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

Haploid maize seeds prediction using deep learning and using mock reference genomes for genomic prediction of hybrids

José Felipe Gonzaga Sabadin

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

Piracicaba
2021

José Felipe Gonzaga Sabadin
Agronomist

Haploid maize seeds prediction using deep learning and using mock reference genomes for genomic prediction of hybrids

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2021

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

Sabadin, José Felipe Gonzaga

Haploid maize seeds prediction using deep learning and using mock reference genomes for genomic prediction of hybrids / José Felipe Gonzaga Sabadin. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

72 p.

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

1. Duplo haploide 2. *R1-navajo* 3. Rede neural convolucional 4. Polimorfismo de nucleotídeo único 5. Seleção genômica I. Título

DEDICATÓRIA

Aos meus filhos, Heitor e Beatriz,
e à minha amada esposa Paula

AGRADECIMENTOS

Primeiramente, agradeço à Deus por mais uma conquista, pelas bençãos recebidas e pela insistência de conduzir-me no caminho que sempre almejei.

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ / USP) pela minha formação acadêmica e por todo o suporte durante o período em que estive aqui.

Ao meu orientador Professor Dr. Roberto Fritsche Neto por ter me recebido em seu laboratório e aceitado me orientar. Agradeço a oportunidade, paciência, conselhos e ensinamentos que fizeram-me compreender a verdadeira atividade científica. Sou muito grato!

Ao Professor Dr. Antônio Augusto Franco Garcia que me recebeu no início da minha estadia no Departamento de Genética. A sua sinceridade, apoio e compreensão foram fundamentais para eu realinhar meus objetivos.

Aos meus filhos que me ensinam diariamente que a essência é mais importante que a forma. Estar com vocês é a experiência mais engrandecedora de todas. Obrigado por iluminarem os meus dias e torná-los mais felizes!

À minha esposa Paula pelo amor, carinho, paciência e compreensão. Obrigado pelo incentivo desde o início da nossa jornada e pela família que estamos construindo. Sem você, eu não teria conseguido. Tenha certeza de que essa conquista também é sua!

Aos meus pais pelo amor, dedicação, carinho e apoio incondicional. Os ensinamentos que me deram foram e sempre serão importantes. Sou eternamente grato!

À Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro durante o período de doutorado.

Aos meus colegas do Laboratório de Genética e Melhoramento de Plantas Alógamas pelas discussões e conselhos que aprimoraram a minha formação. Agradeço ao convívio e amizade que tornaram esses dias mais alegres.

Por fim, agradeço a todos que contribuíram para que esse trabalho fosse possível.

CONTENTS

RESUMO

Predição de sementes haploides de milho usando aprendizado profundo e utilização de genomas mock para a predição genômica de híbridos

A predição é um conceito chave para o melhoramento animal e de plantas. Estimativas acuradas dos valores fenotípicos e genéticos são fundamentais para a seleção dos melhores genótipos. Por isso, diversas ferramentas vêm sendo empregadas com o objetivo de melhorar a precisão dessas estimativas, desde marcadores moleculares, usados para acessar a informação genética, até a fenotipagem de alto rendimento, usada para aumentar o tamanho da amostra e a precisão fenotípica. Nesse trabalho, nós apresentamos dois estudos envolvendo o uso de diferentes abordagens e ferramentas no processo de predição. No primeiro capítulo, nós apresentamos um estudo envolvendo o uso de *deep learning* e imagens para a fenotipagem de sementes. Nele, nós construímos um modelo de rede neural convolucional (CNN) para classificar imagens de sementes haploides putativas e verdadeiras de milho baseadas no fenótipo *R1-nj*. Nossos resultados mostram que o modelo CNN foi capaz de classificar as sementes putativas com elevada acurácia (97%). No entanto, o modelo não conseguiu detectar as sementes haploides verdadeiras. Por fim, nós disponibilizamos à comunidade científica um modelo CNN treinado e com alta acurácia para classificar sementes haploides de milho. No último capítulo, nós estudamos a utilização de genomas mock para a descoberta de marcadores e o seu efeito sobre estimativas de diversidade genética e predição genômica de híbridos. Além disso, nós os comparamos com marcadores SNP oriundos de um SNP-array e *genotyping-by-sequencing* (GBS) ancorado no genoma de referência B73. Nossos resultados mostram que a utilização de genomas mock entrega estimativas comparáveis às plataformas padrão, quando consideramos caracteres simples e efeitos aditivos. No entanto, para caracteres complexos e para os efeitos de dominância as estimativas foram um pouco piores. Nós acreditamos que esses trabalhos adicionam conhecimento relevante para a predição fenotípica e genômica aplicado ao melhoramento vegetal.

Palavras-chave: Duplo haploide, *R1-navajo*, Rede neural convolucional, Polimorfismo de nucleotídeo único, Seleção genômica

ABSTRACT

Haploid maize seeds prediction using deep learning and using mock reference genomes for genomic prediction of hybrids

Prediction is a key concept for animal and plant breeding. Accurate estimates of phenotypic and genetic values are crucial for the selection of the best genotypes. For this reason, several tools have been used to improve the accuracy of these estimates, from molecular markers, used to access genetic information, to high-throughput phenotyping, used to increase sample size and phenotypic precision. Here, we present two studies involving the use of different approaches and tools in the prediction process. First, we describe a study using deep learning and images for seed phenotyping. We built a convolutional neural network (CNN) model to classify images from putative and true haploid maize seeds based on the *R1-nj* phenotype. Our results reveal that the CNN model could classify putative haploid maize seeds with high accuracy (97%). However, the CNN model was unable to recognize true haploid seeds. Finally, we provide a highly accurate and trained CNN model to the scientific community to classify haploid maize seeds via *R1-nj*. In the latter, we studied using mock genomes to discover markers and their effect on estimates of genetic diversity and genomic prediction of hybrids. Moreover, we compared them with SNP markers from SNP-array and genotyping-by-sequencing (GBS) scored in the reference genome B73. Our results show that using mock genomes delivers estimates comparable to standard platforms when considering simple traits and additive effects. However, for complex traits and dominance effects, the estimates were slightly worse. We believe that these studies provide relevant knowledge for the phenotypic and genomic prediction applied to plant breeding.

Keywords: Doubled haploid, *R1-navajo*, Convolutional neural network, Single nucleotide polymorphism, Genomic selection

1. INTRODUCTION

Plant breeders seek, via artificial selection, to select more adapted and productive genotypes. However, for genetic progress occurs, it is essential to identify genotypes that have favorable alleles for the desirable trait. For that, breeders evaluate thousands of individuals by their phenotype, which results from the effects of the genotype, environment, and the interaction between them (Falconer & Mackay, 1996). This complex condition requires that experimental designs be used to predict genetic values accurately and phenotypes when desired. Therefore, accurate estimates of these phenotypes and their genetic values drive the selection (Bernardo, 2020).

New strategies and breeding schemes to accelerate and increase the rate of genetic gain are increasingly critical since conventional methods have shown low rates to reach population growth (Rasheed et al., 2017; Ray et al., 2012). Several tools are being implemented to improve genetic values predictions, from using molecular markers to assess individuals' genetic information to using high-throughput phenotyping to increase the sample size and accuracy of phenotypic estimates.

Molecular tools, such as genomic selection, can provide solutions to improve the response to selection. Genotyping platforms capable of delivering a large number of markers, quickly and at low cost have allowed the development and application of genomic selection to better genetic resources exploitation and reduce the phenotype-genotype gap (Rasheed et al., 2017; Varshney et al., 2016). Some studies have shown genomic selection's potential to leverage long-term genetic gains (Gorjanc et al., 2018). Furthermore, many approaches showed to be relevant to increase the predictive ability of genomic prediction models, including training set designs (Fristche-Neto et al., 2018; Isidro et al., 2015), incorporation of genotype by environment interaction (Bandeira e Sousa et al., 2017; Burgueño et al., 2012), and multi-trait genomic models (Lyra et al., 2017). However, a lack of information on how molecular markers, obtained from different platforms, affects genomic prediction of hybrids.

In current years, high-throughput phenotyping has stood out as an ally in increasing the genetic gain of crops (Araus & Cairns, 2014; Furbank & Tester, 2011). Due to the large-scale availability of genomic information, as seen above, the bottleneck is on phenotypic information (Voss-Fels et al., 2019). Several tools have been developed to increase the number of phenotyped individuals, quickly and accurately, which allows increasing the sample size and the number of replicates. Deep-learning models associated with image acquisition have gained prominence since features can be learned automatically from the input data, facilitating the development of applications (Jiang & Li, 2020). These tools can optimize the use of resources, increase the sample size, and automate tasks. Increasing the sample size is essential, because if the number of events assessed increases, the probability of detecting the desired event either (Bernardo, 2010).

Therefore, tools and new approaches that accelerate and improve the best genotypes' predictions are highly desired and directly impact society. Here, two studies involving the concept of prediction will be presented in the following chapters. In the former, the objective was to develop a

convolutional neural network (CNN) model to predict putative and true haploid maize seeds based on images. For this, we used individual seed images from induction crosses, where each seed was classified according to the *R1-nj* marker and then sowing in the field to confirm its haploidy. We built and trained a CNN model and evaluated its performance to classify putative and true haploid seeds. In the latter chapter, we evaluated using mock genomes to assess diversity and population structure of parental lines and estimate the performance for hybrids' genomic prediction. Furthermore, SNP datasets from the array and GBS-based platforms were evaluated to their performance in the genomic prediction of hybrids using additive and additive-dominant models.

References

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, *19*(1), 52–61. https://doi.org/10.1016/j.tplants.2013.09.008

Bandeira e Sousa, M., Cuevas, J., Couto, E. G. de O., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., Burgueño, J., & Crossa, J. (2017). Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3: Genes, Genomes, Genetics*, *7*(6), 1995–2014. https://doi.org/10.1534/g3.117.042341

Bernardo, R. (2010). Breeding for Quantitative Traits in Plants. Stemma Press.

Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*. https://doi.org/10.1038/s41437-020-0312-1

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Falconer, D. S., & Mackay, T. F. C. (1996). Introduction to Quantitative Genetics (Fourth Edition). In *Trends in Genetics*.

Fristche-Neto, R., Akdemir, D., & Jannink, J. L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theoretical and Applied Genetics*, *131*(5), 1153–1162. https://doi.org/10.1007/s00122-018-3068-8

Furbank, R. T., & Tester, M. (2011). Phenomics - technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*(12), 635–644. https://doi.org/10.1016/j.tplants.2011.09.005

Gorjanc, G., Gaynor, R. C., & Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, *131*(9), 1953–1966. https://doi.org/10.1007/s00122-018-3125-3

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *128*(1), 145–158. https://doi.org/10.1007/s00122-014-2418-4

Jiang, Y., & Li, C. (2020). Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review. *Plant Phenomics*, *2020*, 1–22. https://doi.org/10.34133/2020/4152816

Lyra, D. H., de Freitas Mendonça, L., Galli, G., Alves, F. C., Granato, Í. S. C., & Fritsche-Neto, R. (2017). Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Molecular Breeding*, *37*(6). https://doi.org/10.1007/s11032-017-0681-1

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., & He, Z. (2017). Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant*, *10*(8), 1047–1064. https://doi.org/10.1016/j.molp.2017.06.008

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., & Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nature Communications*, *3*, 1–7. https://doi.org/10.1038/ncomms2296

Varshney, R. K., Singh, V. K., Hickey, J. M., Xun, X., Marshall, D. F., Wang, J., Edwards, D., & Ribaut, J. M. (2016). Analytical and Decision Support Tools for Genomics-Assisted Breeding. *Trends in Plant Science*, *21*(4), 354–363. https://doi.org/10.1016/j.tplants.2015.10.018

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669–686. https://doi.org/10.1007/s00122-018-3270-8

2. IMPROVING THE IDENTIFICATION OF HAPLOID MAIZE SEEDS USING CONVOLUTIONAL NEURAL NETWORKS

ABSTRACT

A critical step towards the success of the doubled haploid (DH) technique is the haploid identification within induction crosses. The *R1-nj* marker is the principal mechanism employed in this task enabling the selection of haploids at the seed stage. Even though it seems easy to identify haploid seeds, this task is performed manually by visual classification, which becomes an inefficient process in terms of time and labor. Also, differential phenotypic expression of the *R1-nj* marker results in high rates of false positives among haploid seeds. For the first time, an image-based convolutional neural network (CNN) was trained to identify true positives among putative haploid seeds. The experiment was conducted using 3,000 maize seeds from induction crosses classified as haploid (1,000), diploid (1,000), and inhibited (1,000) class. Images were taken from each seed, and then seeds were planted in the field to confirm their ploidy. Regarding the putative haploids, the CNN accuracy average was 94.39%, for the haploid class was 97.07% and for the diploid class was 91.71%. However, the CNN model was unable to distinguish true haploid seeds among the putative haploid class, which indicates that CNN did not recognize different patterns between them. Finally, we provide a highly accurate and trained CNN model to the scientific community to classify haploid maize seeds via *R1-nj*, which can support maize breeders to optimize DH pipelines, mainly for small breeding programs with limited resources.

2.1. Introduction

As an open-pollinated crop, maize (*Zea mays* L.) breeding consists of obtaining hybrids from crosses between inbred lines. The conventional development of new lines is slow and expensive because it relies on six to eight generations of selfing to reach a high homozygosity level. An efficient alternative way to accelerate maize line development is the DH technique (Prasanna et al., 2012), which consists of three stages: *i)* obtaining haploid individuals, *ii)* identification of haploids within induction crosses, *iii)* artificial chromosomal doubling of haploid seedlings (Melchinger et al., 2014; Prigge et al., 2011). In the DH pipeline, about 1 to 3% of seeds become DH lines, as current haploid inducers produce about 10% of haploid in its crosses, and duplication and growth stages have a 10-30% of success rate, depending on germplasm source (Prasanna et al., 2012).

There are several techniques to identify haploids within induction crosses, such as molecular markers (Belicuas et al., 2007), flow cytometry (Couto et al., 2013), phenotypic markers such as seed expressed anthocyanin color marker, *R1-nj* (Nanda & Chase, 1966), seed oil-content (Melchinger et al., 2013), red root markers (Chaikam et al., 2016), purple sheath (Prigge et al., 2012; Röber et al., 2005),

and seedling traits (Chaikam et al., 2017). The *R1-nj* marker is widely used for haploid identification and is present in all haploid inducers worldwide (Chaikam et al., 2019; Prasanna et al., 2012).

The *R1-nj* gene is involved in the anthocyanin biosynthetic pathway conditioning purple coloration in the aleurone layer of endosperm and the scutellum of the embryo of seeds, resulting in the *Navajo* phenotype (Nanda & Chase, 1966). Thus, haploid seeds exhibit endosperm colored by purple and embryo without purple color, whereas diploid seeds display purple color in both regions (Nanda & Chase, 1966). This phenotypic difference between haploid and diploid seeds facilitates their visual classification. The main advantage of using the *R1-nj* marker would be detecting the haploid in the seed stage, saving resources, labor, and time. However, the *R1-nj* is affected by the genetic background of source germplasm following from differential expressiveness and inhibition of anthocyanin expression by inhibitor genes, mainly in tropical germplasm (Chaikam et al., 2015; Kebede et al., 2011), which results in high rates of false positives (Chaikam et al., 2016; Melchinger et al., 2014; Prigge et al., 2011).

Several techniques have been used to automate the sorting of induction seeds based on the *R1-nj* expression using hyperspectral (Wang et al., 2018), multispectral images (De La Fuente et al., 2017), fluorescence imaging techniques (Boote et al., 2016), and computer vision methods (Altuntaş et al., 2019; Veeramani et al., 2018). Deep-learning is an area of machine learning that has been widely used associated with image-based tasks such as object detection, semantic segmentation, and image classification (Ubbens & Stavness, 2017). The main advantage of deep learning is that features can be learned automatically from the input data, facilitating various applications (Jiang & Li, 2020). Convolutional Neural Networks (CNN) are commonly used among deep learning methods because of their ability to handle large datasets and integrate an image feature extraction step along with neural network classification in a single fully trainable pipeline (Ubbens & Stavness, 2017). These attributes are desirable for tasks involving image processing. Several studies have successfully applied image-based CNN models for plant species identification using leaf images (Grinblat et al., 2016; Lee et al., 2015), plant disease detection and diagnosis system (Ferentinos, 2018; Mohanty et al., 2016), and high-throughput phenotyping (Ampatzidis & Partel, 2019).

All previous studies that used image analysis based on the *R1-nj* marker employed visually manual scores (*R1-nj* phenotype) to train machine learning models (Altuntaş et al., 2019; Veeramani et al., 2018), disregarding false positives. Furthermore, due to the difficulty to obtain seed images dataset from induction crosses, the test of CNN models is performed using the same dataset employed during the training phase, which reduces its transferability evaluation. We have taken a step forward on the haploid maize seed classification by machine learning algorithms, checking if this method can distinguish true haploid among putative haploid seeds scored by the *R1-nj* marker. Furthermore, we tested several scenarios aimed to optimize the haploid maize seeds classification by CNN models. Our objectives were to (i) build and train a CNN model to distinguish haploid maize seeds from induction crosses based on *R1-nj* marker, (ii) verify if CNN models can discriminate true haploids among putative

haploid seeds scored by *R1-nj* marker, and (iii) verify if adding embryo-down images and inhibited seeds can optimize the CNN model performance.

2.2. Materials and Methods

Our work was divided into several stages, from obtaining the maize seeds from induction crosses until their classification by CNN model (Figure 1).

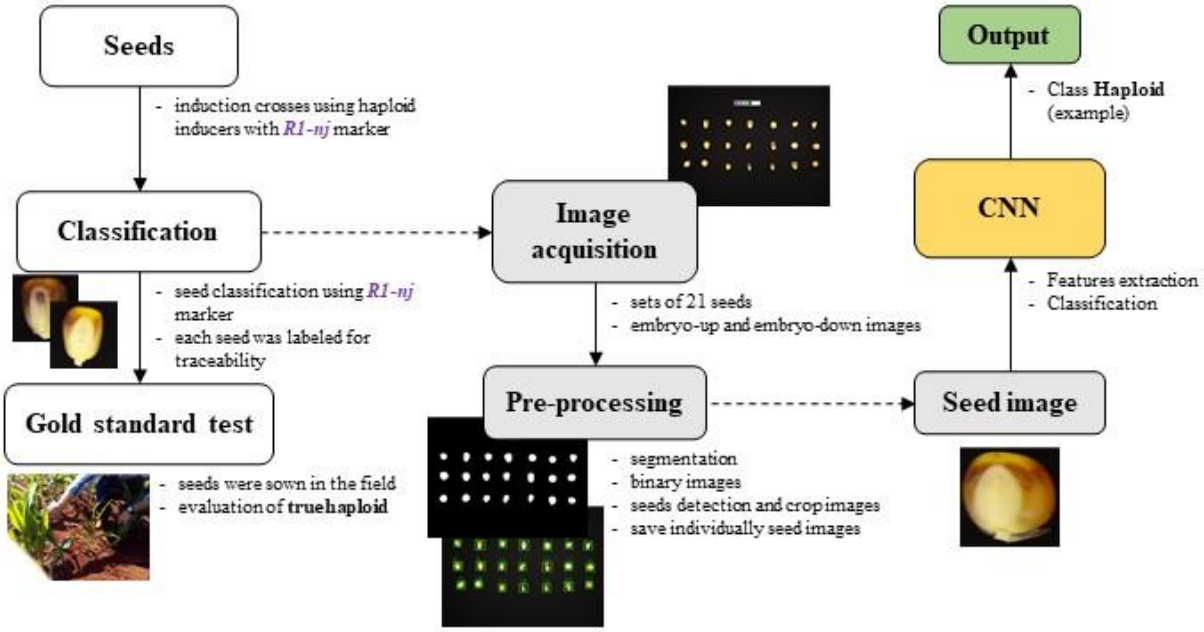

Figure 1. The overall scheme of activities involved during the training of the convolutional neural network (CNN).

2.2.1. Plant material and field trials

To obtain seeds from induction crosses marked by *R1-nj*, we used a haploid inducer population derived from a cross of two inducer lines (W23 and Stock6) with a maize hybrid adapted to tropical conditions. This inducer population has a dominant inheritance for the *R1-nj* marker and a putative haploid induction rate equal to 1.51% (Couto et al., 2019). As donors, we used two commercial single-crosses, one from the flint and another from the dent heterotic group, where 50 unique plants from the haploid inducer were crossed with two plants from each donor. Seeds obtained from these crosses were evaluated individually and classified into putative haploid, putative diploid, or inhibited (Nanda & Chase, 1966) (Figure 2) by an expert researcher to avoid visual classification bias. A set of 1,000 seeds were selected from each class, totalizing 3,000 maize seeds. These seeds exhibited variable anthocyanin expression both in the marked area and purple color intensity. Afterward, seeds were labeled individually for traceability and sown in the field to confirm the ploidy using a gold standard

classification based on their phenotype. Each plant was visually evaluated according to its vigor and leaves and classified as haploid or diploid, 30 d after planting. Haploid plants are smaller, with narrower, erect and pale leaves than their equivalent diploids (Chaikam et al., 2016, 2017; Chase, 1964; Melchinger et al., 2013; Prigge et al., 2011). For each seed, ploidy information based on the *R1-nj* seed phenotype (putative) and the gold standard phenotype (true) were available. Field trials were carried out at the University of São Paulo, located in Piracicaba, SP, Brazil (22º42'30" S; 47º38'30" W), during the summer of 2019.

