GUILHERME DE JONG

GENOMIC PREDICTION AND GENOME-WIDE ASSOCIATION STUDY: AN APPLICATION OF QUANTITATIVE GENETICS IN PLANT BREEDING PROGRAMS

LAVRAS – MG
2021

GUILHERME DE JONG

GENOMIC PREDICTION AND GENOME-WIDE ASSOCIATION STUDY: AN APPLICATION OF QUANTITATIVE GENETICS IN PLANT BREEDING PROGRAMS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Prof. Dr. Renzo Garcia Von Pinho
Orientador

Prof. Dr. Timothy Mathes Beissinger
Coorientador

LAVRAS – MG
2021

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

de Jong, Guilherme.
Genomic Prediction and Genome-Wide Association Study: An Application of Quantitative Genetics in Plant Breeding Programs / Guilherme de Jong. - 2021.
83 p.

Orientador(a): Renzo Garcia Von Pinho.
Coorientador(a): Timothy Mathes Beissinger.
Tese (doutorado) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. Maize (Zea mays L). 2. Genomic Selection. 3. Association Mapping. I. Garcia Von Pinho, Renzo. II. Beissinger, Timothy Mathes. III. Título.

GUILHERME DE JONG

GENOMIC PREDICTION AND GENOME-WIDE ASSOCIATION STUDY: AN APPLICATION OF QUANTITATIVE GENETICS IN PLANT BREEDING PROGRAMS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

APROVADA em 28 de setembro de 2021.

Dr. José Maria Villela Pádua	UFLA
Dra. Marcela Pedroso Mendes Resende	UFG
Dr. Roberto Fritsche-Neto	IRRI

Prof. Dr. Renzo Garcia Von Pinho
Orientador

Prof. Dr. Timothy Mathes Beissinger
Coorientador

LAVRAS – MG
2021

Aos meus pais, Robert e Edenéia.
Aos meus irmãos, Gabriele e Gustavo.
Dedico

AGRADECIMENTOS

À Deus, agradeço.

Ao meu orientador, Renzo Garcia Von Pinho, pela orientação, incentivo e confiança. Pela sua pronta disponibilidade em atender todas a demandas necessárias para a realização deste trabalho.

Ao Marcio Balestre (in memoriam) pela idealização deste trabalho. Agradeço também pela orientação confiança, ensinamentos e conversas ao longo dos anos que pudemos conviver.

Ao Tim Bessinger (University of Göttingen), por ter aceitado prontamente fazer parte da minha orientação. Agradeço pelas conversas, ensinamentos e oportunidade.

Ao John Hickey (The Roslin Institute) pela oportunidade fazer parte do seu grupo durante o doutorado sanduíche. Obrigado pelos ensinamentos e conselhos.

Ao Chris Gaynor (The Roslin Institute) por sempre prontamente estar disposto em me ajudar. Agradeço pelas conversas e ensinamentos. Ao Gregor, pela paciência e por sempre me instigar a pensar de outra forma. Obrigado por contribuírem para o meu crescimento pessoal e profissional.

À Universidade Federal de Lavras, ao Departamento de Biologia, pela oportunidade de realização deste trabalho. Ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas e aos professores que contribuíram para o meu desenvolvimento acadêmico e pessoal.

Ao Núcleo de Estudos em Genética e Melhoramentos de Plantas – GEN pela aprendizagem e pelas amizades proporcionadas ao longo dos anos. Agradeço a contribuição na minha formação pessoal e profissional.

Aos meus amigos, Maiara e Vitor, pela amizade ao logo desses anos. Muito obrigado pelos momentos de discussão científica, mas também pelos momentos de descontração.

Aos meus amigos que conheci durante o doutorado sanduíche. Romana, Owen, Christian, Luciano, Lorena, Jon e Troy. Obrigado pelas conversas, apoio e paciência. Muito obrigado por me fazerem me sentir em casa mesmo distante de casa.

