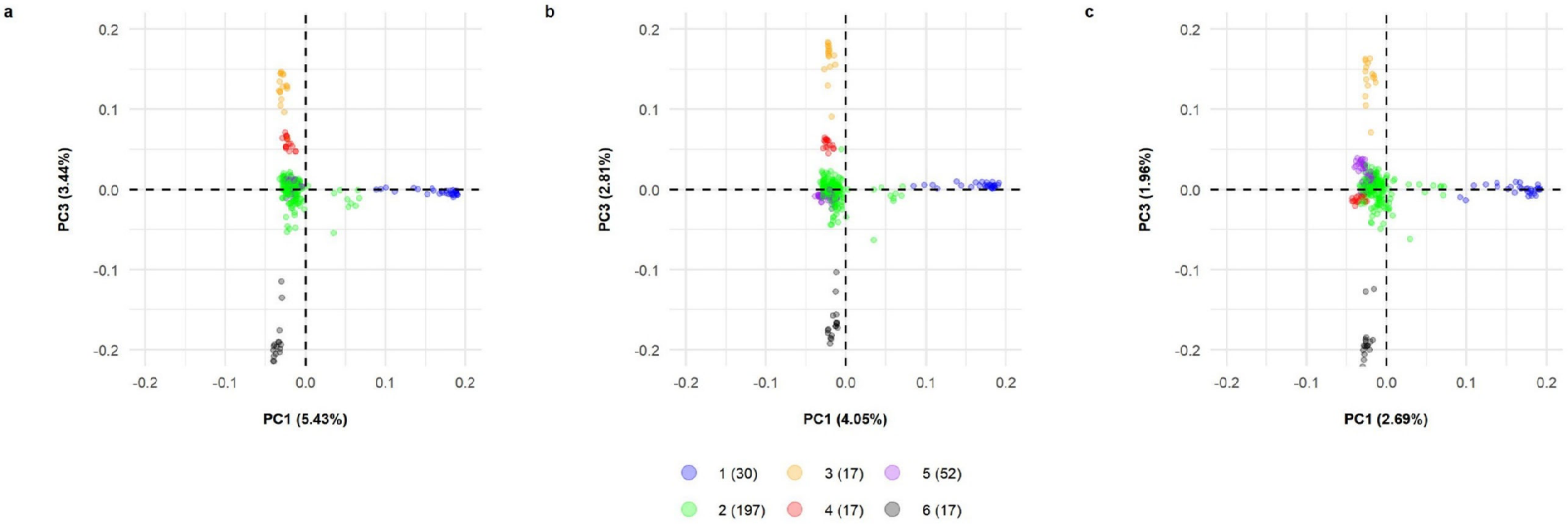

Figure S1 Bi-plot the first and third principal components using all datasets for 330 tropical parental lines a SNP-array; b GBS-B73 and c GBS-Mock. Explained variance percentages of each principal component are in parentheses. Clusters were used to color-coded parental lines.