Figure 2. Examples of maize haploid seeds classified using the *R1-nj* marker. The first three (left) seeds are diploid, three central seeds are haploid, and three last (right) seeds are inhibited.

2.2.2. Image acquisition and segmentation

Previously the image acquisition, seeds were dried at approximately 13% moisture to avoid that seed moisture affecting anthocyanin expression (Rotarenco et al., 2010). The image acquisition and segmentation process were inspired by Altuntaş et al. (2019), however, with minor adjustments. First, we built a simple structure for image acquisition, where a camera and white LED lamps were positioned to the top of the box to promote homogenous light saturation. We used a 16.1 Megapixel Sony DSC-W630 digital camera set for autofocus with an aperture of F2.6, ISO 80, a shutter speed of 1/10, and a distance of 30 cm. Shots were taken with seeds grouped in sets of 21 seeds with the embryo pointing up and downwards. Thus, images have captured both sides. The background was a blackboard to facilitate image segmentation. This setup resulted in 300 RGB images with a resolution of 4608 × 3456 pixels in JPEG format, comprising 1.76 GB.

The image segmentation process aimed to crop individual seed images from the original photo. We analyzed the color distribution from the seeds and background pixels to identify the best threshold value. In the red channel, pixels with values greater than 60 were assigned to seeds otherwise for the background, resulting in a binary image in which 0 represents the background and 1 represents the seed pixels. Occasionally, background pixels were segmented as seeds and, to correct these eventualities, we used morphological transformation operations. In these binary images, we assessed the center and limits (top, bottom, left, and right) of each seed, and the longest distance, between the center and the extremes, was used to define the boundary boxes around the seed, which were employed to crop the seeds from the original image. Finally, each seed image was stored in a separate file correctly identified to maintain traceability.

In order to reduce noise in downstream analyzes, images with damaged, blurred, or misplaced seeds were removed from the dataset and remained with 825 putative haploids, 913 diploids, and 905 inhibited image seeds. However, the imbalance was corrected for ease by randomly selecting 825 images from the diploid and inhibited classes, totalizing 4,950 images (2,475 embryo-up and 2,475 embryo-down). Afterward, images were resized to 100 × 100 pixels by the bicubic interpolation method. The entire segmentation process was made using Python scripts and the OpenCV package (Bradski, 2000).

2.2.3. Convolutional Neural Network

A CNN model was built from scratch for haploid maize seed classification tasks. CNN's overall architecture was composed of four convolutional layers, each followed by max-pooling and batch-normalization layers, three fully connected layers, and an output layer (Figure 3). Each convolutional layer was composed of a 2-dimensional convolution operation with n kernels with 5×5 size, using stride and padding equal to 1, followed by a max-pooling layer with kernel size 2×2, and a batch-normalization layer. All convolutional layers have the same parameters, except the number of kernels from the convolutional operation, which were 32 (conv1), 64 (conv2), 128 (conv3), and 128 (conv4) kernels, respectively (Figure 3). After the convolutional step, three fully connected layers were used, with 2048, 512, and 128 neurons, followed by an output layer. The output layer had one or three neurons, depending on the number of classes. For binary and three output problems, we employed the sigmoid and softmax activation functions, respectively.

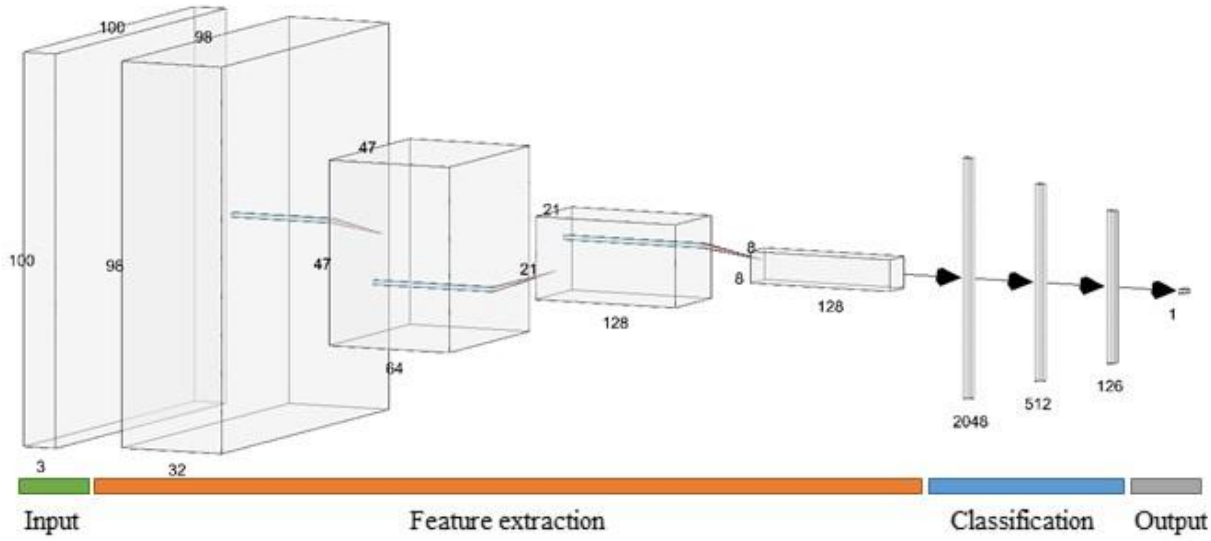

Figure 3. Convolutional Neural Network architecture built to haploid maize classification. Seed images (input) pass through all CNN layers resulting in its probability to refer to each class. The convolutional layers (feature extraction) extract useful features (shapes, edges, patterns, etc) and shrink the size to improve efficiency. After, a neural network layer (classification) outputs probability for each class. This example depicts a one output network (sigmoid activation).

During the CNN training, the *RMSprop* optimizer was applied with an initial learning rate of 0.0001, a learning rate decay of 0.5, and 100 epochs per decay. To avoid the overfitting model during training and improve the CNN performance, we applied data augmentation using rotation, flip, zoom in, zoom out, image brightness, and shearing distortions. Additionally, we implemented dropout layers (0.3) before each fully connected layer. At last, there were 1,356,417 trainable parameters, which were weighable using a batch size with 32 samples and 300 epochs. The CNN model was implemented in Python using the *TensorFlow* and *Keras* packages (Chollet, 2015).

2.2.4. CNN training and evaluation scenarios

The CNN training was divided into two steps. In the former, we created three scenarios (Scenario 1, 2, and 3) using the *R1-nj* phenotype (seed phenotype) to train the CNN model to test its ability to discriminate putative haploid maize seeds using images. Also, we investigated if adding embryo-down and inhibited seeds images can improve CNN performance. In the latter, we created a scenario (Scenario 4) using only the putative haploid seeds classified by *R1-nj* phenotype and the gold standard phenotype to train the CNN. In this scenario, we aimed to test if CNN can discriminate true haploids among putative haploid seeds. The entire work consisted of four scenarios (Figure 4).

For all scenarios, the CNN model was trained using 70% of the images as a training set and 30% as a validation set (VLS), where the classes (haploid, diploid, and inhibited given by *R1-nj* or gold standard phenotype) were assigned to both sets equally. CNN hyperparameters were tuned using embryo-up images from haploid and diploid seeds classified by the *R1-nj* phenotype. First, we trained the CNN model using the embryo-up images from haploid and diploid seeds (Scenario 1). Then, we tested CNN's performance in the image set provided by Altuntaş et al. (2019). This test set is independent of our dataset in terms of genetic background and consisted of 1,230 haploid and 1,770 diploid images seeds obtained by crossing a maternal haploid inducer RWS / RWK76 (Röber et al., 2005) with 107 genotypes. To predict this external dataset, we selected a random sample of 700 images from each class, totalizing 1,400 images, to compose a test set (TTS). The background from these images was set to black color. The same researcher classified our image dataset to check the classifier effect visually reclassified the TTS resulting in the RTS. The independence between the test set (TTS) and dataset used to CNN training is a powerful asset to check the CNN robustness, especially in this case, since we know that the source germplasm (Kebede et al., 2011) and the inducer (Prigge et al., 2011) influence the *R1-nj* expressiveness. Afterward, we carried out the CNN training with the embryo-up images from haploid, diploid, and inhibited seeds (Scenario 2), seeking to verify whether the inhibited class addition increased CNN's predictive ability. Later, as the *R1-nj* marker exhibits irregular anthocyanin expression in the endosperm, we test if the embryo-down images influence CNN performance. For that, the CNN model was trained with embryo-up and embryo-down images from these three classes (Scenario 3). Finally,

we tested if CNN would detect true haploids (gold standard phenotype) among putative haploid seeds. For that, embryo-up haploid seed images were assigned into two classes: true positive (true haploids) and false positives (diploid into putative haploid seeds) (Scenario 4), according to the gold standard phenotype. In Scenario 4, we had 103 true positives (TP) and 722 false positives (FP) images. CNN training was performed using a 70-30 scheme, i.e., 72 TP and 505 FP in the training set and 31 TP and 217 FP in the validation set. To handle the unbalanced dataset *class_weight* parameter in *Keras* was set with the frequency of each class. In this scenario, we trained the CNN model through 500 epochs. For all scenarios, we used the lower validation loss to select the best model weights.

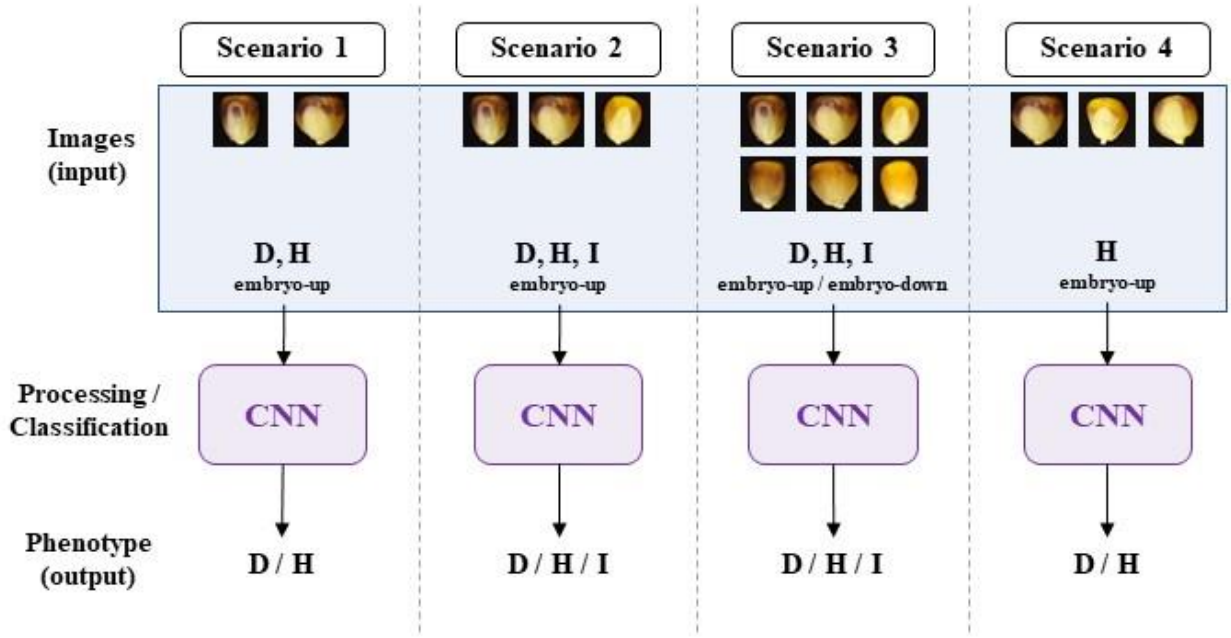

Figure 4. Scenarios for CNN training. D = diploid, H = haploid, I = inhibited, CNN = convolutional neural network. Scenario 1: Embryo-up images from haploid and diploid seeds classified by the *R1-nj* marker; Scenario 2: Embryo-up images from haploid, diploid, and inhibited seeds classified by the *R1-nj* marker; Scenario 3: Embryo-up and embryo-down images from haploid, diploid, and inhibited seeds classified by *R1-nj* marker, Scenario 4: Embryo-up haploid seed images (*R1-nj*) are divided into two classes: true positive (true haploids) and false positives (diploid into putative haploid seeds) according to gold standard phenotype.

2.2.5. CNN performance metrics

For the evaluation of the CNN model, we calculated the confusion matrix to estimate the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values for each scenario. In binary classifications, we consider the haploid class as positive. Therefore, TP and TN are the numbers of correctly predicted haploids and diploids, whereas FP and FN are incorrectly predicted haploids and diploids. Furthermore, we estimated the following metrics: accuracy, sensitivity, specificity, precision, F-score, false discovery rate (FDR), and false negative rate (FNR) (Supplemental Material Appendix). Accuracy is the most intuitive performance measure and is the percentage of

correctly predicted observations to the total samples. However, when we have unbalanced classes or one class is more important than another, this metric can take to a wrong conclusion. For that, we considered other metrics to evaluate CNN performance as sensitivity, specificity, precision, and F-score. In our study, sensitivity (true positive rate) is the number of haploid images predicted as haploid, whereas specificity (true negative rate) is the proportion of diploid images predicted as diploid. Precision is the ratio of correctly predicted haploids to the total haploid prediction, whereas F-score is the harmonic mean between precision and sensitivity. FDR is the FP proportion of the haploid predictions, whereas FNR is the FN proportion into the haploid class. Furthermore, we calculated the receiver operating characteristic (ROC) curve, which shows the true positive rate (TPR) against the false positive rate (FPR) for various threshold values. For scenarios with three classes (Scenario 2 and 3), these metrics were calculated considering each one individually.

2.3. Results

2.3.1. CNN performance using the *R1-nj* phenotype

After calibrating the hyperparameters, we conducted the CNN model's training with embryo-up images of the putative haploid and diploid classes based on the *R1-nj* phenotype (Scenario 1). In the first 100 epochs, the accuracy increased from 0.61 to 0.92. Conversely, the opposite happened with the loss curve, reducing from 0.70 to 0.16 (Figure 5). Both curves produced similar tendencies for training and validation sets, confirming that the CNN architecture effectively prevented overfitting. Following, we classified the validation set (VLS), test set (TTS), and reclassified test set (RTS) using the trained CNN model to evaluate its performance (Table 1). In the VLS, the CNN model showed sensitivity higher than the specificity value, indicating more accuracy to predict the haploid than the diploid class. Also, the FNR (2.93%) was lower than the FDR value (7.87%), indicating few putative haploids classified by the researcher were not by CNN.

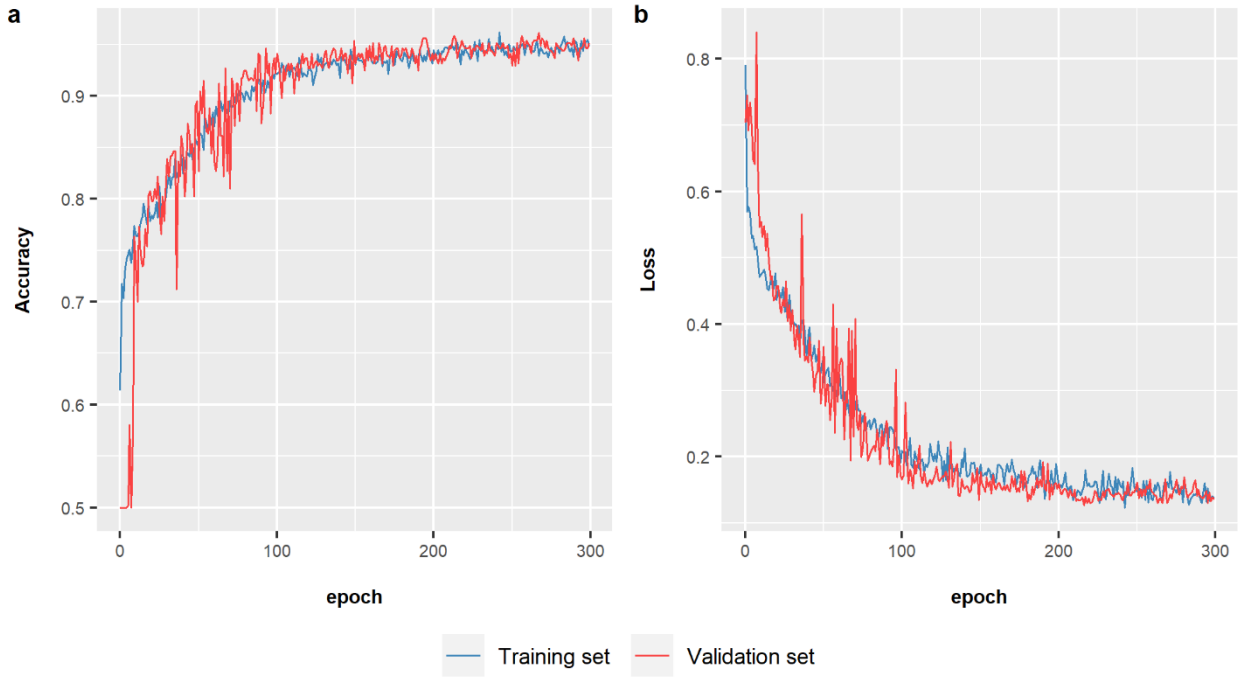

Figure 5. Curves of accuracy (**a**) and loss (**b**) over the epochs of the CNN model training considering Scenario 1. In Scenario 1, CNN was trained using embryo-up images and labels from haploid and diploid seeds classified by the *R1-nj* marker.

Table 1. CNN performance for the validation set, test set, and test set with reclassified images. Sensitivity (Se), specificity (Sp), false discovery rate (FDR), false negative rate (FNR), precision (Prec), and accuracy (Acc) are presented in percentage (%).

Sets	Se	Sp	FDR	FNR	Prec	F-Score	Acc
Validation	97.07	91.71	7.87	2.93	92.13	94.54	94.39
Test	90.43	90.86	9.18	9.57	90.82	90.62	90.64
Test (reclass.)	93.00	96.71	3.41	7.00	96.59	94.76	94.86

[1] VLS: 205 haploid and 205 diploid seed images; TTS: 700 haploid and 700 diploid seed images; RTS: 700 haploid and 700 diploid seed images

The accuracy presented by CNN when classified the TTS was lower than VLS but showed high values (90.64%) (Table 1). Also, sensitivity and specificity exhibited similar values, indicating that both classes were predicted with similar accuracy. Comparing CNN performance on TTS and RTS allows us to verify the classifier effect (Table 1). Accuracy, sensitivity, and specificity of the RTS increased by 4.66%, 2.84%, and 6.44%, respectively, compared to the TTS (Table 1). Regarding haploid predictions, FDR reduced 62.8%, whereas FNR reduced 26,8% from TTS to RTS. False positives decreased from 64 to 23, and false negatives reduced from 67 to 49, considering TTS and RTS, respectively (Figure 6). The area under the ROC curve (AUC) was 0.99 for the VLS, 0.96 for the TTS, and 0.99 for the RTS (Supplemental Figure S1).

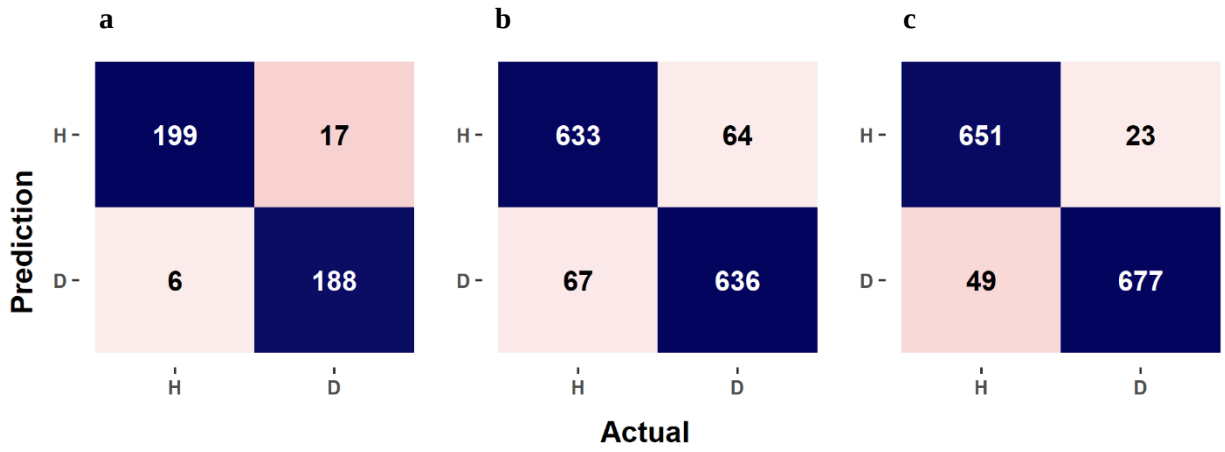

Figure 6. Confusion matrices of CNN for (a) VLS, (b) TTS, and (c) RTS. D = diploid, H = haploid, VLS = validation set, TTS = test set (Altuntas' dataset), RTS = reclassified test set (Altuntas' dataset reclassified by our dataset's classifier).

2.3.2. Effect of embryo-down images and inhibited class on CNN using the *R1-nj* phenotype

We verify the effect of combining the inhibited class on the CNN model's prediction ability through Scenario 2. In this scenario, CNN was trained using embryo-up images from three classes (haploid, diploid, and inhibited) classified by *R1-nj* phenotype. The training set's accuracy curve showed a lower tendency than the validation set from the 70 to upward (Supplemental Figure S2). Moreover, the validation set's accuracy curve depicted a high variation from 10 to 70 epochs, and the loss curve of the validation set showed a lower tendency than the training set. The average accuracy of Scenario 2 was slightly higher than Scenario 1 due to the high specificity value of the inhibited category (Table 1 and 2). However, for the haploid class, the specificity presented a value lower than Scenario 1, whereas the FNR increased by 49.8%. Furthermore, the confusion matrix (Supplemental Figure S4a) revealed that 9.26% of haploid predictions are diploid seeds. In contrast, D and I classes did not perform any classification errors.