À Coordenação de Aperfeiçoamento de Pessoal do Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro.

A aqueles que diretamente ou indiretamente puderam contribuir para o meu crescimento pessoal e profissional.

Muito obrigado!

RESUMO

O desenvolvimento de novas ferramentas e avanços em tecnologias genômicas de alto rendimento têm facilitado a seleção genômica e a identificação regiões causais, especialmente de características complexas. Dessa forma, a disponibilidade de marcadores abundantes e baratos possibilitou a exploração das informações dos marcadores em programas de melhoramento. As ferramentas mais comuns usadas em programas de melhoramento que exploram a cobertura densa de marcadores são a predição genômica e estudos de associação genômica. Na predição genômica, os parâmetros do marcador são estimados a partir de um conjunto de dados de treinamento com indivíduos genotipados e fenotipados. Posteriormente, o modelo treinado é usado para predizer o desempenho de indivíduos que são apenas genotipados. Já os estudos de associação do genoma testam associações marcador-característica que podem ser responsáveis pela variação causal. Nós investigamos o desempenho de diferentes modelos de predição genômica para selecionar pais no estágio inicial de um programa de melhoramento híbrido usando a capacidade geral de combinação estimada e seu impacto na precisão da seleção e no ganho genético de longo prazo. Avaliamos o desempenho de cinco modelos de predição genômica sob diferentes densidades de marcadores SNP ou genótipos QTL usando simulações estocásticas de um programa de melhoramento híbrido completo. Nós também investigamos a capacidade de GWAS univariada e multivariada de identificar marcadores ligados a loci que contribuem para a resistência à podridão da espiga causada por Diplodia ou Fusarium ou ambas as doenças em linhagens de milho. Avaliamos as abordagens univariada e multivariada usando um painel diverso de milho avaliado para três características diferentes.

Palavras-chave: Milho (Zea mays L.); Predição Genômica; Seleção de Pais; Mapeamento Associativo; Podridão de Espiga.

ABSTRACT

The development of new tools and advances in high throughput genomic technologies have facilitated genomic selection the identification of sources of variation, especially of complex traits. Therefore, the availability of abundant and cheap markers made it possible to exploit the marker information in breeding programs. The most common tools used in breeding programs that exploit the dense marker coverage are genomic prediction and genome-wide association studies. In the genomic prediction, marker parameters are estimated from a training dataset with genotyped and phenotyped individuals. Subsequently, the trained model is used to predict performance for individuals that are only genotyped. On the other hand, genome-wide association studies test marker-trait associations that may be responsible for the causal variation of interest. We investigated the performance of different genomic prediction models to select parents in the early stage of a hybrid breeding program using estimated general combining ability and their impact on selection accuracy and long-term genetic gain. We evaluated the performance of five genomic prediction models under different SNP marker densities or QTL genotypes using stochastic simulations of an entire hybrid breeding program. We also investigated the ability of univariate and multivariate GWAS identifying markers linked to loci that contribute to resistance to Diplodia ear rot or Fusarium ear rot or both diseases in maize inbred lines. We evaluated the univariate and multivariate approaches using a maize diverse panel evaluated for three different traits.

Keywords: Maize (*Zea mays* L.); Genomic Prediction; Parent Selection; Association Mapping; Ear Rot.

SUMÁRIO

PRIMEIRA PARTE

1. GENERAL INTRODUCTION

Plant breeding programs have been successful and efficient in providing improved cultivars over the centuries. Basically, the breeding process relies upon the identification of natural and mutant induced genetic variation and in the selection and the recombination of the best individuals. The classical evaluation and selection methods have essentially been based on phenotypic values (PROHENS, 2011). In the last decades, the development of novel tools and advances in high throughput genomic technologies have facilitated the identification of sources of variation, especially of complex traits. With this, it became possible to exploit the markers information in breeding programs. The dense maker coverage made it possible to develop the genomic selection, which estimates all marker effects together and predicts the total genetic variance by summing all estimated marker effects (BERNARDO; YU, 2007; HEFFNER; SORRELLS; JANNINK, 2009; MEUWISSEN; HAYES; GODDARD, 2001). Jointly with genomic selection, modern genomic technologies made possible the development of association mapping tools. This approach exploits natural diversity through the linkage disequilibrium between markers and QTL to identify marker-trait associations (YU et al., 2006; YU; BUCKLER, 2006; ZHU et al., 2008).