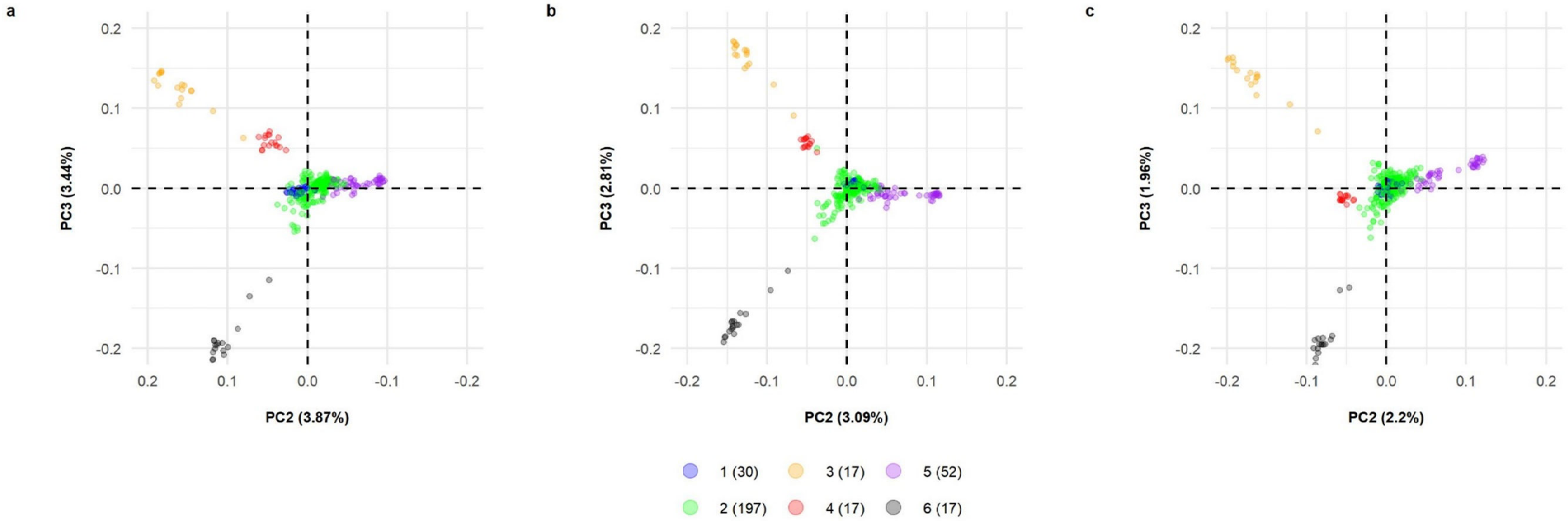

Figure S2 Bi-plot the second and third principal components using all datasets for 330 tropical parental lines a SNP-array; b GBS-B73 and c GBS-Mock. Explained variance percentages of each principal component are in parentheses. Clusters were used to color-coded parental lines.

Table S1 Variance components and genomic heritability of traits from SNP datasets

		GY	PH	EH
	SNP-array	0.14	34.27	31.55
$\hat{\boldsymbol{\sigma}}_{\mathbf{a}}^{\mathbf{2}}$	**GBS-B73**	0.15	30.36	31.34
	GBS-Mock	0.14	35.82	32.49
	SNP-array	0.07	5.83	1.95
$\hat{\boldsymbol{\sigma}}_{\mathbf{d}}^{\mathbf{2}}$	**GBS-B73**	0.09	7.88	3.08
	GBS-Mock	0.08	3.62	1.01
	SNP-array	0.17	17.92	6.70
$\hat{\boldsymbol{\sigma}}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$	**GBS-B73**	0.16	18.02	7.03
	GBS-Mock	0.16	20.55	7.61
	SNP-array	0.56	0.69	0.83
H^2	**GBS-B73**	0.60	0.68	0.83
	GBS-Mock	0.57	0.66	0.81
	SNP-array	0.36	0.59	0.78
h^2	**GBS-B73**	0.37	0.54	0.76
	GBS-Mock	0.37	0.60	0.79

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

GY: grain yield; PH: plant height: EH: ear.

$\hat{\sigma}_a^2, \hat{\sigma}_d^2$ and $\hat{\sigma}_\varepsilon^2$: Additive, dominance and residual variances, respectively: H^2: broad-sense heritability; h^2: narrow-sense heritability.

Table S2 SCA's correlation of the lines between the SNP datasets

		GBS-B73	GBS-Mock
GY	SNP-array	0.97**	0.96**
	GBS-B73	-	0.97**
PH	SNP-array	0.97**	0.96**
	GBS-B73	-	0.97**
EH	SNP-array	0.97**	0.96**
	GBS-B73	-	0.97**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Significant at the 0.01 probability level by the t-test.

Table S3 GCA's correlation of the lines between the SNP datasets

		GBS-B73	GBS-Mock
GY	SNP-array	0.99**	1.00**
	GBS-B73	-	0.99**
PH	SNP-array	1.00**	1.00**
	GBS-B73	-	1.00**
EH	SNP-array	1.00**	1.00**
	GBS-B73	-	1.00**

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock: genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

** Significant at the 0.01 probability level by the t-test.