Table 2. CNN performance for Scenario 2 and 3. Sensitivity (Se), specificity (Sp), false discovery rate (FDR), false negative rate (FNR), precision (Prec), and accuracy (Acc) are presented in percentage (%).

	Class[1]	Se	Sp	FDR	FNR	Prec	F-Score	Acc
Scenario 2	D	91.22	99.01	2.09	8.78	97.91	94.42	95.28
	H	95.61	95.12	9.26	4.39	90.74	93.11	
	I	99.02	98.71	2.40	0.98	97.60	98.30	
Scenario 3	D	91.22	98.76	2.60	8.78	97.40	94.21	94.96
	H	95.12	94.88	9.72	4.88	90.28	92.64	
	I	98.54	98.71	2.41	1.46	97.58	98.06	

[1] class: D – diploid, H – haploid, and I – inhibited

In Scenario 3, despite adding embryo-down images in CNN training, the accuracy and loss curves showed similar to Scenario 2. However, differences between these tendency curves were smaller (Supplemental Figure S3). The average accuracy and the specificity of the haploid class reduced when compared with Scenario 2. The confusion matrix was similar to Scenario 2, keeping higher confusion between diploid and haploid classes (8.3%) and without error between diploid and inhibited classes predictions (Supplemental Figure S4b).

2.3.3. CNN performance using the gold standard phenotype

Using Scenario 4, we tested if the CNN model can distinguish true haploids (gold standard phenotype) among putative haploid seeds classified by the *R1-nj*. In this scenario, the accuracy of the training set ranged from 0.46 to 0.64. Conversely, the validation set showed higher variation in the first 100 epochs, ranging from 0.12 to 0.89 (Figure 7a). The training set's loss value was close to zero during all epochs, whereas the validation set was close to 0.70 (Figure 7b). After the training process, we tested CNN's performance in the validation set, and the accuracy was 0.88, sensitivity was 0.00, and specificity was 1.00. The confusion matrix showed that the model predicted all images as diploids (most frequent class), resulting in accuracy equal to 0.88 (Figure 7c). However, the sensitivity was 0.00, which reports the model's inefficiency in distinguishing true haploids (gold standard phenotype) among putative haploid seeds classified by the *R1-nj* marker.

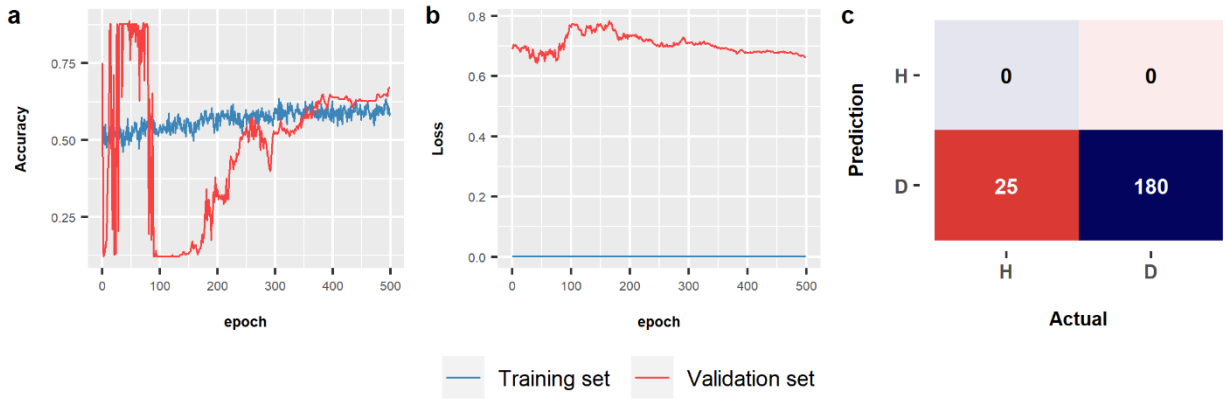

Figure 7. Curves of accuracy (a) and loss (b) over the epochs of the CNN model training, and (c) confusion matrix, considering Scenario 4. D = diploid (false positive), H = haploid (true positive). In Scenario 4, CNN was trained using embryo-up haploid seed images (R1-nj) and labels from gold standard phenotype.

2.4. Discussion

2.4.1. Classification of haploid maize seed by CNN

Deep learning methods can optimize resources and labor in several activities of plant breeding programs. Our study shows that CNN models can accurately distinguish putative haploid maize seeds from induction crosses using the *R1-nj* marker. Moreover, our CNN model showed higher accuracy for haploids than diploids (Table 1). Furthermore, 7.87% of haploid predictions were diploid seeds (FDR), whereas only 2.93% of haploid seeds were predicted as diploid (FNR). The higher FDR may be explained by the differential expressiveness of anthocyanin in the embryo region, hindering the purple color detection in this region. Regarding the low occurrence of haploid seeds within induction crosses, FNR becomes more relevant than FDR. Failure to detect haploid seeds can necessitate the sorting of a larger number of seeds, burdening the DH production. On the other hand, diploid seeds classified as haploid (FDR) have minor importance on the DH pipeline since diploids can be removed quickly in the next pipeline steps. Our CNN model presented less than 3% of undetected haploid seeds, proving its effectiveness in detect putative haploid seeds similarly when the expert researcher does.

Benchmarking with other deep learning models developed with the same objective, our CNN presents higher performance. Veeramani et al. (2018), using a DeepSort CNN model, obtained 96.8% accuracy and 91.6% sensitivity (haploid class). Altuntaş et al. (2019) evaluated some CNN architectures for haploid seeds' classification task and reached 94.2% accuracy and 94.6% sensitivity using the VGG19 model. Our CNN model is noteworthy on the haploid classification with 97.07% of sensitivity and average accuracy of 94.39%. Moreover, all previously developed models did not test their performance using a genetically unrelated dataset to the training data, limiting their transferability evaluation. We were the first to do this using the image dataset provided by Altuntaş et al. (2019) as a test set for our CNN model. The accuracy in this set (TTS) was 90.64%, with slightly higher accuracy in the diploid class than in the haploid (Table 1). However, when we reclassified the images from this set (RTS), the accuracy was 94.86%, higher for the diploid than the haploid class. This outcome suggests the influence of the classifier on the discrimination of putative haploid seed based on *R1-nj*. Since thousands of kernels are sorted in the DH lines development by different human experts, it can bring some bias that can be avoided using a CNN model for this purpose, as the CNN training is performed using only one reference.

Afterward, we verified whether including inhibited class (seeds without anthocyanin expression) could improve the CNN model's performance, significantly increasing the haploid class prediction (Scenario 2). Adding inhibited class images increased the average accuracy because CNN distinguishes them better than other categories (Table 2). Also, the CNN model did not confuse diploid and inhibited classes (Supplemental Figure S4a). This phenomenon could be explained by the absence of anthocyanin spots becoming the most distinct class. On the other hand, the haploid class sensitivity

was lower (95.61%) than Scenario 1 (97.07%), whereas the FNR increased by 49.8%, from 2.93% to 4.39%. Combining embryo-up and embryo-down images (Scenario 3) did not increase the CNN performance. For the haploid class, the accuracy and sensitivity exhibited similar values of Scenario 2, whereas the FNR was slightly lower than using only embryo-up images (Table 2). This slight enhancement can be explained by the anthocyanin expressiveness, where there may be spots on the embryo-down side that would facilitate the distinction between haploid and inhibited seeds. Some studies have used hyperspectral images (Wang et al., 2018) or NIR (Cui et al., 2019) of both sides of seeds from induction crosses to classify them into haploid and diploid seeds. These studies showed that the distinction between embryo-up and embryo-down images was more accurate than the haploid and diploid classes. Therefore, adding photos of the inhibited category or embryo-down did not improve haploid classification using the CNN model.

Given the anthocyanin expressiveness and the high rate of false positives of the *R1-nj* system (Chaikam et al., 2016; Prigge et al., 2011), seeds classified as putative haploids may have diploid plants. Our study is the first to verify CNN models' ability to detect false positives among putative haploid kernels classified by *R1-nj*. We used Scenario 4 and the gold standard phenotype to test if the CNN model can recognize any pattern able to identify false positives in putative haploid fraction. According to the gold standard phenotype, we found only 103 true positives among 825 putative haploid seeds. The high rate of false positives (722 out of 825 seeds) can be explained by the low level of improvement of the inducer population. Apart from the low putative haploid induction rate (~1.5%), anthocyanin expression was highly variable, which makes the purple color identification more difficult, especially in the embryo region. However, CNN showed a high accuracy value in Scenario 1, confirming the deep-learning model recognized the same pattern between classes as the classifier. Unfortunately, our results showed that CNN could not distinguish them. The confusion matrix showed that all images were predicted as diploids, i.e., the most frequent class (Figure 7c). The classification of true haploids among putative haploid seeds is impossible for the human eye, and our CNN model was also unable either. Therefore, to increase the anthocyanin expressiveness in haploid inducers, reducing false positives mainly in flint germplasms (Röber et al., 2005), and the correlation between true and putative haploid induction rate is a more effective strategy to overcome this issue. Furthermore, plant breeders might use tools as oil content (Melchinger et al., 2013) and red root marker (Chaikam et al., 2016) associated with the *R1-nj* marker to reduce false positives and optimize DH pipelines. However, these latter technologies are under patents, which involves high costs to acquire these inducers.

2.4.2. Applications in plant breeding

One of the main bottlenecks in the DH technique in maize is the identification of haploid seeds. Several morphological markers are used in this identification, where the most common is the *R1-nj*

system (Chaikam et al., 2019; Prasanna et al., 2012). Nevertheless, anthocyanin expression caused by *R1-nj* in embryo and endosperm tissues varies in the purple color area and intensity (Chaikam et al., 2015), which hinders the visual classification of seeds producing inconsistencies between classifiers (Prigge et al., 2011). Furthermore, the sorting of all kernels from induction crosses is visually carried out one by one by a human expert, which makes it an expensive and time-consuming activity. For this reason, the use of tools that automates and reliably perform the classification is promising, mainly for breeding programs with a limited budget.

Our CNN model showed accuracy higher than other deep-learning models in identifying the putative haploid seed from induction crosses using the *R1-nj* marker. Furthermore, our CNN was unique to be tested using images from induction crosses genetically unrelated to the training set, ensuring reliable CNN transferability estimates. That is an important test if you desire to use the CNN to classify seeds from induction crosses obtained from different inducers and source germplasm. The CNN developed by us, especially when using the RTS image dataset, presented performance close to (sometimes higher than) the values obtained by Altuntaş et al. (2019), which reveals that our model offers a possibility of use for other germplasms and breeding programs. Although the CNN failed to identify true haploids (gold standard phenotype) among putative haploid seeds (*R1-nj*), this result indicates to focus efforts on increasing the *R1-nj* expression and correlation between true and putative haploid induction rates instead to detect false positives. Moreover, if the *R1-nj* phenotypes will be easy to discriminate them, consequently, the CNN performance will increase.

CNN models for image-based high-throughput plant phenotyping have been highlighted as a new method to speed up phenotypes' acquisition. The availability to the scientific community of a trained CNN to classify haploid maize seeds by *R1-nj* can support maize breeders to optimize DH lines production, mainly for small breeding programs with limited resources. Automated tools for image acquisition of individual seeds may streamline the classification of thousands of kernels. However, to capture the embryo-up seed image is a challenge of working with maize kernels due to their irregular shape (De La Fuente et al., 2017). We built a CNN model with high accuracy, especially for the haploid class, to classify putative haploid maize seeds based on the *R1-nj* system. Lastly, we are making available the CNN trained and a basic tutorial that can be accessed by https://github.com/sabadinfelipe/DHsort.

2.5. Conclusion

Using CNN models to sort haploid maize seeds from induction crosses by inducers with the *R1-nj* system is promising. Our CNN model showed accuracy higher than other deep-learning models, especially for the haploid class (97%). Also, our CNN can be used to classify seeds from induction crosses of different genetic backgrounds. However, the CNN model cannot detect the true haploids

among the putative haploid seeds. Therefore, it seems that the development of inducers with more prominent anthocyanin expression should be a better strategy for reducing false positives. Finally, we provided a trained CNN and a basic tutorial for the scientific community supporting breeders to develop new DH lines.

References

Altuntaş, Y., Cömert, Z., & Kocamaz, A. F. (2019). Identification of haploid and diploid maize seeds using convolutional neural networks and a transfer learning approach. *Computers and Electronics in Agriculture*, *163*(40), 1–11. https://doi.org/10.1016/j.compag.2019.104874

Ampatzidis, Y., & Partel, V. (2019). UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. *Remote Sensing*, *11*(4). https://doi.org/10.3390/rs11040410

Belicuas, P. R., Guimarães, C. T., Paiva, L. V., Duarte, J. M., Maluf, W. R., & Paiva, E. (2007). Androgenetic haploids and SSR markers as tools for the development of tropical maize hybrids. *Euphytica*, *156*(1–2), 95–102. https://doi.org/10.1007/s10681-007-9356-z

Boote, B. W., Freppon, D. J., De La Fuente, G. N., Lübberstedt, T., Nikolau, B. J., & Smith, E. A. (2016). Haploid differentiation in maize kernels based on fluorescence imaging. *Plant Breeding*, *135*(4), 439–445. https://doi.org/10.1111/pbr.12382

Bradski, G. (2000). The OpenCV Library. *Dr Dobbs Journal of Software Tools*. https://doi.org/10.1111/0023-8333.50.s1.10

Chaikam, V., Lopez, L. A., Martinez, L., Burgueño, J., & Boddupalli, P. M. (2017). Identification of in vivo induced maternal haploids in maize using seedling traits. *Euphytica*, *213*(8). https://doi.org/10.1007/s10681-017-1968-3

Chaikam, V., Martinez, L., Melchinger, A. E., Schipprack, W., & Boddupalli, P. M. (2016). Development and validation of red root marker-based haploid inducers in maize. *Crop Science*, *56*(4), 1678–1688. https://doi.org/10.2135/cropsci2015.10.0653

Chaikam, V., Molenaar, W., Melchinger, A. E., & Boddupalli, P. M. (2019). Doubled haploid technology for line development in maize: technical advances and prospects. *Theoretical and Applied Genetics*, *132*(12), 3227–3243. https://doi.org/10.1007/s00122-019-03433-x

Chaikam, V., Nair, S. K., Babu, R., Martinez, L., Tejomurtula, J., & Boddupalli, P. M. (2015). Analysis of effectiveness of R1-nj anthocyanin marker for in vivo haploid identification in maize and molecular markers for predicting the inhibition of R1-nj expression. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *128*(1), 159–171. https://doi.org/10.1007/s00122-014-2419-3

Chase, S. S. (1964). MONOPLOIDS AND DIPLOIDS OF MAIZE: A COMPARISON OF GENOTYPIC EQUIVALENTS. *American Journal of Botany*, *51*(9), 928–933. https://doi.org/10.1002/j.1537-2197.1964.tb06719.x

Chollet, F. (2015). Keras: The Python Deep Learning library. *Keras.Io*.

Couto, E. G. de O., Cury, M. N., e Souza, M. B., Granato, Í. S. C., Vidotti, M. S., Garbuglio, D. D., Crossa, J., Burgueño, J., & Fritsche-Neto, R. (2019). Effect of F1 and F2 generations on genetic variability and working steps of doubled haploid production in maize. *PLoS ONE*, *14*(11), 1–16. https://doi.org/10.1371/journal.pone.0224631

Couto, E. G. de O., Davide, L. M. C., Bustamante, F. de O., Pinho, R. G. Von, & Silva, T. N. (2013). Identification of haploid maize by flow cytometry, morphological and molecular markers. *Ciência e Agrotecnologia*, *37*(1), 25–31. https://doi.org/10.1590/S1413-70542013000100003

Cui, Y., Ge, W., Li, J., Zhang, J., An, D., & Wei, Y. (2019). Screening of maize haploid kernels based on near infrared spectroscopy quantitative analysis. *Computers and Electronics in Agriculture*, *158*(February), 358–368. https://doi.org/10.1016/j.compag.2019.01.038

De La Fuente, G. N., Carstensen, J. M., Edberg, M. A., & Lü bberstedt, T. (2017). Discrimination of haploid and diploid maize kernels via multispectral imaging. *Plant Breeding*, *136*(1), 50–60. https://doi.org/10.1111/pbr.12445

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*(January), 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Grinblat, G. L., Uzal, L. C., Larese, M. G., & Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, *127*, 418–424. https://doi.org/10.1016/j.compag.2016.07.003

Jiang, Y., & Li, C. (2020). Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review. *Plant Phenomics*, *2020*, 1–22. https://doi.org/10.34133/2020/4152816

Kebede, A. Z., Dhillon, B. S., Schipprack, W., Araus, J. L., Bänziger, M., Semagn, K., Alvarado, G., & Melchinger, A. E. (2011). Effect of source germplasm and season on the in vivo haploid induction rate in tropical maize. *Euphytica*, *180*(2), 219–226. https://doi.org/10.1007/s10681-011-0376-3

Lee, S. H., Chan, C. S., Wilkin, P., & Remagnino, P. (2015). Deep-plant: Plant identification with convolutional neural networks. *2015 IEEE International Conference on Image Processing (ICIP)*, 452–456. https://doi.org/10.1109/ICIP.2015.7350839

Melchinger, A. E., Schipprack, W., Utz, H. F., & Mirdita, V. (2014). In vivo haploid induction in maize: Identification of haploid seeds by their oil content. *Crop Science*, *54*(4), 1497–1504. https://doi.org/10.2135/cropsci2013.12.0851

Melchinger, A. E., Schipprack, W., Würschum, T., Chen, S., & Technow, F. (2013). Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. *Scientific Reports*, *3*, 1–5. https://doi.org/10.1038/srep02129

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, *7*(September), 1–10. https://doi.org/10.3389/fpls.2016.01419

Nanda, D. K., & Chase, S. S. (1966). An Embryo Marker for Detecting Monoploids Of Maize (Zea Mays L.) 1. *Crop Science*, *6*(2), 213–215. https://doi.org/10.2135/cropsci1966.0011183X000600020036x

Prasanna, B. M., Chaikam, V., & Mahuku, G. (2012). Doubled haploid technology in maize breeding: theory and practice. CIMMYT.

Prigge, V., Sánchez, C., Dhillon, B. S., Schipprack, W., Araus, J. L., Bänziger, M., & Melchinger, A. E. (2011). Doubled haploids in tropical maize: I. effects of inducers and source germplasm on in vivo haploid induction rates. *Crop Science*, *51*(4), 1498–1506. https://doi.org/10.2135/cropsci2010.10.0568

Prigge, V., Schipprack, W., Mahuku, G., Atlin, G. N., & Melchinger, A. E. (2012). Development of in vivo haploid inducers for tropical maize breeding programs. *Euphytica*, *185*(3), 481–490. https://doi.org/10.1007/s10681-012-0657-5

Röber, F. K., Gordillo, G. A., & Geiger, H. H. (2005). In vivo haploid induction in maize: Performance of new inducers and significance of doubled haploid lines in hybrid breeding. *Maydica*, *50*(3), 275–283.

Rotarenco, V., Dicu, G., State, D., & Fuia, S. (2010). New inducers of maternal haploids in maize. *Maize Genetics Cooperation Newsletter*, *84*(3), 1–7.