Genomic prediction consists of setting a statistical model and estimating its parameters from a training dataset with genotyped and phenotyped individuals. Subsequently, the trained model is used to predict phenotype performance for individuals that are only genotyped (MEUWISSEN; HAYES; GODDARD, 2001). The approach allows carrying selection without the need of further phenotypic evaluations. Genomic selection is an advantageous approach when there are too many candidates to phenotype and when the breeder wants to increase the genetic gain per unit time by increasing the generations of selection by year (BERNARDO, 2020). In hybrid breeding programs, the genomic prediction has been useful in predicting hybrid performance. Most studies have focused on predicting hybrid performance in the late stage of breeding programs, in which hybrids are generated from advanced inbred lines and testers from the opposite heterotic pools (ALBRECHT et al., 2011, 2014; MASSMAN et al., 2013; TECHNOW et al., 2012, 2014). Although hybrid prediction of untested hybrids in late stages of breeding programs have shown to be effective, it is still necessary to evaluate the genomic prediction of general combining ability (GCA) for inbred lines in the early stages of

breeding programs. The GCA of inbred lines is estimated based on phenotypic information and the new parents are selected in the late stages of the breeding program. Therefore, genomic prediction can be used to identify promising inbred lines sooner and use them as parents of subsequent breeding cycles earlier in a breeding program.

In the genome-wide association study, differently of genomic selection, there is a systematic search in the genome for a causal genetic variation. A large number of markers are tested for association with the trait of interest. This approach identifies quantitative trait loci (QTL) by examining the marker-trait associations that can be attributed to the strength of the linkage disequilibrium between markers and functional polymorphisms. This approach uses a set of diverse germplasm, so it is possible to exploit evolutionary and historical recombination events (NORDBORG; TAVARÉ, 2002; ZHU et al., 2008). Several studies have shown that GWAS is an effective approach for identifying loci underlying variation for traits (XIAO et al., 2017). Although many traits are evaluated, the association analyses are performed using a single trait. Multivariate approaches have been suggested to exploit the complexity and quantity of traits available. In plant breeding programs, diseases are one of the most limiting factors in production. Ear rots are one of the most important diseases in maize. Therefore, the multivariate GWAS is an interesting approach for the identification of markers linked to loci related to the resistance.

In the first chapter, we investigated the performance of different genomic prediction models to select parents in the early stage of a hybrid breeding program using estimated general combining ability and their impact on selection accuracy and long-term genetic gain. We evaluated the performance of five genomic prediction models under different SNP marker densities or QTL genotypes using stochastic simulations of an entire hybrid breeding program. In the second chapter, we investigated the ability of univariate and multivariate GWAS to identifying markers linked to loci that contribute to resistance to Diplodia ear rot or Fusarium ear rot or both diseases in maize inbred lines. We evaluated the univariate and multivariate approaches using a maize diverse panel evaluated for three different traits.

REFERENCES

ALBRECHT, T. et al. Genome-based prediction of testcross values in maize. **Theoretical and Applied Genetics**, v. 123, n. 2, p. 339–350, 1 jul. 2011.

ALBRECHT, T. et al. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. **Theoretical and Applied Genetics**, v. 127, n. 6, p. 1375–1386, 1 jun. 2014.

BERNARDO, R. N. **Breeding for quantitative traits in plants**. 3rd ed ed. Woodbury, Minn: Stemma Press, 2020.