Ubbens, J. R., & Stavness, I. (2017). Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Frontiers in Plant Science*, *8*(July). https://doi.org/10.3389/fpls.2017.01190

Veeramani, B., Raymond, J. W., & Chanda, P. (2018). DeepSort: Deep convolutional networks for sorting haploid maize seeds. *BMC Bioinformatics*, *19*(Suppl 9), 1–9. https://doi.org/10.1186/s12859-018-2267-2

Wang, Y., Lv, Y., Liu, H., Wei, Y., Zhang, J., An, D., & Wu, J. (2018). Identification of maize haploid kernels based on hyperspectral imaging technology. *Computers and Electronics in Agriculture*, *153*(February), 188–195. https://doi.org/10.1016/j.compag.2018.08.012

SUPPLEMENTAL MATERIAL

Appendix

		Actual	
		Haploid	**Diploid**
Predicted	**Haploid**	**TP**	**FP**
	Diploid	**FN**	**TN**

TP: true positive, TN: true negative, FP: false positive, FN: false negative

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}}$$

$$\boldsymbol{False\ discovery\ rate\ (FDR)} = \frac{\boldsymbol{FP}}{\boldsymbol{TP + FP}}$$

$$\boldsymbol{False\ negative\ rate\ (FNR)} = \frac{\boldsymbol{FN}}{\boldsymbol{TP + FN}}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\boldsymbol{Fscore} = \boldsymbol{2} \times \frac{\boldsymbol{Precision \times Sensitivity}}{\boldsymbol{Precision + Sensitivity}}$$

Supplementary Figures

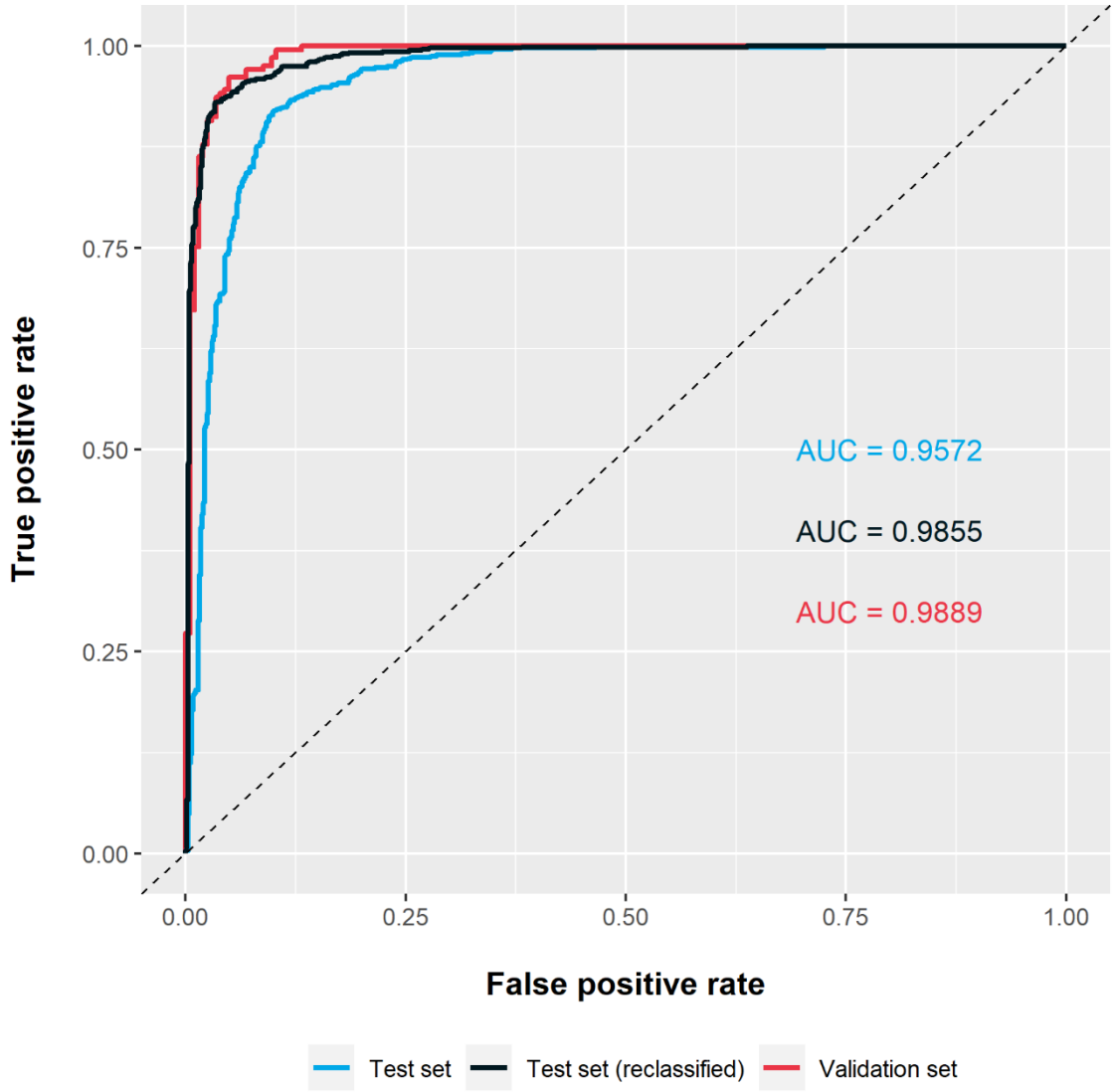

Figure S1. ROC curves and area under the curve (AUC) of validation set (VLS), test set (TTS), and reclassified test set (RTS). TTS refer to Altuntas' original dataset, whereas RTS refer to reclassified Altuntas' dataset (reclassified by our dataset's classifier).

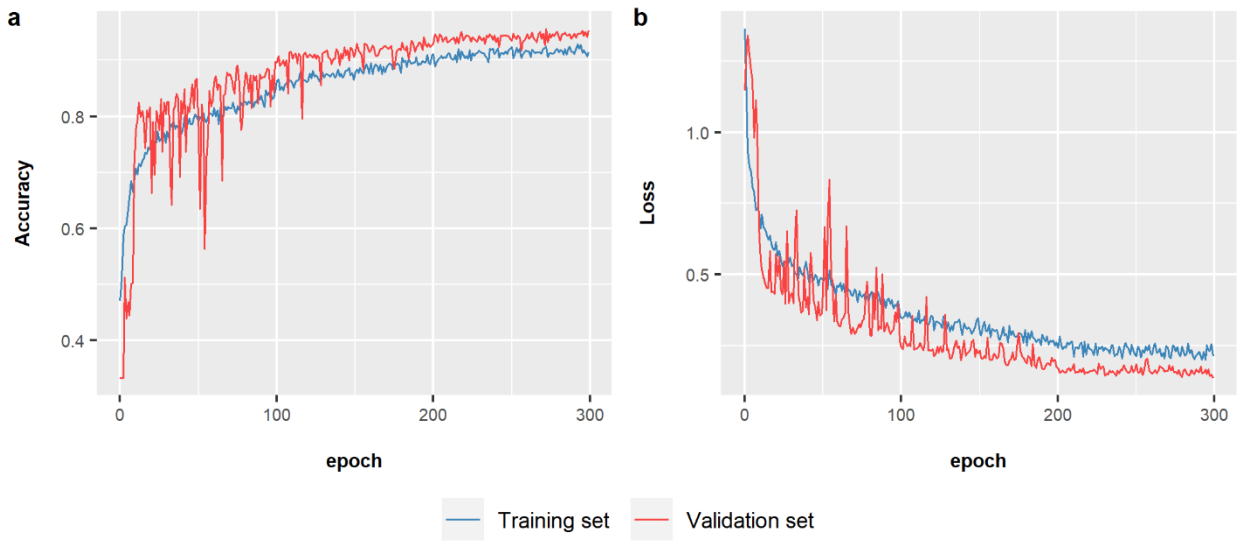

Figure S2. Curves of accuracy (**a**) and loss (**b**) over the epochs of the CNN model training considering Scenario 2. In Scenario 2, CNN was trained using embryo-up images and labels from haploid, diploid and inhibited seeds classified by the *R1-nj* marker.

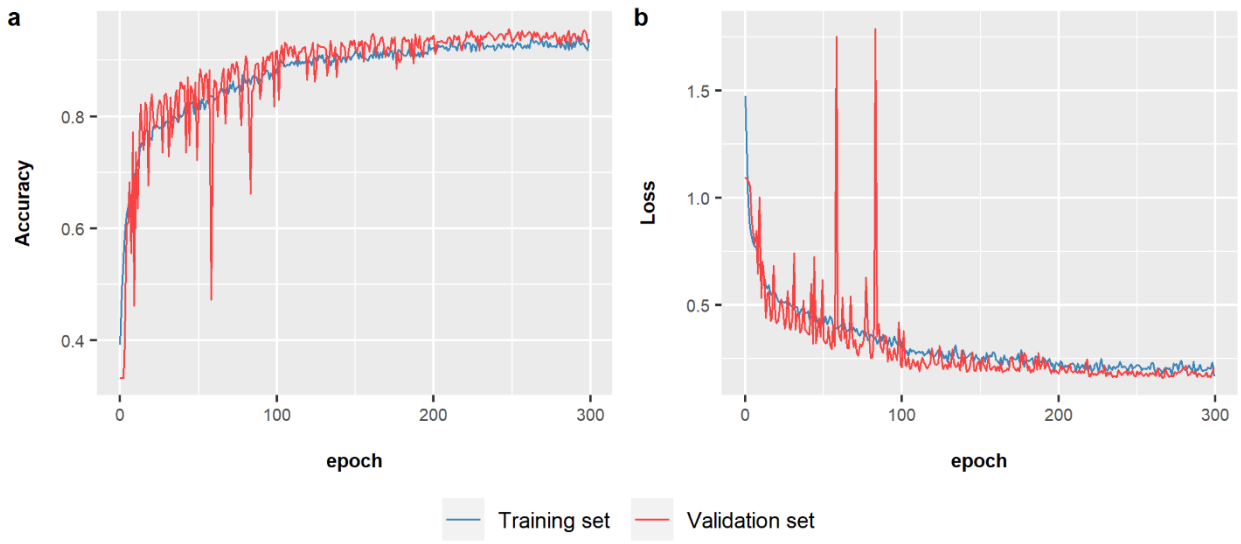

Figure S3. Curves of accuracy (**a**) and loss (**b**) over the epochs of the CNN model training considering Scenario 3. In Scenario 3, CNN was trained using embryo-up and embryo-down images and labels from haploid, diploid and inhibited seeds classified by the *R1-nj* marker.

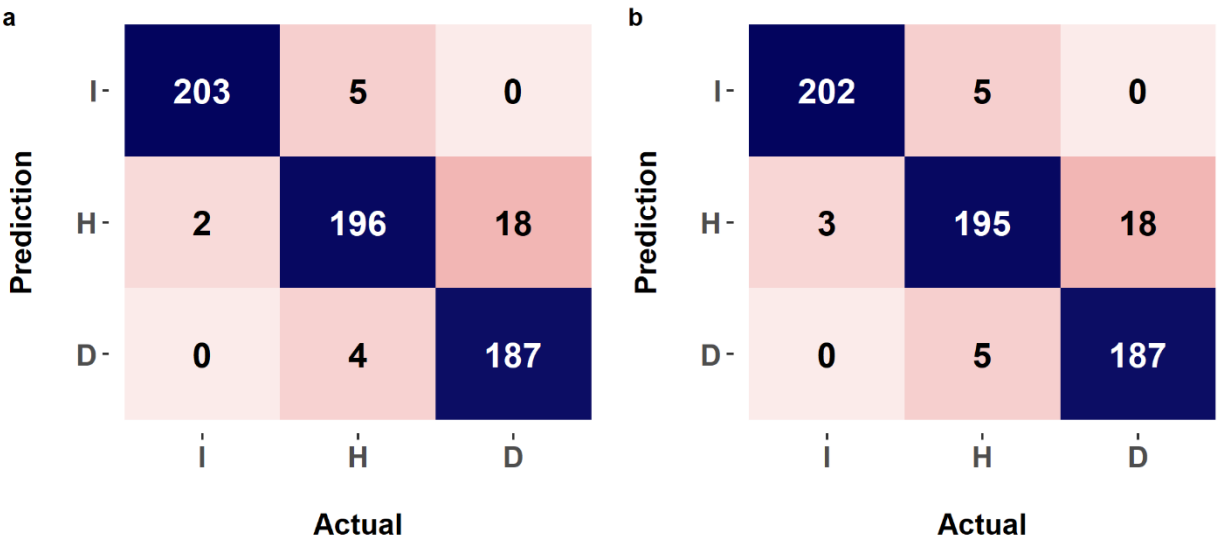

Figure S4. Confusion matrices of CNN for (**a**) Scenario 2, and (**b**) Scenario 3. D = diploid; H = haploid; and I = inhibited; Scenario 2: Embryo-up images from haploid, diploid, and inhibited seeds classified by the R1-nj marker; Scenario 3: Embryo-up and embryo-down images from haploid, diploid, and inhibited seeds classified by the *R1-nj* marker.

3. ON THE USEFULNESS OF MOCK REFERENCE GENOMES TO DIVERSITY STUDIES AND PREDICT SINGLE-CROSSES VIA ADDITIVE-DOMINANCE MODELS

ABSTRACT

Genomic prediction, based on molecular markers, enables speed breeding schemes and increases the response to selection. Even though there are several genotyping platforms for obtaining single nucleotide polymorphism (SNP) markers, lacking comparative information on how these platforms affect hybrid prediction or the inclusion of non-additive effects. Moreover, SNP discovery techniques are commonly based on a unique reference genome, which can introduce an ascertainment bias when tested germplasms are distant from reference. We employed a tropical maize single-crosses panel and genomic data from two genotyping platforms: array and genotyping-by-sequencing, both based on the B73 genome (temperate). Also, we used a pipeline to build a mock reference genome for SNP discovery aiming to capture unique SNP markers within the tropical maize population, independent from an external reference genome. Our results indicate that mock reference genomes deliver reliable estimates for genetic diversity and population structure assessment. Furthermore, genomic prediction estimates were comparable to standard approaches, especially when considering additive effects or simple traits. However, mock genomes were slightly worse to predict complex traits and estimate dominance effects, but still with similar GBS performance using B73 as the reference genome. Nevertheless, the SNP-array methods achieved the best predictive ability and reliability to estimate variance components. Finally, the mock genomes can be a worthy alternative to perform genetic diversity and genomic selection studies, especially for those species where the reference genome is not available.

Keywords: genotyping-by-sequencing, SNP-array, genomic selection, GBLUP, orphan crops

3.1. INTRODUCTION

Annually, maize (*Zea mays* L.) breeding programs develop thousands of new inbred lines. Hence, a crucial challenge for plant breeders is to assess the performance of all possible combinations (hybrids) between them (Hallauer *et al.*, 2010). Also, these hybrids must be evaluated in several locations, which makes carrying out field testing unfeasible. Thus, the application of biotechnological tools, such as molecular markers, has been highlighted in hybrid breeding. Regarding molecular makers, single nucleotide polymorphism (SNP) are abundant and uniformly distributed throughout the genome of crop species (Gupta *et al.*, 2008). The lower cost, high read accuracy, and competitive sequencing systems are increasing the availability of SNP markers for crop improvement purposes (Thomson, 2014; Kang *et al.*, 2016), especially in genetic diversity analysis (Frascaroli *et al.*, 2013; Zhang *et al.*, 2016),

genome-wide association studies (GWAS) (Morosini *et al.*, 2017; Vidotti *et al.*, 2019; Galli *et al.*, 2020), and genomic prediction (Technow *et al.*, 2012; Lyra *et al.*, 2017).

Currently, there are different types of high-throughput genotyping platforms for obtaining SNPs widely distributed in the genome. SNP arrays and next-generation sequencing (NGS) platforms are the most suitable, as they are capable of genotyping hundreds or thousands of samples with many markers (Rasheed *et al.*, 2017). There are several array-based genotyping platforms for maize (Ganal *et al.*, 2011; Unterseer *et al.*, 2014; Rousselle *et al.*, 2015; Xu, Ren, *et al.*, 2017) producing from 3,000 to 600,000 SNP markers. Array-based platforms have advantages such as the range of multiplex levels that provide rapid high-density scans and robust SNP calling with high call-rate (Rasheed *et al.*, 2017). Moreover, arrays are fixed platforms (SNP subset), which can be positive or negative depending on the purpose. On the other hand, they are expensive, making them inaccessible for small breeding programs.

SNP identification during the array development process is carried out in a discovery panel consisting of a small sample of individuals from a population (Moragues *et al.*, 2010). A limited sample of individuals allows only a fraction of polymorphisms to be discovered. When a larger sample of individuals are genotyped for these SNPs, ascertainment bias occurs (Nielsen, 2000; Clark *et al.*, 2005). Considering a small discovery panel, if the probability of identifying a distinct SNP is a function of allelic frequency, rare SNPs have a lower chance of discovery than more frequent ones (Moragues *et al.*, 2010). Therefore, if the target germplasm is genetically distant from the germplasm used for SNP discovery, ascertainment bias can be introduced (Albrechtsen *et al.*, 2010).

A widely used alternative for obtaining SNPs for genomic studies is the genotyping-by-sequencing (GBS). This approach consists of reducing the genome complexity using restriction enzymes, followed by adapter ligation, polymerase chain reaction (PCR), and sequencing (Elshire *et al.*, 2011). A further protocol modification was proposed by Poland et al. (2012) using two restriction enzymes. Both approaches can produce a large number of SNP markers with a low cost per data point. Furthermore, SNP discovery and scoring occur concurrently, minimizing ascertainment bias (Rasheed *et al.*, 2017). However, GBS typically generates a large number of low-quality markers with a high rate of missing data because DNA fragments are sequenced at low depth (Heslot *et al.*, 2013). Furthermore, the maize reference genome commonly used is a temperate inbred line B73, and, both array or re-sequencing strategies that use it, can introduce significant ascertainment bias when we analyze tropical maize germplasm (Lu *et al.*, 2009; Hirsch *et al.*, 2014; Xu, Li, *et al.*, 2017). Besides, maize is a species with high diversity caused by its dynamic genome (Hirsch *et al.*, 2014) and some structural variations often associated with phenotype variation (Wallace *et al.*, 2014; Dolatabadian *et al.*, 2017).

Consequently, both approaches have their advantages and disadvantages, impacting not only costs and procedures but also the downstream genetic analyzes. The decision of the best genotyping platform will depend on the target germplasm, the genetic study to be carried out, and the budget available for the acquisition of the genomic information. Recent studies have compared these SNP platforms on estimates of genetic diversity, GWAS, and genomic prediction in wheat (Elbasyoni *et al.*,

2018; Chu *et al.*, 2020), barley (Darrier *et al.*, 2019), and inbred maize lines (Negro *et al.*, 2019). However, all of them use inbred lines, lacking information on how these SNP platforms affect genomic studies involving hybrids, the final product of maize breeding. Furthermore, the discovery of intrinsic polymorphisms within a population without using an external genome could improve the accuracy of genetic estimates. Melo et al. (2016) developed a pipeline using bioinformatics tools and clustering strategies to build a population-tailored mock reference using the same GBS data for downstream SNP calling. A mock genome could be employed to SNP calling and perform some genomic studies (Munjal *et al.*, 2018; Matias *et al.*, 2019), which did not require a physical position in a constant reference as the genomic prediction.

In our study, we consider maize as a model species to verify how these SNP datasets can affect diversity and genomic prediction studies. For this reason, we used the phenotypic data from a panel of tropical maize hybrids in which both array and GBS platforms genotyped parental lines. Additionally, we used a pipeline to build a mock reference genome based on the GBS data (Melo *et al.*, 2016), intending to capture the polymorphism independently of an external genome to perform the SNP calling. Furthermore, we evaluated three distinct traits with different genetic architecture and heritabilities. Our objectives were to evaluate the efficiency of using the mock genome to assess the diversity and population structure of inbred lines, estimate the performance of using the mock genome in the genomic prediction of hybrids, and compare array-based and GBS-based SNPs for performance in the genomic prediction of hybrids via additive and additive-dominance models.

3.2. MATERIAL AND METHODS

3.2.1. Phenotypic data

The dataset used consists of 903 maize single-crosses derived from a full diallel mating design among 49 tropical inbred lines. These inbred lines were phenotypically characterized as flint (31), semi-flint (9), dent (2), and semi-dent (7) pools and selected based on their nitrogen use efficiency (Mendonça *et al.*, 2017). The hybrids were evaluated in an augmented block design (unreplicated trial), each block consisting of 16 unique hybrids and two checks. Field trials were performed in Anhembi (22º50’51’’S, 48º01’06’’W) and Piracicaba (22º42’23’’S, 47º38’12’’W), at São Paulo State, Brazil, during the second growing season (January to June) of 2016 and 2017. In both sites and years, the hybrids were evaluated under two nitrogen (N) application regimes, low with 30 kg N ha^{-1}, and ideal with 100 kg N ha^{-1}. The combination of site × year × N level was defined as unique environments.

Each plot consisted of 7-meter rows spaced 0.50 m with a sowing density of about of 57.000 kernels per hectare, under conventional fertilization, weed, and pest control. The traits evaluated were grain yield (GY, Mg ha^{-1}), plant height (PH, cm) and, ear height (EH, cm). Plots were manually harvested, and GY was corrected to 13% moisture. PH and EH were measured from the soil surface to

the flag leaf collar and the first ear's insertion, respectively. It was performed on five representative plants within each plot and averaged.

The joint analysis of each phenotype was performed to estimate genotype means across environments. Then, we fitted the following mixed model to obtain the best linear unbiased estimator (BLUE) for each genotype and then estimate the adjusted means across environments.

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

where $\boldsymbol{y}$ is the vector of phenotypic values of hybrids and checks; $\boldsymbol{l}$ is the vector of fixed effects of environment (combination of site × year × N level); $\boldsymbol{b}$ is the vector of random effect of block nested within environment, where $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_b^2)$; $\boldsymbol{c}$ is the vector of fixed effect of checks; $\boldsymbol{g}$ is the vector of fixed effects of hybrids; $\boldsymbol{i}$ is the vector of random effects of interaction checks × environments, where $\boldsymbol{i} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{cl}^2)$; $\boldsymbol{\varepsilon}$ is the vector of random residuals from checks and hybrids by environment effects, which are confounded in the final residual term, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{D_e})$. $\boldsymbol{Q}$, $\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are the incidence matrices for $\boldsymbol{l}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{g}$, and $\boldsymbol{i}$. We assumed an unstructured (US) covariance matrix across environments for the residual term ($\boldsymbol{D_e}$).

Additionally, the same model was fitted regarding genotype as a random effect for variance component estimation employing restricted maximum likelihood (REML/BLUP). The broad-sense heritability (H^2) was estimated at the entry mean level by $H^2 = \hat{\sigma}_g^2 / \left(\hat{\sigma}_g^2 + \hat{\sigma}_{cl}^2/e + \hat{\sigma}_\varepsilon^2/re\right)$, where $\hat{\sigma}_g^2$ is the genetic variance, $\hat{\sigma}_{cl}^2$ is the variance due to G×E interaction, $\hat{\sigma}_\varepsilon^2$ is the residual variance averaged across environments, e is the number of environments (e = 8), and r is the number of replicates (r = 1). To assess the significance of model effects, we performed the Wald test for fixed, and the likelihood ratio test (LRT) for random effects. The mixed model equations, Wald test, and LRT were performed using the *ASReml-R* package (Butler *et al.*, 2018).

3.2.2. Genotypic data

The 49 parental inbred lines were genotyped using two high-density SNP platforms: *i)* Affymetrix® Axiom Maize Genotyping array containing about 616,000 SNPs (SNP-array) (Unterseer et al., 2014) discovered by mapping whole genome sequencing reads of 30 temperate maize inbred lines against the B73 reference genome. This SNP-array was optimized for European and American temperate maize to ensure its suitability for broad range applications; *ii)* genotyping-by-sequencing (GBS) method following the protocol described by Poland et al. (2012). Individual genomic DNA samples were digested by two restriction enzymes *Pst*I and *Mse*I and samples were included in a sequencing plate. DNA fragments from each sample were ligated to specific barcode adapters and multiplexed per flow

cell. The sequencing was performed on the Illumina NextSeq 500 platform (Illumina Inc., San Diego, CA, United States).

The GBS raw data was used for two purposes: 1) call the SNPs using the B73 RefGen_v2 reference genome, and 2) build a "mock reference genome" to perform the SNP calling using it as the reference genome. The B73 RefGen_v2 was chosen for equivalency motives, given that this version was employed for the SNP-array development (Unterseer *et al.*, 2014), and another version could induce bias in further analysis.

The mock reference was built according to the pipeline proposed by Melo et al. (2016). This pipeline integrates custom parsing and well-known filtering procedures, using bioinformatics tools, adopting clustering strategies to build a population-tailored mock reference, using the same GBS data for downstream SNP calling. We executed the pipeline stages solely to build a mock reference. First, we parsed the raw reads compressed FASTQ files to verify GBS fragments (reads), indicated by the Illumina common adapter's presence, coupled with the appropriate cut site residue. Reads with uncalled bases (i.e., N's) are discarded. Then, we trimmed reads based on quality (Phred 33) and adaptors, according to recommended by pipeline. Afterward, reads are demultiplexed and separated into a genotype-specific FASTQ file. The next step was to cluster reads and assemble the mock reference, using the parsed and quality-filtered reads (150bp) from the previous stage. In this step, the script calls upon VSEARCH (Rognes *et al.*, 2016) to cluster the reads using the *consout* clusterization algorithm based on a 0.93 similarity threshold, thereby producing a reduced list of non-redundant consensus sequences (clusters) with a minimum length of 32bp. The algorithm identifies clusters within each genotype, and then across genotypes. After, the clusters are linked via poly-A boundaries to constitute a unique sequence of nucleotides called the mock reference. The parameters used to perform the parsing, trimming, and clustering steps were recommended by Melo et al. (2016). Two mock references were assembled: considering all inbred lines (Mock-All) and considering only the single most-read abundant line (Mock-L56).

Finally, we evaluated four SNP datasets: 1) SNP-array; 2) GBS-scored in B73 (GBS-B73); 3) GBS-scored in Mock-All (GBS-Mock-All), and 4) GBS-scored in Mock-L56 (GBS-Mock-L56). For datasets from GBS, SNPs were scored from the raw sequence data using the TASSEL 5.0 GBSv2 pipeline (Glaubitz *et al.*, 2014) under default parameters values. Tags were aligned against the reference genome (B73, Mock-All, and Mock-L56) using BWA aligner aiming to reduce bias since the same used to development of SNP-array (Unterseer *et al.*, 2014).

For all SNP datasets, markers with low call rate (<90%) and non-biallelic were removed from the datasets, and remaining missing data was imputed by *Beagle 5.0* algorithm (Browning *et al.*, 2018). Pairwise linkage disequilibrium (LD) was calculated as the squared allele frequencies correlation (r^2), and values higher than 0.99 were removed from datasets using the *SNPRelate* package (Zheng *et al.*, 2012). The filtered SNP datasets from inbred lines contained 316.688 (SNP-array), 12.077 (GBS-B73), 2.597 (GBS-Mock-All), and 544 (GBS-Mock-L56) remaining markers. Afterward, heterozygous loci

on at least one individual were removed, and high-quality polymorphic SNPs from parental lines were combined to build the artificial hybrid genomic matrix. Additionally, duplicated markers across chromosomes were removed to avoid overestimates caused by multicollinearity. Lastly, markers with minor allele frequency (MAF) lower than 0.05 were removed from the hybrid genomic matrices, scoring 62.409 (SNP-array), 5.594 (GBS-B73), 311 (GBS-Mock-All), and 22 (GBS-Mock-L56) remaining SNP markers for hybrids.

3.2.3. Population structure and diversity analyses

SNP datasets from the 49 parental inbred lines were used to evaluate the population structure. In this specific analysis, we consider heterozygous loci and rare variants (MAF < 0.05), aiming to capture all diversity and variability to perform the principal components analysis (PCA) and calculate the relatedness among inbred lines. The k-means clustering was applied to determine the optimal number of groups (k) concerning the lowest associated Bayesian information criterion (BIC) value. This procedure aims to determine the number of groups that minimizes the within-group variance associating with a likelihood criterion. We used the *adegenet* package (Jombart and Ahmed, 2011) to perform the k-means clustering, concerning k ranging from 1 to 6, and 10.000 iterations and 1.000 starting points for each k value. PCA was accomplished, and bi-plot graphs were built to assess the population structure.

For each SNP datasets, we computed the genetic distances (GD) among inbred lines, using Roger's distance. Mantel correlation was conducted in the GD matrices to test the relatedness among them. Moreover, we identified the shared and non-shared SNP markers between SNP-array and GBS-B73. As approaches had the same reference genome (B73v2), this evaluation is straightforward due to the unique coordinate positions of the SNPs, whereas for GBS-Mock-All and GBS-Mock-L56 it was not possible, because each one has a different reference genome built according to the previous section.

3.2.4. Genomic prediction

We used the SNP datasets to create genomic relationship matrices (GRM), and we applied additive GBLUP (eq.2) and additive-dominance GBLUP (eq.3) models to perform the genomic prediction, following the equations:

$$\hat{\boldsymbol{g}} = \mathbf{1}\mu + \boldsymbol{Z_a a} + \boldsymbol{\varepsilon} \quad \text{(eq.2)}$$

$$\hat{\boldsymbol{g}} = \mathbf{1}\mu + \boldsymbol{Z_a a} + \boldsymbol{Z_d d} + \boldsymbol{\varepsilon} \quad \text{(eq.3)}$$

where $\hat{\boldsymbol{g}}$ is the vector of adjusted means of hybrids from the joint analysis; μ is the mean (intercept); $\boldsymbol{a}$ is the vector of additive genetic effects of the individuals, where $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{G_a}\sigma_a^2)$; $\boldsymbol{d}$ is the vector of dominance effects, where $\boldsymbol{d} \sim N(\mathbf{0}, \boldsymbol{G_d}\sigma_d^2)$; and $\boldsymbol{\varepsilon}$ is the vector of random residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_\varepsilon^2)$. $\mathbf{1}$ is the vector of ones; $\boldsymbol{Z_a}$ and $\boldsymbol{Z_d}$ are incidence matrices for $\boldsymbol{a}$ and $\boldsymbol{d}$. σ_a^2 is the genomic additive variance, σ_d^2 is the genomic dominance variance, and σ_ε^2 is the residual variance. $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ are the additive and dominance genomic relationship matrices, following the equations: $\boldsymbol{G_a} = \boldsymbol{W_A}\boldsymbol{W_A'}/2\sum_{i=1}^{n} p_i\,(1-p_i)$ and $\boldsymbol{G_d} = \boldsymbol{W_D}\boldsymbol{W_D'}/4\sum_{i=1}^{n}\big(p_i(1-p_i)\big)^2$, where p_i is the frequency of one allele of the locus i and $\boldsymbol{W}$ is the incidence matrix of markers (VanRaden, 2008). The $\boldsymbol{W_A}$ matrix was coded as 0 for homozygote A_1A_1, 1 for the heterozygote A_1A_2, and 2 for the homozygote A_2A_2. For $\boldsymbol{W_D}$, genotypes were coded as 0 for both homozygotes and 1 to heterozygote. The GRM was built using the *snpReady* package (Granato *et al.*, 2018), and the genomic prediction models were performed using the *sommer* package (Covarrubias-Pazaran, 2016).

3.2.5. Model comparison

To evaluate model performance, we employed a fivefold cross-validation scheme to compose training and validation sets. In this scheme, genotypes were randomly divided into five subsets, where four were combined to form the training test (80%), and the remaining subset as the validation set (20%). Permutations of five subsets led to five training and validation sets combinations. The same partitions were analyzed with all models. The phenotypic data from the validation set were omitted from data and then predicted by fitting the genomic prediction model, using all genomic data, and training set phenotypic data. This procedure was repeated 50 times, totaling 250 runs for each scenario (genomic prediction model × SNP dataset × trait combination). For each run, we calculated the predictive ability (PA) as the Pearson correlation between the genomic estimated breeding values (GEBV) and adjusted means from the validation set, and we averaged them for each combination. To compare PA values estimated from SNP datasets, we applied the Fisher z-transformation in the predictive ability from SNP datasets, and the Tukey test compared the means at 5% significance. Additionally, we used linear regression between GEBVs estimated from the SNP datasets to verify further associations.

Moreover, we computed the coincidence of selection for each scenario. The coincidence of selection is the percentage of common genotypes that would have been selected by their adjusted means from phenotypic analysis, and their GEBV from the genomic prediction model, regarding the same selection intensity. Moreover, we computed the proportion of the 5% top hybrids (based on adjusted means rank) that would have been captured by the genomic prediction model. For GY and EH, hybrids with the highest values were selected, whereas PH was the opposite way.

3.3. RESULTS

3.3.1. Phenotypic analysis

The joint analysis revealed significance for all sources of variation (Table S1), in all traits evaluated. The significant environmental effect suggests differential performance among them, which can be explained due to the edaphoclimatic and N level differences. The significant hybrid effect reveals genetic variability for all traits evaluated. The adjusted values ranged from 3.29 to 8.15 Mg ha^{-1}, with mean 6.35 Mg ha^{-1} for GY. Regarding the PH, values ranged from 180.6 to 230.4 cm, with a mean of 210.5 cm. Finally, for PH, and from 93.0 to 128.9 cm with a mean 113.1 cm for EH.

The broad-sense heritabilities on an entry-mean level were 0.74, 0.92, and 0.94 for GY, PH, and EH, respectively. The lowest heritability value for GY results from the fact that it is a polygenic trait, while the highest values for PH and EH confirms their oligogenic nature. The high heritability values suggest excellent reliability of the field phenotypes.

3.3.2. Quality control and allele frequencies

The number of SNP markers varied among the SNP datasets, ranging from 616,201 to 2,119 markers for SNP-array and GBS-Mock-L56, respectively (Table S2). Concerning the eliminated SNP markers by missing data (>10%), GBS based methods showed higher percentages (73.1%, 51.9%, and 50.6% for GBS-B73, GBS-Mock-All, and GBS-Mock- L56, respectively) when compared to the SNP-array (5.7%) (Fig. S1). Conversely, when we consider removed markers by the LD criteria, SNP-array showed 30.7%, while the methods based on GBS data presented values close to 5%. After quality control, 316,688, 12,077, 2,597, and 544 SNP markers were scored for the inbred lines and 62,409, 5,594, 311, and 22 hybrids considering SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56, respectively (Table S2).

Considering inbred lines, the SNP datasets based on a mock reference genome achieved the highest percentages of SNP markers with MAF lower than 5%, comprising 27.8% and 32.2% for GBS-Mock-All and GBS-Mock-L56, respectively (Fig. S2a). Similarly, the SNP datasets for hybrids based on GBS data showed the highest proportions of markers with MAF between 5% and 10% (Fig. S2b). Moreover, only 300 SNP markers were coincident between SNP-array and GBS-B73, which proposes different polymorphism regions captured by these datasets. However, both datasets showed similar distributions across the genome with higher density for the SNP-array (Fig. 1a).

3.3.3. Genetic diversity and population structure

For all SNP datasets, the k-means clustering method did not identify subpopulations among inbred lines (data not shown). Regarding PCA analysis, the SNP datasets showed similar performances concerning the explained variance by the principal components. Considering the first ten principal components (PC), GBS-B73 showed the highest value (43.7%), while SNP-array, GBS-Mock-All, and GBS-Mock-L56 presented values of 40.2%, 35.5%, and 35.7%, respectively (Fig. S3a). The PCA revealed that the first four eigenvectors from SNP-array, GBS-B73, and GBS-Mock-All exhibited similar captured variance patterns, verified by the coefficient of determination (R^2) observed in Fig. S4a, b, and d. The first four eigenvectors from the SNP-array and GBS -B73 showed high R^2 values (Fig. 1b), however, for the remaining eigenvectors, there was less similarity between the captured variance patterns (Fig. S4a). The same was repeated when analyzing SNP-array and GBS-Mock-All, however, with lower R^2 values and a change in the order of captured variance by eigenvectors (PC2 and PC3) (Fig. S4b). The spatial distribution analysis was performed using the first three PCs, along with the grain type information. For all SNP datasets, there was no clustering pattern, supporting the k-means clustering analysis. However, we detect scattering differences among inbred lines, suggesting that SNP datasets capture distinct patterns of variance (Fig. S5).

The Rogers distance (GD) matrices average values were 0.43, 0.40, 0.25, and 0.22 for SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56, respectively (Fig. S6). Regarding the Mantel correlations among GDs, high values were observed between SNP-array and GBS-B73 (r = 0.79), and GBS-B73 and GBS-Mock-All (r = 0.75). They were intermediates for SNP-array and GBS-Mock-All (r = 0.59), and GBS-Mock-All and GBS-Mock-L56 (r = 0.63). Finally, they were low values between SNP-array and GBS-Mock-L56 (r = 0.24), and GBS-B73 and GBS-Mock-L56 (r = 0.29) (Table S3). The GBS based data, especially GBS-B73 and GBS-Mock-All, detected a close genetic similarity between L07 and L08, L31 and L32, and L43 and L44 inbred lines when compared to SNP-array (Fig. S6b, c), which suggests they capture important polymorphism regions not represented by the SNP-array. This discrepancy may be caused by array-based SNPs ascertainment bias, especially when the target germplasm was not considered in the array development (Albrechtsen *et al.*, 2010; Frascaroli *et al.*, 2013).

3.3.4. Variance components and GRMs

Concerning the additive genomic relationship matrices ($\boldsymbol{G_a}$) between hybrids, SNP-array, GBS-B73, and GBS-Mock-All showed high paired correlations. Mantel's correlations were 0.93 between SNP-array and GBS-B73, 0.88 between SNP-array and GBS-Mock-All, and 0.94 between GBS-B73 and GBS-Mock-All. All comparisons with GBS-Mock-L56 performed correlation values

close to 0.55 (Table S3). In contrast, the dominance genomic relationship matrices ($\boldsymbol{G_d}$) showed pairwise Mantel's correlation values considerably lower than $\boldsymbol{G_a}$. Solely SNP-array and GBS-B73 (r = 0.69) $\boldsymbol{G_d}$ matrices performed a correlation value higher than 0.5, confirming the distinct dominance patterns estimated by SNP datasets (Fig. S7e, f, g, and h).

As expected, the proportion of explained variance by additive effects was consistently higher for PH and EH (simple traits) than GY (complex traits). For the additive model, σ_a^2 proportion ranged from 10.6% to 45.4% with mean 34.3% for GY, from 63.2% to 78.1% with mean 69.7% for PH, and from 62.2% to 87.1% with mean 76.4% for EH. GBS-Mock-All performed the highest values for GY (45.4%) and EH (87.1%), whereas GBS-Mock-L56 for PH (78.1%). However, GBS-Mock-L56 captured σ_a^2 proportions substantially lower for GY (10.6%) and EH (62.2%) when compared to other SNP datasets (Fig. 2a). Concerning the additive-dominance model, σ_a^2 proportion ranged from 10.5% to 47.1% with mean 36.8% for GY, from 65.3% to 78.1% with mean 71.5% for PH, and from 62.2% to 87.8% with mean 77.3% for EH. The captured variance proportion by dominance effects (σ_d^2) ranged from 0% to 18.9% with mean 9.8% for GY, from 0% to 10.3% with mean 5.3% for PH, and from 0% to 7.2% with mean 3.0% for EH. These σ_d^2 proportions were higher for GY than PH and EH. Furthermore, SNP-array captured the highest values, whereas GBS-B73 and GBS-Mock-All showed minor proportions. Moreover, the σ_d^2 captured by GBS-Mock-L56 was close to zero for all traits evaluated (Fig. 2b).

3.3.5. Genomic Prediction

Overall, the predictive abilities estimated by the additive-dominance model were higher than the additive model for all traits and SNP datasets, excepting GBS-Mock-L56, wherein PA values were equivalent for both models (Fig. 3). The highest PA differences between genomic prediction models were observed for EH, intermediate for PH, and lowest for GY, supporting the results from variance genetic components. PA mean values were 0.52 for GY, 0.74 for PH, and 0.82 for EH. Also, the standard deviations from the PA estimates were lower for EH, when compared to PH and GY. Analyzing the additive model, SNP-array, GBS-B73, and GBS-Mock-All did not reveal statistical differences between PA estimates for all traits (Fig. 3a). Conversely, the additive-dominance model showed statistical differences between these SNP datasets, being highest for GY (Fig. 3b). Concerning the additive-dominance model, SNP-array presented the highest PA values, comprising 0.66 for GY, 0.79 for PH, and 0.87 for EH. GBS-Mock-L56 showed PA values substantially lower than other SNP datasets, especially for GY (complex trait). PA mean values for SNP-array, GBS-B73, and GBS-Mock-All were approximately three times higher for GY, 25% for PH, and 26% for EH than from GBS-Mock-L56.

The linear regression analyzes among the GEBVs obtained from SNP datasets showed relevant similarities among SNP-array, GBS-B73, and GBS-Mock-All (Table S4). The β_0 (intercept) and β_1

(regression coefficient) parameters were close to 0 and 1, respectively. Moreover, very high Pearson correlations (r > 0.99) among GEBVs indicate the high similarity among them. All possibilities for linear regression analyzes were performed between SNP datasets, and the GBS-Mock-L56 yielded the lowest correlation values, especially for GY. GEBVs from the additive-dominance model, among SNP-array, GBS-B73, and GBS-Mock-All, performed lower correlations than the additive model, although with high magnitudes, concerning 0.95 for GY, 0.98 for PH, and 0.99 for EH (Fig. 4).

3.3.6. Coincidence of selection

The coincidence of selection increased considerably when selection intensity varied from 1 to 10%. The coincidence values were higher for simple traits (PH and EH) when compared with GY (complex trait) (Fig. 5a). SNP-array, GBS-B73, and GBS-Mock-All described similar coincidences values across selection intensities, especially for PH and EH. For GY, using the additive model, GBS-B73 obtained slightly lower coincidence values when compared to SNP-array and GBS-Mock-All. However, considering the additive-dominance model, SNP-array achieved the highest coincidence values for GY across all selection intensities. GBS-Mock-L56 showed coincidence values lower than the other SNP datasets for all traits, especially for GY (Fig. 5a).

The increase of selection intensity increased the genomic models' ability to capture the 5% top hybrids selected by the phenotypic value (Fig. 5b). GY showed the highest gain (27.9% to 68.6%) when the selection intensity varied from 5 to 20%. SNP-array, GBS-B73, and GBS-Mock-All showed similar tendencies to capture the 5% top hybrids with the increase of selection intensity using the additive model. When we considered the additive-dominance model, the SNP-array showed the highest ability to capture the 5% top hybrids in all selection intensities for GY (Fig. 5b). On the other hand, GBS-Mock-L56 showed the lowest percentages for all scenarios, especially for GY.

3.4. Discussion

Genotypic data, especially from SNP markers, have been intensively used by breeders to perform genomic studies. Currently, arrays and GBS are the leading genotyping platforms to obtain SNP markers. Both approaches have their advantages and disadvantages, affecting costs, procedures, and further genetic analyzes. Recent studies have been carried to verify the influence of SNP platforms in genetic diversity, GWAS, and genomic prediction, but restricted mainly to inbred lines (Elbasyoni *et al.*, 2018; Darrier *et al.*, 2019; Negro *et al.*, 2019; Chu *et al.*, 2020). In our study, a wide range of analysis was performed to study the influence of the genotyping method on the assessment of parental lines diversity and prediction of hybrids. Additionally, we used a method to build a mock reference genome based on their GBS data (Melo *et al.*, 2016), aiming to capture the intrinsic variability within the

genotypes. This approach provides independence from an external genome to perform the SNP discovery. We used two ways to build the mock reference: *i)* using all the genotypes (GBS-Mock-All), and *ii)* considering only the most representative genotype (GBS-Mock-L56) (the most reads). In the next few paragraphs, we discuss our results in two stages: how the mock reference genome influences genetic diversity and population structure assessment and how it influences hybrid genomic prediction.

3.4.1. Impact of the mock reference genome on genetic diversity and population structure

Genetic diversity and population structure studies based on molecular markers are essential for plant breeders to understand the working germplasm. Moreover, the population structure knowledge permits the plant breeder to describe heterotic groups (Wu *et al.*, 2016) and select more interesting combinations between inbred lines. Recent studies aimed at verifying how the array and GBS scored markers affect genetic diversity (Darrier *et al.*, 2019; Chu *et al.*, 2020). However, our focus is to investigate how the use of a mock reference during the SNP calling affects the discovered markers and how they assess the populational structure.