BERNARDO, R.; YU, J. Prospects for Genomewide Selection for Quantitative Traits in Maize. **Crop Science**, v. 47, n. 3, p. 1082–1090, maio 2007.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J.-L. Genomic Selection for Crop Improvement. **Crop Science**, v. 49, n. 1, p. 1–12, jan. 2009.

MASSMAN, J. M. et al. Genomewide predictions from maize single-cross data. **Theoretical and Applied Genetics**, v. 126, n. 1, p. 13–22, jan. 2013.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, abr. 2001.

NORDBORG, M.; TAVARÉ, S. Linkage disequilibrium: what history has to tell us. **Trends in Genetics**, v. 18, n. 2, p. 83–90, fev. 2002.

PROHENS, J. Plant Breeding: A Success Story to be Continued Thanks to the Advances in Genomics. **Frontiers in Plant Science**, v. 2, 2011.

TECHNOW, F. et al. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. **Theoretical and Applied Genetics**, v. 125, n. 6, p. 1181–1194, out. 2012.

TECHNOW, F. et al. Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. **Genetics**, v. 197, n. 4, p. 1343–1355, ago. 2014.

XIAO, Y. et al. Genome-wide Association Studies in Maize: Praise and Stargaze. **Molecular Plant**, v. 10, n. 3, p. 359–374, mar. 2017.

YU, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, v. 38, n. 2, p. 203–208, fev. 2006.

YU, J.; BUCKLER, E. S. Genetic association mapping and genome organization of maize. **Current Opinion in Biotechnology**, v. 17, n. 2, p. 155–160, abr. 2006.

ZHU, C. et al. Status and Prospects of Association Mapping in Plants. **The Plant Genome**, v. 1, n. 1, p. plantgenome2008.02.0089, jul. 2008.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 - COMPARISON OF GENOMIC PREDICTION MODELS FOR GENERAL COMBINING ABILITY IN EARLY STAGES OF HYBRID BREEDING PROGRAMS

ABSTRACT

Genomic prediction studies in maize have a primary focus on the prediction of hybrids in the late stages of the breeding program and have largely ignored the selection of inbred lines for use as parents in subsequent breeding cycles. The aim of this study was to evaluate the performance of different genomic prediction models to select parents in the early stage of a hybrid breeding program using estimated general combining ability and their impact on selection accuracy and long-term genetic gain. We evaluated the performance of genomic prediction models for the selection of parents in subsequent breeding cycles using estimated general combining ability. Five genomic prediction models were evaluated under different scenarios, SNP marker densities or true QTL genotypes, using stochastic simulations of entire maize breeding programs. The simulated maize breeding programs were modeled over 20 years of breeding using AlphaSimR. The performance of the genomic prediction models was measured at the double haploid stage by tracking genetic gain and heterosis for all possible hybrids formed by crossing all inbred lines from different heterotic pools. The results showed that the performance of the genomic prediction models depends on the marker density. Under low-density, the models that included pool-specific effects showed higher genetic gain than the models that have a common additive effect for both heterotic pools. In contrast, the performance of the models was similar when high-density markers were used. True QTL genotypes showed the superiority of the models that included dominance effects. For heterosis, under low-density, the genomic prediction models showed a different performance. While using high-density markers and true QTL genotypes, the models that included dominance effects showed higher heterosis than the models that included only additive effects. To sum up, the genetic gain and heterosis increased with higher marker density, the performance of the genomic prediction models depends on the marker density and, there is a significant difference in the genetic gain between markers and true QTL genotypes.

Keywords: Stochastic Simulations; Genomic Prediction; Parent Selection.

1. INTRODUCTION

In this study, we evaluated the performance of genomic prediction models to select inbred lines as parents of subsequent breeding cycles based on their general combining ability. Breeding programs are structured in different components. One of these components is inbred line development, in which inbred lines are selected based on general combining ability (GCA), which is the average performance of a line in hybrid combinations. GCA is used to identify the best inbred lines that can be used as parents of new hybrids or new inbred lines. Genomic prediction has revolutionized the evaluation of per-se and hybrid performance. Consequently, genomic prediction could also identify promising inbred lines and use them as parents of subsequent breeding cycles in the earlier stages of a hybrid breeding program.

Breeding programs are divided into different components of which inbred line development and hybrid development are arguably the most important (EATHINGTON et al., 2007). The inbred line development component focuses on the early stages of a breeding program to develop new inbred lines. The two main goals at this stage are to identify inbred lines as parents of new hybrids and as parents of subsequent breeding cycles. The hybrid development comes at the late stage of a breeding program to test inbred line combinations and commercialize the best ones.

At the early stages, the highest performing inbred lines are selected based on their combining ability. The combining ability is used to rank inbred lines according to their hybrid performance (HALLAUER; CARENA; MIRANDA FILHO, 2010). The concept was first presented by Sprague and Tatum (1942), where they defined general combining ability (GCA) as the average performance of a line in hybrid combinations and specific combining ability (SCA) as deviation in performance from the GCA of the lines in hybrid combinations. To evaluate the combining abilities of inbred lines we would require a complete factorial trial to test the possible hybrids between the inbred lines from opposite heterotic pools. However, complete factorial trials are impractical with a large number of inbred lines. Especially with the broad use of the double haploid (DH) technology in plant breeding, the development of new inbred lines has accelerated (GEIGER; GORDILLO, 2009). As a result, general combining ability of inbred lines is assessed via testcross trials in the line development component. In testcross trials, inbred lines are crossed with a limited number of testers and their hybrid performance is evaluated to estimate GCA of the inbred lines. These testcross trials are planted in a limited number of locations and years to balance an expected genetic gain against costs of these trials.

Genomic prediction is now an established strategy to evaluate and select the best individuals in plant breeding programs (BERNARDO; YU, 2007; HEFFNER et al., 2010). Genomic prediction consists of setting a statistical model and estimating its parameters from a training dataset with genotyped and phenotyped individuals. Subsequently, the trained model is used to predict phenotype performance for individuals that are only genotyped (MEUWISSEN; HAYES; GODDARD, 2001). So, it is possible to predict the performance of individuals without testing them all in trials. Most genomic prediction models consider only breeding values. Although breeding values capture a large part of total genetic variation (FALCONER; MACKAY; FRANKHAM, 1996; HILL; GODDARD; VISSCHER, 2008), they miss a part of dominance and epistatic components of genetic variation, which are important in hybrid crops. The inclusion of both additive and dominance effects may increase genomic prediction accuracy, and guide exploitation of complementarity between heterotic pools and heterosis (SU et al., 2012; TORO; VARONA, 2010; VARONA et al., 2018; WELLMANN; BENNEWITZ, 2012).

In hybrid breeding programs, genomic prediction has been useful in predicting hybrid performance. Most studies have focused on predicting hybrid performance in the late stage of breeding programs, in which hybrids are generated from advanced inbred lines and testers from the opposite heterotic pool (ALBRECHT et al., 2011, 2014; MASSMAN et al., 2013; TECHNOW et al., 2012, 2014). In another study, (Kadam et al. (2016) evaluated the potential of genomic prediction to identify superior hybrids at the early stage of a breeding program. Although these studies have shown the value in genomic prediction of untested hybrids, we still have to evaluate genomic prediction of GCA for inbred lines. Traditionally, the GCA of inbred lines was estimated from phenotype information collected from their testcrosses. Furthermore, new parents of subsequent breeding cycles are selected late in a breeding program to accumulate the phenotype information. Within this context, genomic prediction can be used to identify promising inbred lines sooner and use them as parents of subsequent breeding cycles earlier in a breeding program. Thus, genomic prediction would be an efficient tool to reduce the breeding cycle and with this increase genetic gain per year (HEFFNER; SORRELLS; JANNINK, 2009).