Despite the higher frequency of low-quality SNP markers from GBS datasets (Elshire *et al.*, 2011; Poland *et al.*, 2012), SNP-array, GBS-B73, and GBS-Mock-All revealed similar performance concerning the genetic diversity and population structure of parental lines (Fig. S4 and Fig. S5). The PCA confirmed that these SNP datasets capture similar variance patterns, especially regarding the first four eigenvectors (Fig. S4a, b, and d). However, we observed that the captured variance is more consistent when we compare GBS-B73 and GBS-Mock-All than comparisons with SNP-array. These discrepancies may be due to ascertainment bias of the SNP-array (Albrechtsen *et al.*, 2010; Moragues *et al.*, 2010; Frascaroli *et al.*, 2013; Heslot *et al.*, 2013) since it was built using 30 temperate inbred lines from Europe and the USA (Unterseer *et al.*, 2014). Furthermore, there were only 300 matching markers between SNP-array and GBS-B73, which reveals different polymorphism regions captured by these approaches (Fig. 1a). Darrier et al. (2019) studied the influence of the genotyping platform (array versus GBS) on the assessment of genetic diversity and observed only 496 common SNP, indicating that these genotyping platforms capture different regions of polymorphism in wheat. Besides, we noticed that GBS-B73 and GBS-Mock-All identified a higher rare variant frequency (MAF < 0.05), which is more interesting for diversity studies, where these rare alleles can be a new source of variation (Darrier *et al.*, 2019; Chu *et al.*, 2020).

Genetic distances between inbred lines exhibited similar patterns, with a higher average distance for the SNP-array (Table S3). Accordingly, fewer markers in the GBS datasets increase the individual marker effect in parameters based on allelic frequency (Moragues *et al.*, 2010), which can cause a bias in the estimation of the genetic distance between some genotypes.

Notwithstanding the genetic divergence between temperate and tropical germplasm in maize (Lu *et al.*, 2009), the reference genome [B73 temperate genome or a mock genome (with all lines)] did not exhibit significant differences either the population structure or the relationship among inbred lines (Fig. S4d, Fig. S5). Furthermore, the mock reference genome performed similarly to a high-density SNP array. On the other hand, the mock reference built with one genotype discovered few high-quality polymorphic markers, affecting their diversity estimates. Therefore, we recommend using the genomic data (GBS data) from all genotypes to build the mock reference genome to avoid the ascertainment bias common to SNP-arrays.

3.4.2. The impact from mock reference genome on genomic prediction

A major challenge for plant breeders is to evaluate the performance of all possible combinations among hundreds or thousands of inbred lines developed by breeding programs (Hallauer *et al.*, 2010). Genetic values estimation of non-phenotyped individuals became feasible with the increasing availability of molecular markers. However, it is not clear to the scientific community how the SNP datasets obtained from different genotyping platforms influence the hybrids' predictive ability. In this sense, simple traits (PH and EH) exhibited the highest proportions of additive variance captured by the $\boldsymbol{G_a}$ matrices than for GY, as expected, considering the genetic control of the traits (Fischer *et al.*, 2008; Hallauer *et al.*, 2010). Also, the GBS-based approaches captured a higher σ_a^2 proportion than SNP-array, especially concerning mock genomes (Fig. 2). The high amount of markers may explain it with low MAF (MAF < 0.10), since the range of allele frequencies of a given locus, maximizes allelic substitution effect and, consequently, the additive variance (Bernardo, 2010; Galli *et al.*, 2020). In contrast, SNP-array captured a higher proportion of the variance caused by the dominance deviations than GBS-based approaches, especially for GY (Fig. 2b).

The $\boldsymbol{G_a}$ matrices of hybrids revealed high correlations among SNP-array, GBS-B73, and GBS-Mock-All, which indicates comparable additive relationships between hybrids by the three approaches (Table S3). Elbasyoni et al. (2018), investigating the influence of SNPs from different genotyping platforms on the genomic prediction, observed a high correlation between the additive relationship kernels among 282 wheat inbred lines (r = 0.77). We obtained higher correlation values, r = 0.93 between SNP-array and GBS-B73, r = 0.88 between SNP-array and GBS-Mock-All, and r = 0.94 for GBS-B73 and GBS-Mock-All. These results suggest that GBS-Mock-All captures additive relationships for hybrids comparable to SNP-array and GBS-B73 (Fig. S7a, b, c, and d). In contrast, the correlations among $\boldsymbol{G_d}$ matrices were lower, except for the comparison between SNP-array and GBS-B73 (r = 0.69). However, it is worth noting that $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ based on the GBS-Mock-L56 (built with only one inbred line) showed very low correlations with other SNP datasets, especially considering the dominance of deviations relationships.

As expected, simple traits (PH and EH) presented higher PA than GY, yielding 0.82 for EH, 0.74 for PH, and 0.52 for GY. The entry-mean heritabilities followed the fashion, where the higher ones were achieved in high heritable traits (Combs and Bernardo, 2013). Furthermore, the inclusion of dominance effects on the genomic prediction model revealed better performance for all traits, especially for GY. The incorporation of non-additive genetic effects on genomic prediction models increases the PA for hybrids, mainly, for complex traits that are more influenced by dominance deviations (Technow *et al.*, 2012, 2014; Alves *et al.*, 2019).

Regarding the SNP datasets, the PA estimated from the additive model did not reveal statistical differences for all addressed traits, except for GBS-Mock-L56 (Fig. 3a). The severe decline from GBS-Mock-L56 PA estimates may be due to the number of markers that decrease the ability to capture QTLs (Daetwyler *et al.*, 2010). Nevertheless, the difference of SNP number between SNP-array and GBS-Mock-All (62,409 to 311) did not lead to a PA increase. Previous studies have shown that increasing marker density above a certain level, around 200-250 markers for maize biparental populations (Lorenzana and Bernardo, 2009; Zhang *et al.*, 2015; Sousa *et al.*, 2019), does not lead to an increase in predictive ability, showing that the response would reach a plateau (de los Campos *et al.*, 2013). As a result of hybrids originated from a full diallel, and their parental inbred lines were related, there may be large common blocks of haplotypes and would need a few markers to estimate the relationship between them. However, for the additive-dominance GBLUP, the PA means from SNP datasets showed slight differences, exhibiting statistical significance in some comparisons between SNP-array, GBS-B73, and GBS-Mock-All, especially for GY (Fig. 3b). These results follow the proportion of the variance captured by the dominance effects in the additive-dominance model mentioned above. PA differences between SNP-array and GBS-B73 presented statistical significance only for GY, according to the inclusion of dominance effects in the genomic prediction model is more relevant for complex traits (Technow *et al.*, 2012; Alves *et al.*, 2019).

The correlations between GEBVs estimated by the SNP-array, GBS-B73, and GBS-Mock-All, exhibited high values of coefficient of determination for all traits and genomic prediction models (Table S4). The lowest, but still high R^2 value (0.91), was observed for GY (complex trait) using the additive-dominance model. These results reveal that the SNP markers obtained from these approaches are equivalent to the GEBVs estimation of hybrids, especially when only additive genetic effects are considered. Also, the outcomes demonstrate that building a mock reference genome using all genotypes from the population was adequate to capture the haplotypic variability from hybrids and achieve performance comparable to the standard approaches. Furthermore, the relationship of the population used in the SNP identification and genomic analysis does not seem to be a crucial factor for genomic prediction. It is drawn from the fact that the analyzed germplasm was historically unrelated to the germplasm used in array development and the B73 inbred line, and highly correlated to the mock reference genome. Nevertheless, all genotyping methods resulted in equivalent genetic values estimates.

In this context, a mock genome, employed as a reference genome for SNP discovery, can be a worthy alternative, especially for species where the reference genome is not available. There are several species important for food security, especially in Africa, that still do not have a reference genome (Hendre *et al.*, 2019). These species are called orphan crops, and the lack of biotechnological tools limits the genetic gain, and the productivity increase (Armstead *et al.*, 2009). These mock genomes could be employed to discover polymorphisms and perform genetic studies with reliable estimates of population structure and genomic prediction. Plant breeders could determine subpopulations (heterotic groups) to improve crosses' choice and speed up breeding schemes using genomic selection. Certainly, mock genomes have pitfalls, e.g., the lack of physical position in a constant reference, which would hinder its use to studies like GWAS and candidate genes discovery. Nevertheless, our results suggest that mock genomes are suitable for diversity studies, population structure assessment, and genomic selection, which are central applications of genomic information in breeding programs.

Our results indicate that genotyping-by-sequencing methods captured a higher frequency of rare variants, becoming most desirable to genetic studies. Nevertheless, the SNP-array methods achieved the best predictive ability and reliability to estimate de variance components. The usefulness of mock reference genomes to discover polymorphisms within the population delivered reliable estimates for genetic diversity and population structure assessment, which can be a worthy alternative, especially for those species where the reference genome is not available. Furthermore, genomic prediction estimates were comparable with standard approaches, mainly when considering simple traits and additive effects. However, mock genomes were slightly worse to predict complex traits and estimate dominance effects, still had similar performance to GBS-B73. In this context, we strongly recommend using all individuals from the population to build a mock reference genome to avoid sampling bias from the polymorphism within the population. We believe that this knowledge supports plant breeders choosing the genotyping platform to perform their genomic studies and brings an alternative to SNP discovery for diversity studies and genomic selection on orphan crops.

References

Albrechtsen A, Nielsen FC, Nielsen R (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27**: 2534–2547.

Alves FC, Granato ÍSC, Galli G, Lyra DH, Fritsche-Neto R, De Los Campos G (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* **15**: 1–18.

Armstead I, Huang L, Ravagnani A, Robson P, Ougham H (2009). Bioinformatics in the orphan crops. *Brief Bioinform* **10**: 645–653.

Bernardo R (2010). *Breeding for Quantitative Traits in* Plants. Stemma Press: Woodbury, MN.

Browning BL, Zhou Y, Browning SR (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**: 338–348.

Butler D, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2018). ASReml-R Reference Manual Version 4. : 176.

Chu J, Zhao Y, Beier S, Schulthess AW, Stein N, Philipp N, *et al.* (2020). Suitability of Single-Nucleotide Polymorphism Arrays Versus Genotyping-By-Sequencing for Genebank Genomics in Wheat. *Front Plant Sci* **11**: 1–12.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**: 1496–1502.

Combs E, Bernardo R (2013). Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *Plant Genome* **6**: plantgenome2012.11.0030.

Covarrubias-Pazaran G (2016). Genome-Assisted prediction of quantitative traits using the r package sommer. *PLoS One* **11**: 1–15.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **185**: 1021–1031.

Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, Macaulay M, *et al.* (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front Plant Sci* **10**: 1–14.

Dolatabadian A, Patel DA, Edwards D, Batley J (2017). Copy number variation and disease resistance in plants. *Theor Appl Genet* **130**: 2479–2490.

Elbasyoni IS, Lorenz AJ, Guttieri M, Frels K, Baenziger PS, Poland J, *et al.* (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci* **270**: 123–130.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, *et al.* (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: 1–10.

Fischer S, Möhring J, Schön CC, Piepho HP, Klein D, Schipprack W, *et al.* (2008). Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. *Plant Breed* **127**: 446–451.

Frascaroli E, Schrag TA, Melchinger AE (2013). 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. *Theor Appl Genet* **126**: 133–141.

Galli G, Alves FC, Morosini JS, Fritsche-Neto R (2020). On the usefulness of parental lines GWAS for predicting low heritability traits in tropical maize hybrids (M Causse, Ed.). *PLoS One* **15**: e0228724.

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, *et al.* (2011). A large maize (zea mays L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* **6**.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, *et al.* (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**.

Granato ISC, Galli G, de Oliveira Couto EG, e Souza MB, Mendonça LF, Fritsche-Neto R (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol Breed* **38**.

Gupta PK, Rustgi S, Mir RR (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity (Edinb)* **101**: 5–18.

Hallauer AR, Carena MJ, Filho JBM (2010). *Quantitative Genetics in Maize Breeding*. Springer-Verlag New York: New York.

Hendre PS, Muthemba S, Kariba R, Muchugi A, Fu Y, Chang Y, *et al.* (2019). African Orphan Crops Consortium (AOCC): status of developing genomic resources for African orphan crops. *Planta* **250**: 989–1003.

Heslot N, Rutkoski J, Poland J, Jannink JL, Sorrells ME (2013). Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. *PLoS One* **8**.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, *et al.* (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**: 121–135.

Jombart T, Ahmed I (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070–3071.

Kang YJ, Lee T, Lee J, Shim S, Jeong H, Satyawan D, *et al.* (2016). Translational genomics for plant breeding with the genome sequence explosion. *Plant Biotechnol J* **14**: 1057–1069.

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

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**: 327–345.

Lu Y, Yan J, Guimarães CT, Taba S, Hao Z, Gao S, *et al.* (2009). Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor Appl Genet* **120**: 93–115.

Lyra DH, de Freitas Mendonça L, Galli G, Alves FC, Granato ÍSC, Fritsche-Neto R (2017). Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Mol Breed* **37**.

Matias FI, Xavier Meireles KG, Nagamatsu ST, Lima Barrios SC, Borges do Valle C, Carazzolle MF, *et al.* (2019). Expected Genotype Quality and Diploidized Marker Data from Genotyping-by-Sequencing of Urochloa spp. Tetraploids . *Plant Genome* **12**: 190002.

Melo ATO, Bartaula R, Hale I (2016). 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* **17**: 1–15.

Mendonça L de F, Granato ÍSC, Alves FC, Morais PPP, Vidotti MS, Fritsche-Neto R (2017). Accuracy and simultaneous selection gains for N-stress tolerance and N-use efficiency in maize tropical lines. *Sci Agric* **74**: 481–488.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor Appl Genet* **120**: 1525–1534.

Morosini JS, Mendonça L de F, Lyra DH, Galli G, Vidotti MS, Fritsche-Neto R (2017). Association mapping for traits related to nitrogen use efficiency in tropical maize lines under field conditions. *Plant Soil* **421**: 453–463.

Munjal G, Hao J, Teuber LR, Brummer EC (2018). Selection mapping identifies loci underpinning autumn dormancy in alfalfa (Medicago sativa). *G3 Genes, Genomes, Genet* **8**: 461–468.

Negro SS, Millet EJ, Madur D, Bauland C, Combes V, Welcker C, *et al.* (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol* **19**: 1–22.

Nielsen R (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**: 931–942.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* **7**.

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, *et al.* (2017). Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Mol Plant* **10**: 1047–1064.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**: 1–22.

Rousselle Y, Jones E, Charcosset A, Moreau P, Robbins K, Stich B, *et al.* (2015). Study on essential derivation in maize: III. selection and evaluation of a panel of single nucleotide polymorphism loci for use in European and North American germplasm. *Crop Sci* **55**: 1170–1180.

Sousa MB, Galli G, Lyra DH, Granato ÍSC, Matias FI, Alves FC, *et al.* (2019). Increasing accuracy and reducing costs of genomic prediction by marker selection. *Euphytica* **215**: 18.

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet* **125**: 1181–1194.

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* **197**: 1343–1355.

Thomson MJ (2014). High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breed Biotechnol* **2**: 195–212.

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

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

Vidotti MS, Lyra DH, Morosini JS, Granato ÍSC, Quecine MC, de Azevedo JL, *et al.* (2019). Additive and heterozygous (dis)advantage GWAS models reveal candidate genes involved in the genotypic variation of maize hybrids to Azospirillum brasilense. *PLoS One* **14**: 1–21.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES (2014). Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. *PLoS Genet* **10**.

Wu Y, San Vicente F, Huang K, Dhliwayo T, Costich DE, Semagn K, *et al.* (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor Appl Genet* **129**: 753–765.

Xu Y, Li P, Zou C, Lu Y, Xie C, Zhang X, *et al.* (2017). Enhancing genetic gain in the era of molecular breeding. *J Exp Bot* **68**: 2641–2666.

Xu C, Ren Y, Jian Y, Guo Z, Zhang Y, Xie C, *et al.* (2017). Development of a maize 55 K SNP array with improved genome coverage for molecular breeding. *Mol Breed* **37**.

Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz MA, et *al.* (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb)* **114**: 291–299.

Zhang X, Zhang H, Li L, Lan H, Ren Z, Liu D, *et al.* (2016). Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. *BMC Genomics* **17**: 1–16.

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

FIGURES

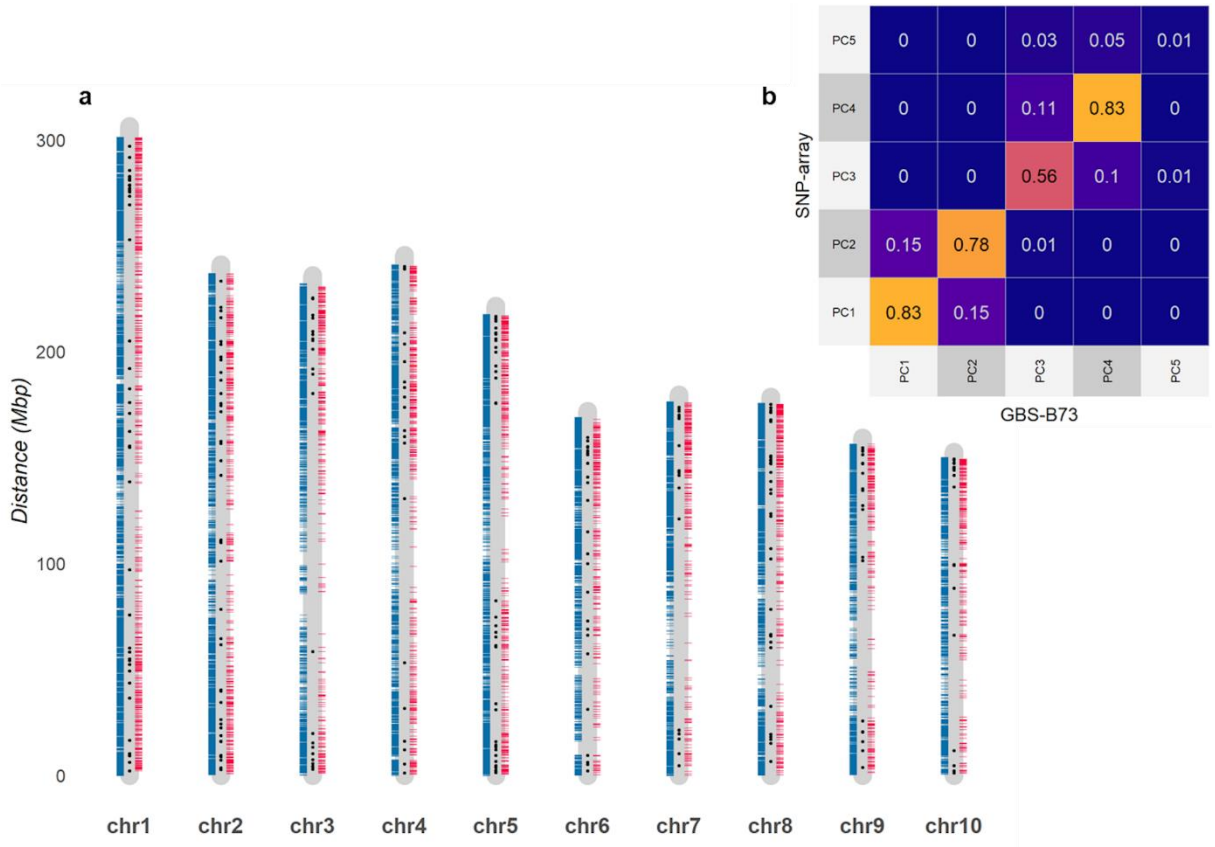

Fig. 1 Comparison between SNP-array and GBS-B73 dataset. a Distribution of SNP-array (blue), GBS-B73 (red), and coincident (black dots) markers across maize chromosomes; **b** Paired Pearson correlation among the five first principal components from SNP-array and GBS-B73 SNP datasets.

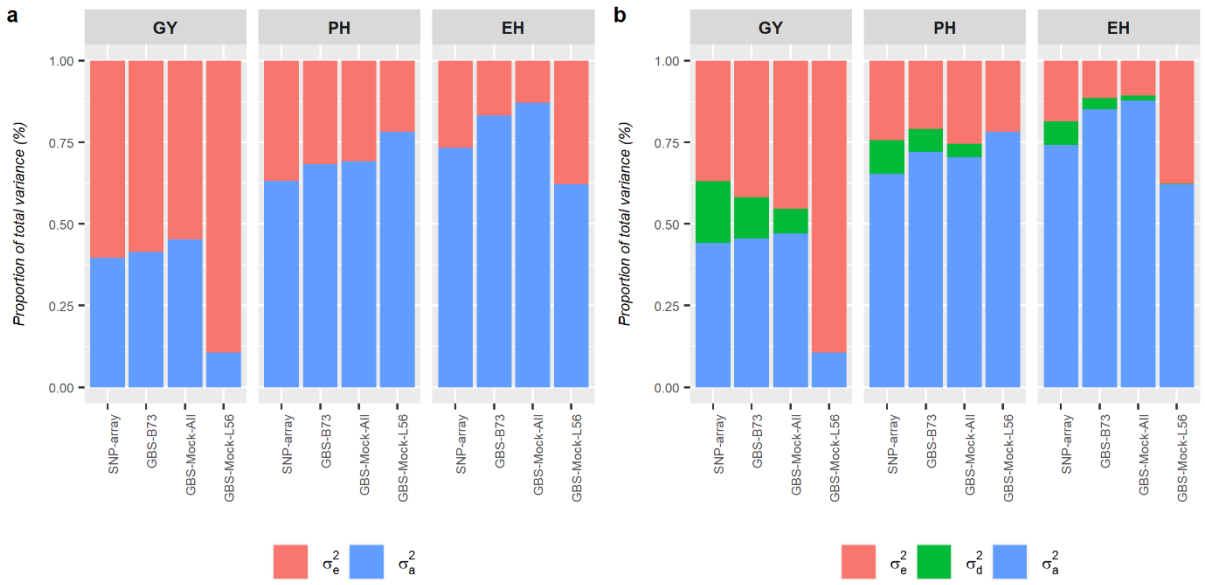

Fig. 2 Proportion of total variance captured by additive ($\boldsymbol{\sigma_a^2}$), dominance ($\boldsymbol{\sigma_d^2}$), and residual ($\boldsymbol{\sigma_e^2}$) effects from genomic prediction models by SNP datasets. The traits evaluated were grain yield (GY), plant height (PH), and ear height (EH). The SNP datasets used to perform (**a**) additive GBLUP and (**b**) additive-dominance GBLUP were SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56.