The aim of this study was to evaluate the performance of different genomic prediction models to select parents in the early stage of a hybrid breeding program using estimated general combining ability and their impact on selection accuracy and long-term genetic gain. We evaluated the performance of five genomic prediction models under different SNP marker densities or QTL genotypes using stochastic simulations of an entire hybrid breeding program. The results show that the performance of the genomic prediction models is dependent on SNP

marker density and type of a model: higher SNP marker densities increase the rate of genetic gain and heterosis, and models with average or additive effects specific to each heterotic pool as well as dominance effects provide a better fit.

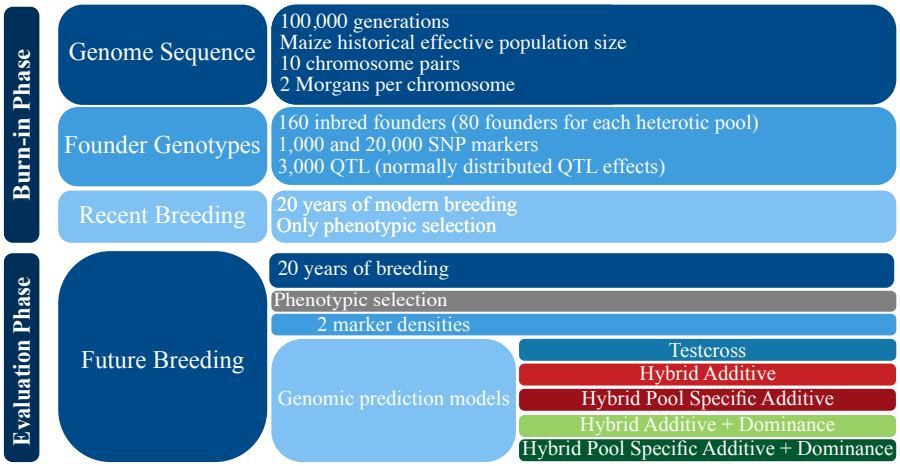
2. MATERIALS AND METHODS

Stochastic simulations of a hybrid crop breeding program were used to compare the performance of different genomic prediction models for selection of inbred lines as parents of subsequent breeding cycles based on their predicted general combining ability. Different scenarios were simulated to test five genomic prediction models using two SNP marker densities (low and high-density) or QTL at the early stage of a hybrid breeding program.

The breeding programs were compared across 40 years using 100 replicates of stochastic simulations. The simulations consisted of two phases, burn-in and future breeding phase (Figure 1). The first phase of the simulations consisted of burn-in, which was identical for all scenarios. In the burn-in a conventional breeding program was simulated for 20 years based only on phenotypic selection. After the burn-in phase, genomic prediction and subsequent selection were evaluated using two SNP marker densities or QTL in the future breeding program for another 20 years. For the future breeding program, a double haploid breeding scheme was used, and genomic prediction guided selection of new parents and also, to advance individuals to the next stages in the breeding program.

The genomic prediction models used in simulations were Testcross model, Hybrid Additive model, Hybrid Pool Specific Additive model, Hybrid Additive + Dominance model, and Hybrid Pool Specific Additive + Dominance model. The performance of the genomic prediction models was measured at the DH stage by tracking genetic gain and heterosis for hybrids formed by crossing inbred lines from different heterotic pools along the 20 years of the future breeding program.

Figure 1 – Overview of the simulation phases for the hybrid breeding program.



2.1. Burn-In Phase

All scenarios shared the same burn-in phase to enable comparison. The burn-in phase was divided into three stages. In the first stage, the whole genome sequence data of maize was simulated. In the second stage, founder genotypes and phenotypes were simulated. And in the third stage, 20 years of the recent breeding program using the conventional breeding strategy was simulated.