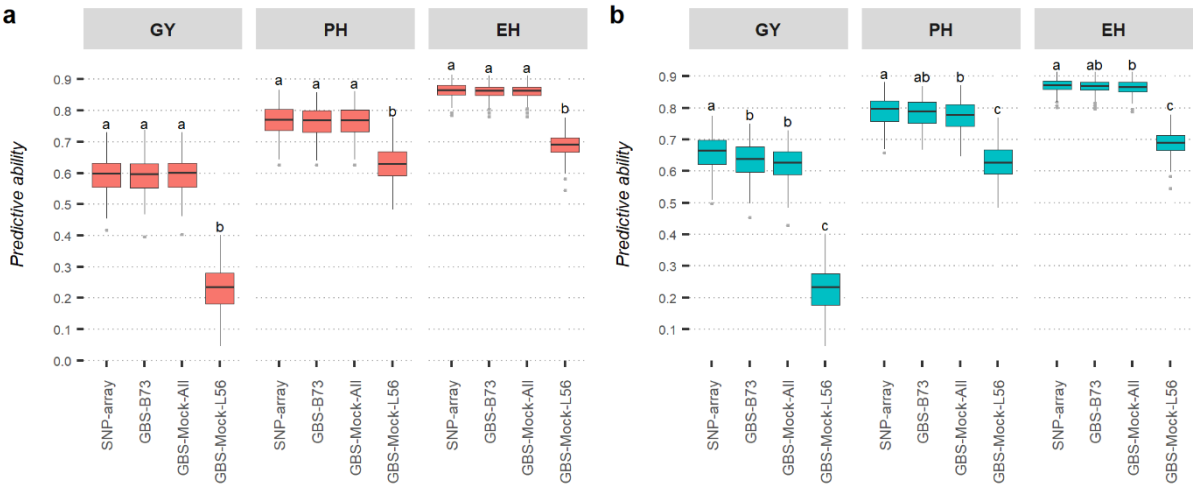

Fig. 3 Comparison of boxplot distributions of predictive ability from SNP datasets. The SNP datasets used to perform (**a)** additive GBLUP and (**b**) additive-dominance GBLUP were SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56. Different letters indicate significant differences among groups (post hoc nonparametric Tukey's test, $P < 0.05$).

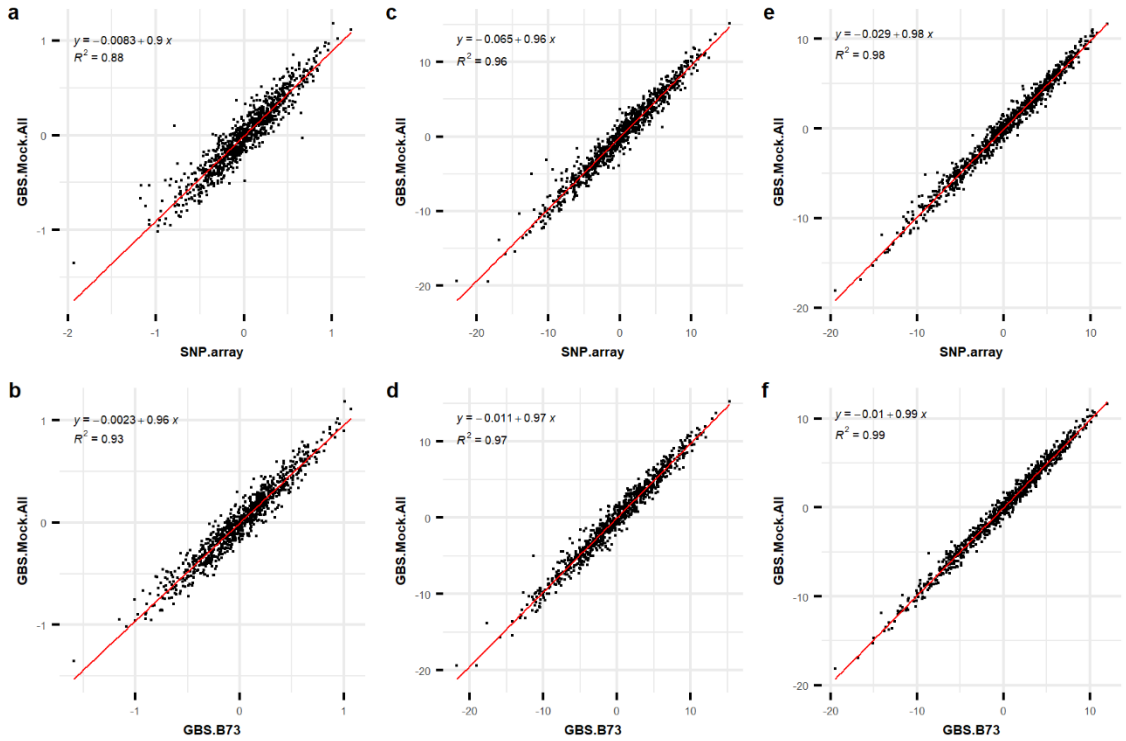

Fig. 4 Relationship between genomic estimated breeding values (GEBV) of 903 tropical maize hybrids estimated by additive-dominance GBLUP. **a**, **c**, and **e** Relationship between GEBVs from SNP-array and GBS-Mock-All datasets; **b**, **d**, and **f** Relationship between GEBVs from GBS-B73 and GBS-Mock-All datasets; **a** and **b** GEBVs of grain yield; **c** and **d** GEBVs of plant height; **e** and **f** GEBVs of ear height. The red line is the regression slope.

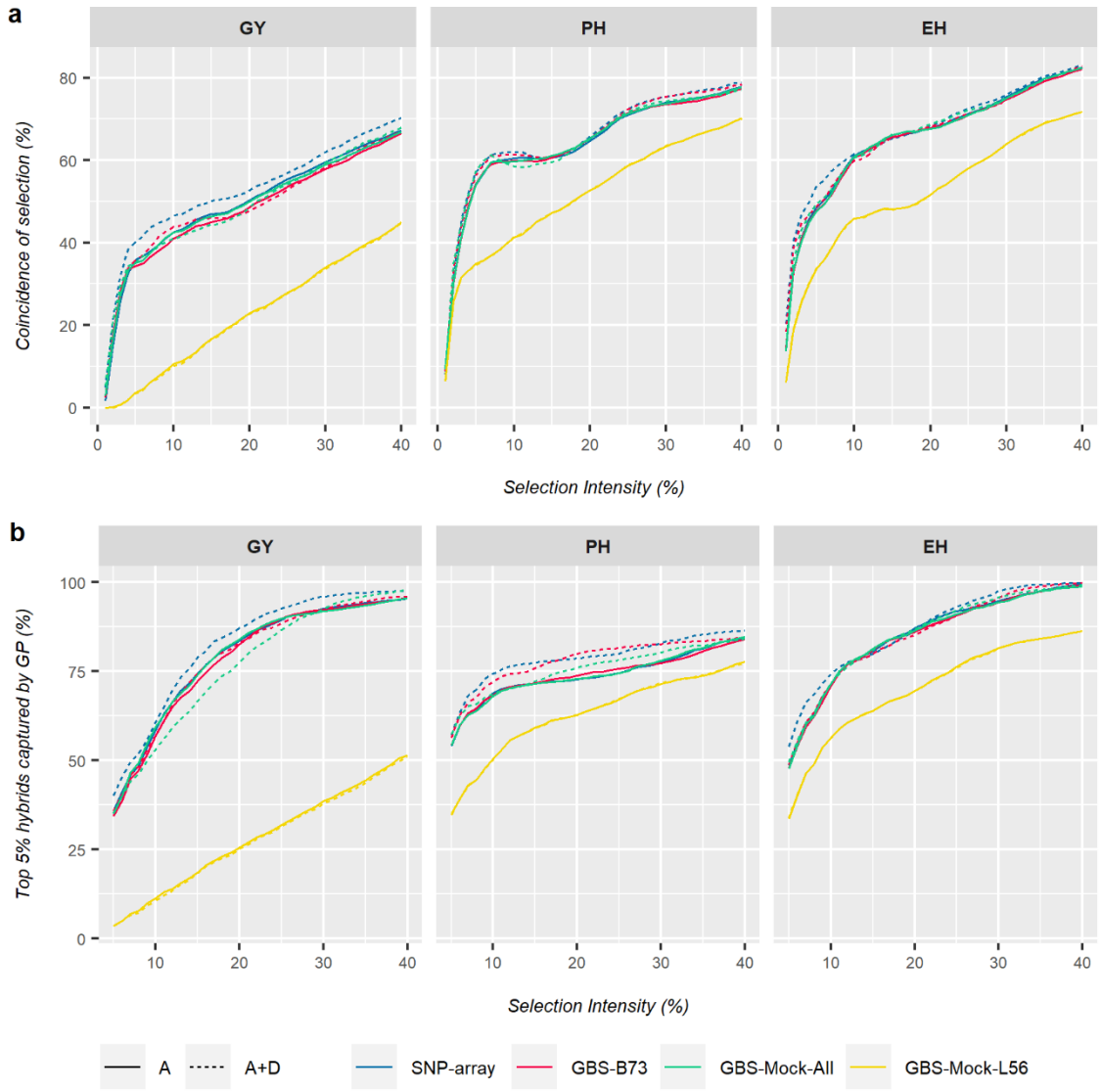

Fig. 5 Coincidence of selection and top 5% hybrids select by genomic prediction. a Coincidence of selection percentage (y-axis) over a series of continuous selection intensities (1-40%) (x-axis); **b** proportion of the 5% top hybrids (based on phenotypic rank) selected according to the genomic prediction models considering a range of selection intensities (5-40%) (x-axis). Each panel corresponds to a combination between the genomic prediction model and trait. Type lines correspond genomic prediction model and color lines represent the SNP datasets. A: additive GBLUP; A+D: additive-dominance GBLUP; GY: grain yield; PH: plant height: EH: ear height.

SUPPLEMENTAY TABLES

Table S1. Wald test of fixed effects, likelihood-ratio test (LRT) of random effects, heritability at entry mean, variance components, and overall means for grain yield (GY, Mg ha^{-1}), plant height (PH, cm), and ear height (EH, cm) in tropical maize single-cross

Source	GY	PH	EH
		Wald statistic	
Environment	1487***	1744***	1881***
Check	42***	8*	351***
Hybrid	4988***	13477 ***	14261***
		Likelihood-ratio test (LRT)	
Block/Environment	373.92***	1757***	1504***
Check x Environment	493.65***	201.9***	64.78***
		Heritability and variance components	
Heritability (H2)	0.74	0.92	0.94
σ_g^2	1.07	171.25	129.67
σ_{cl}^2	1.34	29.79	8.18
σ_ε^2	1.62	85.25	60.76
		Overall mean (adjusted means)	
Mean	6.35	210.5	113.1

*, ***: significant at the 0.05 and 0.001 probability level (by Wald test or LRT), respectively.

Table S2. Number of markers scored (raw data), and the final number of markers used to assess 49 tropical inbred lines and 903 maize single-crosses after quality control for all SNPs datasets

	SNP datasets[a]			
	SNP-array	**GBS-B73**	**GBS-Mock-All**	**GBS-Mock-L56**
Raw data	616,201	59,246	8,383	2,119
Lines[b]	**316,688**	**12,077**	**2,597**	**544**
Hybrids[c]	**62,409**	**5,594**	**311**	**22**

[a] SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock-All: genotyping-by-sequence with SNP calling using the mock reference built with all inbred lines; GBS-Mock-L56: genotyping-by-sequence with SNP calling using the mock reference built with the single most-read abundant inbred line (L56).

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

[c] number of markers used to evaluate maize single crosses (genomic prediction).

Table S3. Mantel correlation of Rogers genetic distance (**GD**), additive genomic relationship ($\boldsymbol{G_a}$), and dominance genomic relationship ($\boldsymbol{G_d}$) matrices estimated from SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56 markers

		SNP datasets[a]		
		GBS-B73	GBS-Mock-All	GBS-Mock-L56
GD[b]	SNP-array	0.79**	0.59**	0.24**
	GBS-B73	-	0.75**	0.29**
	GBS-Mock-All	-	-	0.63**
$\boldsymbol{G_a}$[c]	SNP-array	0.93**	0.88**	0.52**
	GBS-B73	-	0.94**	0.56**
	GBS-Mock-All	-	-	0.55**
$\boldsymbol{G_d}$[c]	SNP-array	0.69**	0.34**	0.06**
	GBS-B73	-	0.46**	0.10**
	GBS-Mock-All	-	-	0.12**

[a] SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock-All: genotyping-by-sequence with SNP calling using the mock reference built with all inbred lines; GBS-Mock-L56: genotyping-by-sequence with SNP calling using the mock reference built with the single most-read abundant inbred line (L56).

** Significant at the 0.01 probability level.

[b] Rogers genetic distance (**GD**) matrices were computed with markers from 49 inbred lines data

[c] $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ matrices were computed with markers from 903 maize singles crosses.

Table S4. Linear regression parameters (intercept and regression coefficient), Pearson correlation and coefficient of determination between genomic estimated breeding values (GEBVs) estimated from SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56

Models	Traits	SNP dataset 1	SNP dataset 2	β_0	β_1	r	R^2
A	**GY**	SNP-array	GBS-B73	-0.0001	0.9960	0.989	0.978
		SNP-array	GBS-Mock-All	-0.0001	0.9929	0.993	0.985
		SNP-array	GBS-Mock-L56	0.0001	1.1609	0.442	0.195
		GBS-B73	GBS-Mock-All	0.0000	0.9901	0.997	0.993
		GBS-B73	GBS-Mock-L56	0.0002	1.1722	0.449	0.202
		GBS-Mock-All	GBS-Mock-L56	0.0002	1.1533	0.439	0.193
	PH	SNP-array	GBS-B73	-0.0001	0.9993	0.994	0.988
		SNP-array	GBS-Mock-All	-0.0006	0.9996	0.996	0.993
		SNP-array	GBS-Mock-L56	0.0006	1.0025	0.819	0.671
		GBS-B73	GBS-Mock-All	-0.0005	0.9951	0.997	0.995
		GBS-B73	GBS-Mock-L56	0.0007	1.0121	0.832	0.692
		GBS-Mock-All	GBS-Mock-L56	0.0012	1.0081	0.827	0.683
	EH	SNP-array	GBS-B73	-0.0003	1.0001	0.996	0.992
		SNP-array	GBS-Mock-All	-0.0005	1.0005	0.997	0.993
		SNP-array	GBS-Mock-L56	0.0014	1.0042	0.796	0.633
		GBS-B73	GBS-Mock-All	-0.0002	0.9988	0.999	0.998
		GBS-B73	GBS-Mock-L56	0.0017	1.0124	0.805	0.648
		GBS-Mock-All	GBS-Mock-L56	0.0019	1.0108	0.804	0.646
A+D	**GY**	SNP-array	GBS-B73	0.0076	0.9935	0.958	0.919
		SNP-array	GBS-Mock-All	0.0106	0.9807	0.938	0.880
		SNP-array	GBS-Mock-L56	0.0206	1.1541	0.400	0.160
		GBS-B73	GBS-Mock-All	0.0032	0.9737	0.965	0.932
		GBS-B73	GBS-Mock-L56	0.0131	1.1681	0.420	0.176
		GBS-Mock-All	GBS-Mock-L56	0.0102	1.1515	0.417	0.174
	PH	SNP-array	GBS-B73	0.0589	0.9923	0.985	0.971
		SNP-array	GBS-Mock-All	0.0727	0.9934	0.978	0.957
		SNP-array	GBS-Mock-L56	0.1771	1.0051	0.797	0.634
		GBS-B73	GBS-Mock-All	0.0150	0.9914	0.983	0.966
		GBS-B73	GBS-Mock-L56	0.1192	1.0152	0.810	0.657
		GBS-Mock-All	GBS-Mock-L56	0.1051	1.0108	0.814	0.662
	EH	SNP-array	GBS-B73	0.0203	0.9964	0.993	0.986
		SNP-array	GBS-Mock-All	0.0309	0.9959	0.990	0.980
		SNP-array	GBS-Mock-L56	0.0789	1.0083	0.789	0.623
		GBS-B73	GBS-Mock-All	0.0108	0.9956	0.993	0.986
		GBS-B73	GBS-Mock-L56	0.0588	1.0172	0.799	0.638
		GBS-Mock-All	GBS-Mock-L56	0.0482	1.0156	0.800	0.639

SNP-array: Affymetrix® Axiom Maize Genotyping array; GBS-B73: genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock-All: genotyping-by-sequence with SNP calling using the mock reference built with all inbred lines; GBS-Mock-L56: genotyping-by-sequence with SNP calling using the mock reference built with the single most-read abundant inbred line (L56); GY: grain yield; PH: plant height; EH: ear height; A: additive GBLUP; A+D: additive-dominance GBLUP.

β_0: linear regression intercept; β_1: coefficient of regression; r: Pearson correlation; R^2: coefficient of determination

SUPPLEMENTARY FIGURES

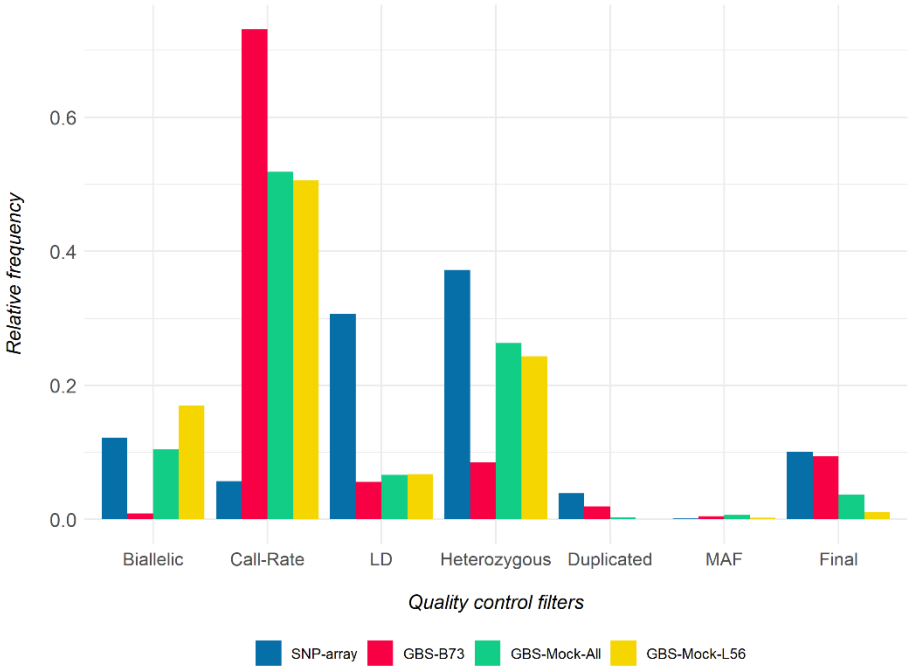

Fig. S1 Relative frequency of removed markers by quality control filters of genomic data for all SNPs datasets.

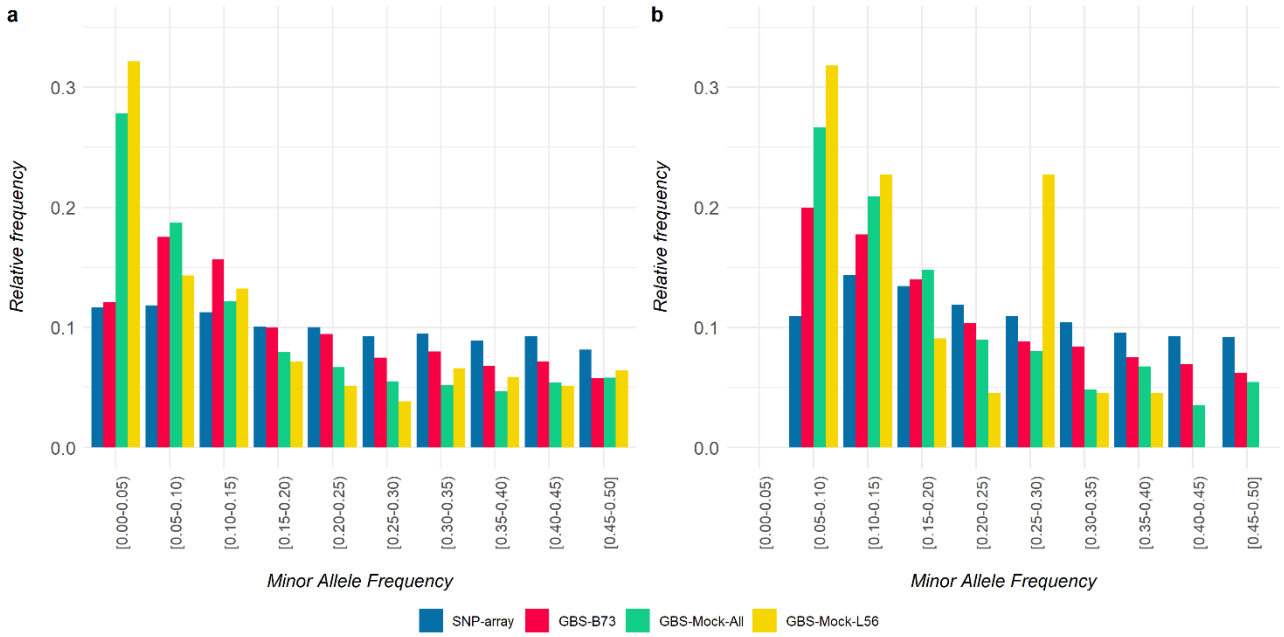

Fig. S2 Distribution of minor allele frequency for SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56 markers. **a** inbred lines; **b** hybrids.

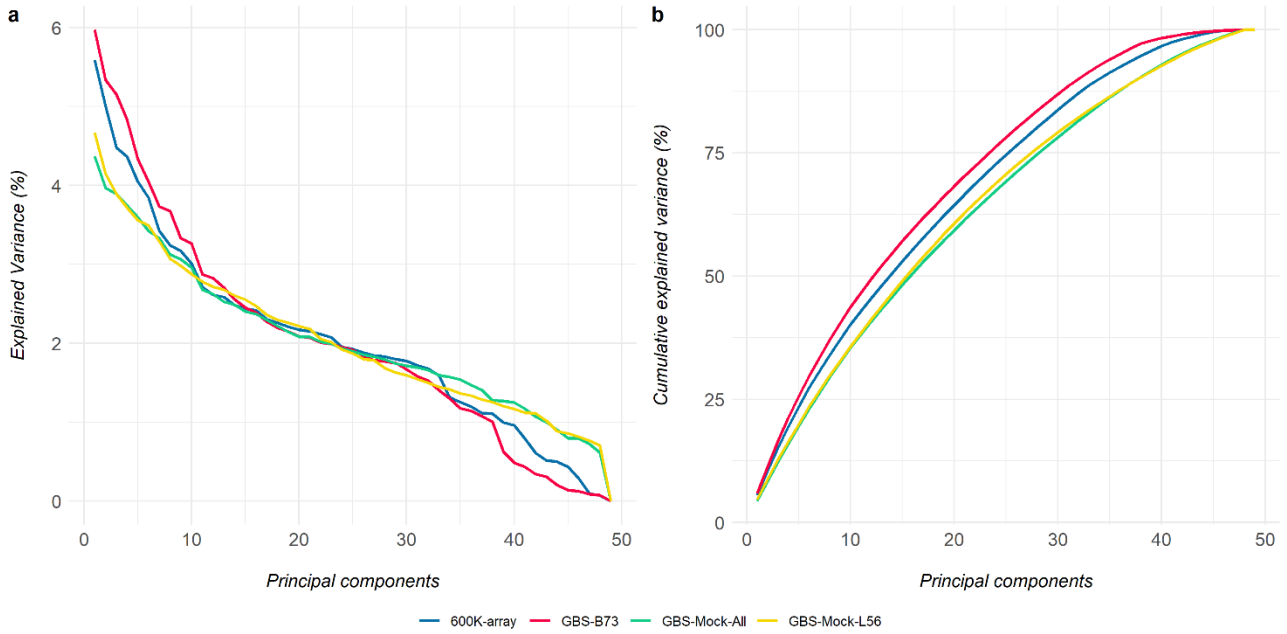

Fig. S3 a Variance explained by the principal components (PCA) from SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56 SNP datasets for 49 tropical inbred lines; **b** Cumulative explained variance estimated by principal components from SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56 SNP datasets for 49 tropical inbred lines.

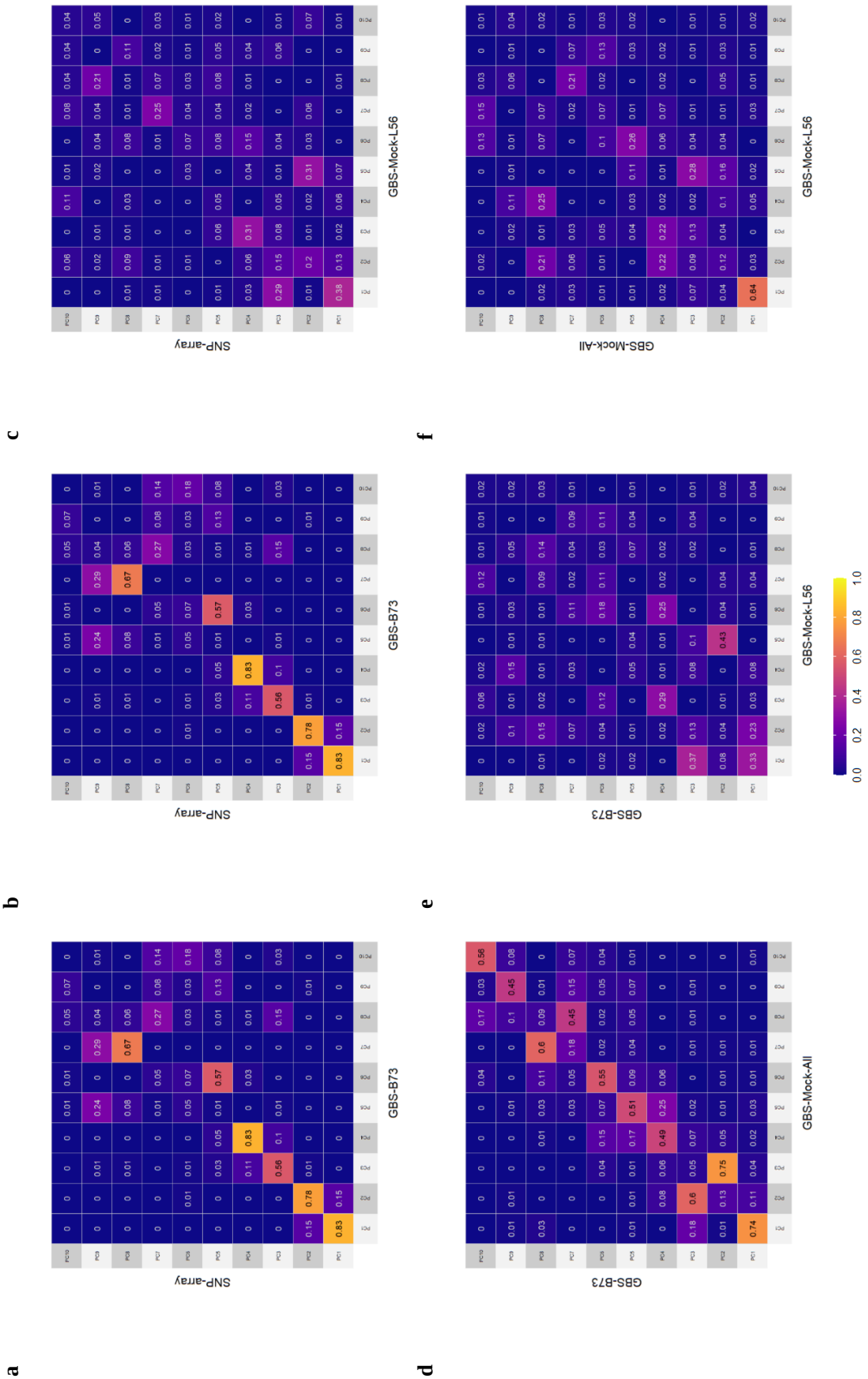

Fig. S4. Heatmap of the coefficient of determination (R^2) of the ten first eigenvectors between **a** SNP-array and GBS-B73; **b** SNP-array and GBS-Mock-All; **c** SNP-array and GBS-Mock-L56; **d** GBS-B73 and GBS-Mock-All; **e** GBS-B73 and GBS-Mock-L56; and **f** GBS-Mock-All and GBS-Mock-L56.

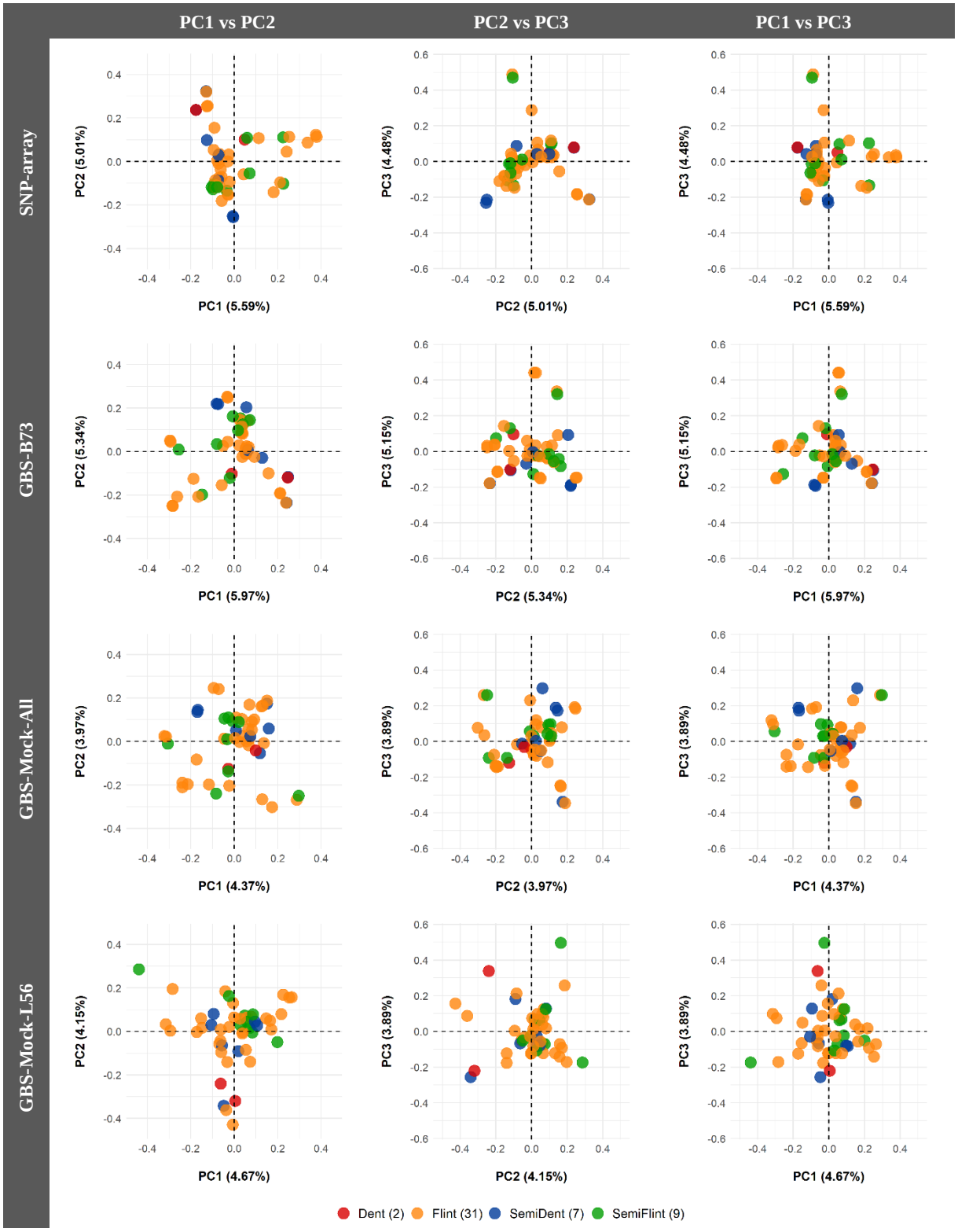

Fig. S5. Bi-plot among three first principal components analysis using the SNP-array, GBS-B73, GBS-Mock-All, and GBS-Mock-L56 SNP datasets for 49 tropical inbred lines. Explained variance percentages of each principal component are into parentheses. Grain types were used to color-coded inbred lines.

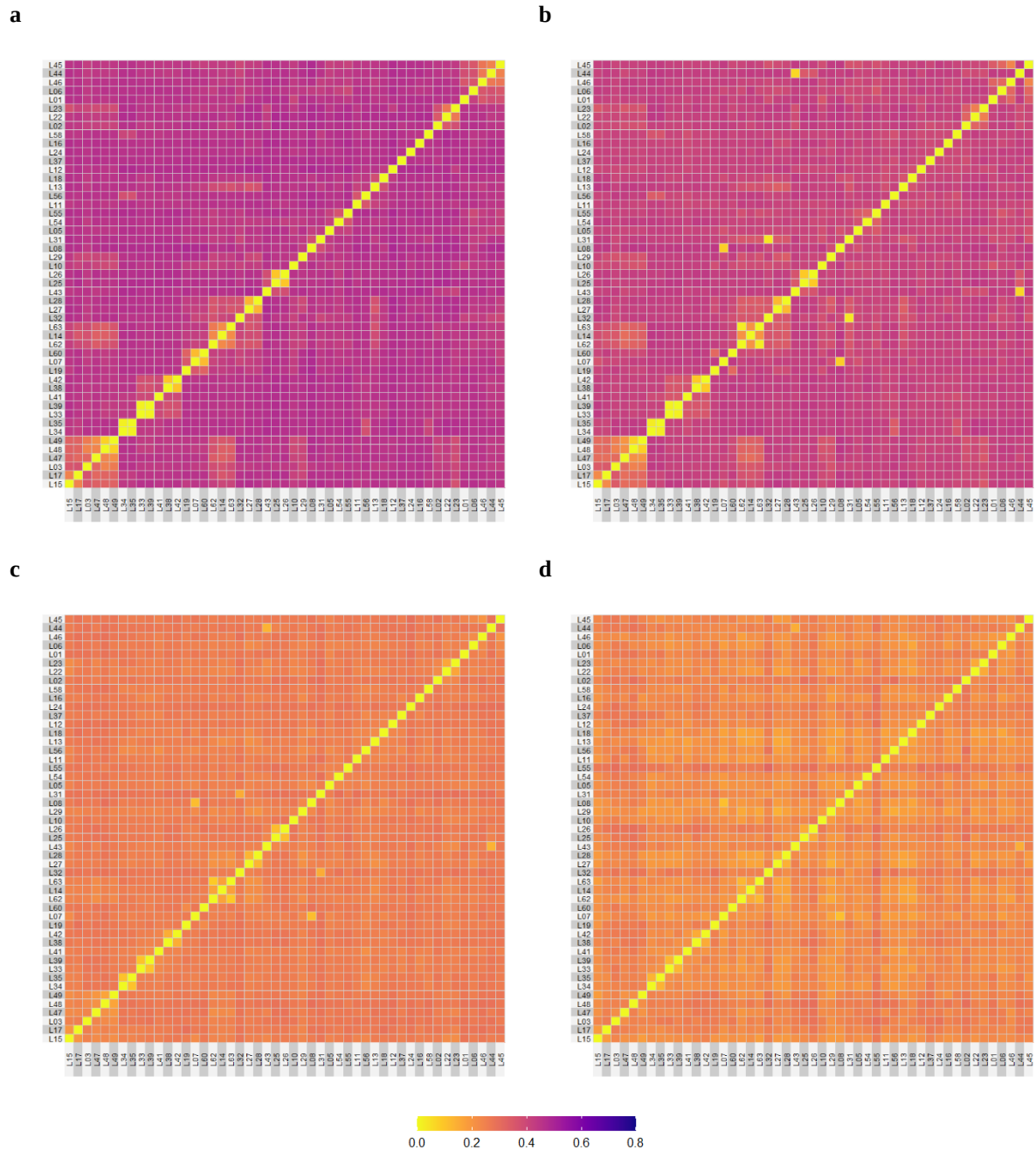

Fig. S6. Heatmaps of the Rogers genetic distance matrices estimated from **a** SNP-array, **b** GBS-B73, **c** GBS-Mock-All, and **d** GBS-Mock-L56 SNP datasets for 49 tropical inbred lines. Lines and columns of each plot were clustered according to the Euclidian distance performed in the Roger genetic distance matrix from the SNP-array dataset.

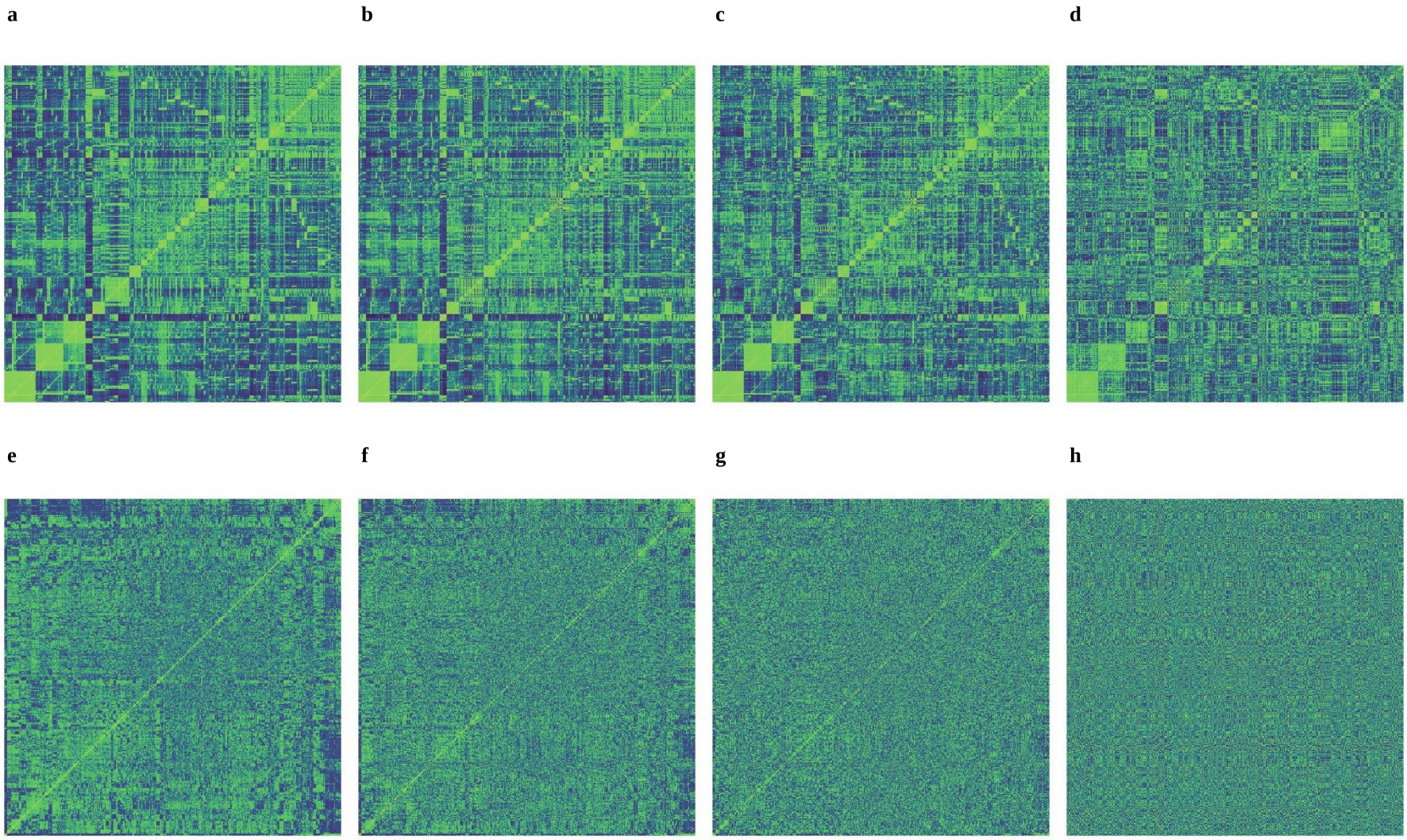

Fig. S7 Heatmaps of the (**a**, **b**, **c**, and **d**) additive genomic relationship ($\boldsymbol{G_a}$), and (**e**, **f**, **g**, and **h**) dominance genomic relationship ($\boldsymbol{G_d}$) matrices estimated from (**a** and **e**) SNP-array, (**b** and **f**) GBS-B73, (**c** and **g**) GBS-Mock-All, and (**d** and **h**) GBS-Mock-L56 SNP datasets for 903 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.

4. GENERAL CONSIDERATIONS

We present two approaches involving the concept of prediction using different methods. The prediction of haploid seeds via CNN models showed high accuracy, especially for the putative haploid class (97%). Furthermore, our CNN model accurately predicted other germplasms, which guarantees its transferability. However, this model was unable to recognize the true haploid seeds within the putative haploid seeds. Finally, we made it available to the scientific community to be used in the DH pipeline.

Using mock genomes for SNP discovery to diversity studies and genomic prediction of hybrids delivered reliable estimates comparable to SNP from standard genotyping platforms. For simple traits and considering only additive effects, the genomic predictions were similar to other genotyping platforms. However, for complex traits and estimation of dominance effects, genomic predictions were slightly worse. Finally, the strategy of using mock genomes can be a worthy alternative, especially for species where the reference genome is not available.

Both studies lead innovative tools that can assist plant breeding. Studies using deep learning models will leverage breeding pipelines. In the case of DH production, specifically in haploid identification, future studies related to false positives reduction are crucial for process optimization. Besides, studies about the automation of haploid screening will be attractive to leverage efficiency in the process. Concerning mock genomes for SNP calling, new studies regarding a higher diversity panel should add knowledge about population structure assessment using these genomes. Furthermore, using mock genomes to call SNPs for orphan species can leverage new breeding schemes and increase their genetic gain.