Whole Genome Sequence Data

Firstly, the whole genome sequence data was generated. To mimic the maize genome, 10 chromosome pairs were simulated and the genome sequence for each maize chromosome was simulated using the Markovian Coalescent Simulator (CHEN; MARJORAM; WALL, 2009) within AlphaSimR (GAYNOR; GORJANC; HICKEY, 2020) package in R (R CORE TEAM, 2019). Each chromosome had a genetic length of 2 Morgans and a physical length of 2×10^8 base pair. Recombination rate was set to 1.25×10^{-8} base pair and mutation rate was set to 1×10^{-8} per base pair. Simulation involved variable historical effective population size over time as follows; 100,000 at 6,000 generations ago, 10,000 at 2,000 generations ago, 5,000 at 1,000 generations ago, 1,000 at 100 generations ago. To mimic the genetic separation of two heterotic groups we induced a population split 100 generations ago. The final effective population size at the end of the coalescent simulation was set to 100 for each heterotic group.

Founder Genotypes

The founders were created to serve as initial parents in the recent breeding phase. The founder genotypes were created by randomly sampling gametes from the simulated genome sequences. Segregating sites in the founder sequences were randomly selected to serve as 100 and 2,000 Single Nucleotide Polymorphism (SNP) markers per chromosome (1,000 and 20,000 SNP markers in total per genome) for low- and high-density coverage, respectively. Additionally, 300 Quantitative Trait Loci (QTL) per chromosome were sampled from the segregating sites (3,000 QTL in total per genome). The segregating sites for SNP markers and QTL did not overlap. The founders were converted to inbred lines by simulating formation of double haploid lines and split into two heterotic pools, with 80 parents in each heterotic pool.

Founder Phenotypes

The phenotypes for a trait representing grain yield, in bushels per acre, were simulated. The genetic value of the trait was assumed to be controlled by 3,000 QTL. To each QTL we assigned an additive effect and a dominance effect resulting from the interactions between alleles at a heterozygous locus. In this simulation the epistatic gene action was not considered. The QTL additive effects (a_i) were sampled from a normal distribution as:

$$\alpha_i \sim N\left(0, \sigma_a^2\right)$$

where σ_a^2 is the variance of QTL additive effects which was scaled to achieve a genetic variance of 20 among the founder inbred lines. The QTL dominance effects (d_i) followed:

$$d_i = \delta_i \left| a_i \right|,$$

where δ_i is a QTL specific dominance deviation and a_i is the absolute value of the QTL additive effect. QTL Dominance deviations (δ_i) were sampled from a normal distribution as:

$$\delta_i \sim N(\mu_\delta, \sigma_\delta^2),$$

where is μ_δ the mean dominance deviation, which was set to 0.92, and σ_δ^2 is the variance of dominance deviations, which was set to 0.2. These values were chosen to obtain relative rates of genetic gain in inbred lines and hybrids that approximately match historical trends observed in a maize commercial breeding program (TROYER; WELLIN, 2009). The genetic value of each individual was defined as the sum of additive and dominance effects across all QTL.

The phenotypic values for each individual were generated by adding random error to the genetic value. The random error was sampled from a normal distribution with mean zero. The error variance varied according to the stage of evaluation in the breeding program. In order to account for increasing accuracy in evaluation, due to the increase of plot size and replications per entry, the values for these error variances were set so as to achieve target levels of heritability as indicated in the following.

2.2. Recent Breeding Program (Burn-in) and Phenotypic Selection (Future Breeding Program)

Recent breeding program (burn-in) was simulated for 20 years of breeding with a conventional program (without genomic prediction). The design of the burn-in program was modelled based on existing commercial maize breeding programs (BERNARDO, 2020; Table 1.2; POWELL et al., 2020). The main features of the recent (burn-in) breeding program were:

i) a crossing block of inbred lines in each heterotic pool used to develop biparental populations each year;
ii) development of double haploids from each biparental cross;
iii) a three-year cycle-time from crossing to the selection of new parents; and
iv) a six-year production interval from crossing to release a new commercial hybrid.

The selection performed in the burn-in program was based on yield trial phenotypes. The levels of heritability at each selection stage were adapted from the number of DH lines and locations reported in Bernardo (2020; Table 1.2). The overall design of the burn-in breeding program is given in Figure 2 and a detailed description follows. The progression of the breeding program was simulated using *AlphaSimR* package (GAYNOR; GORJANC; HICKEY, 2020) and R (R CORE TEAM, 2019). The design of the hybrid breeding program was based on the study of Powell et al. (2020).

Year 1

At this stage, 80 biparental populations were created by crossing the inbred lines within each heterotic pool. Each of the 80 parental lines was used as male or female only once. From each cross 50 F_1 double haploid lines were produced in each heterotic pool (Geiger and Gordillo, 2009). The 4,000 DH inbred lines were planted in separate plots and no selection was performed. Each DH inbred line was crossed to a single inbred tester from the opposite heterotic pool.

Year 2

The 4,000 DH testcrosses were evaluated in the testcross 1 (TC1) stage. The testcrosses were evaluated in unreplicated trials, two-row plots at six locations. The selection in the TC1 stage was modelled as the selection on a yield phenotype with heritability of 0.54. The 400 best DH lines were selected based on general combining ability (GCA) to advance to the next trial.

Each of 400 selected individuals were crossed to three inbred testers from the opposite heterotic pool.

Year 3

The 1,200 DH testcrosses were evaluated in the testcross 2 (TC2) stage. The testcrosses were evaluated in unreplicated trials, two-row plots at 12 locations. The selection in the TC2 stage was modelled as the selection on a yield phenotype with heritability of 0.71. The 40 best DH lines were selected based on general combining ability (GCA) to advance to the next trial. Each of 40 selected individuals were crossed to five "elite" DH inbred lines from the opposite heterotic pool.

Year 4

The 200 experimental hybrids were evaluated in the elite yield trial (EYT) stage. The experimental hybrids were evaluated in unreplicated trials, two-row plots at 24 locations. The selection in the YET stage was modelled as the selection on a yield phenotype with heritability of 0.82. The best performing 40 experimental hybrids were advanced to the next trial.

Year 5

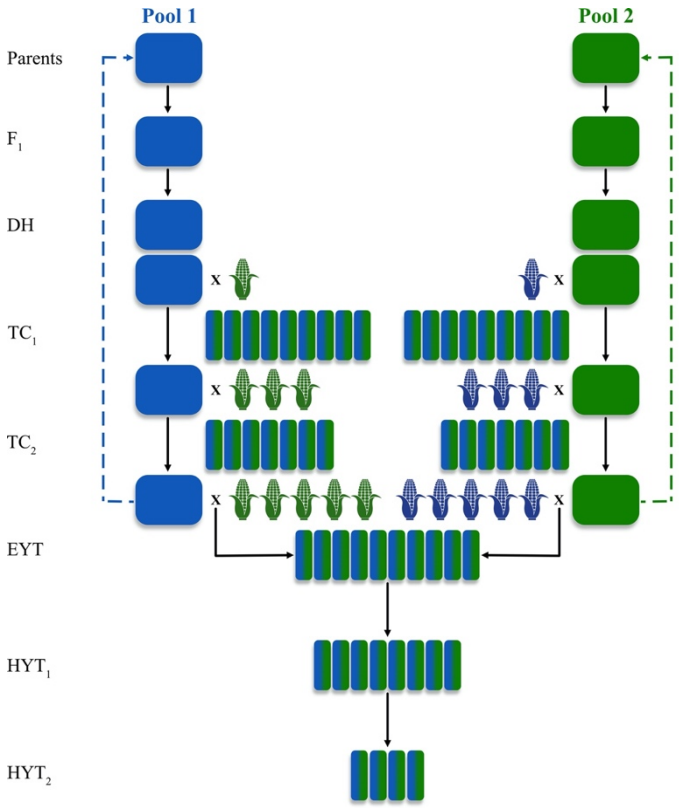
The 40 experimental hybrids were evaluated in the hybrid yield trial 1 (HYT1) stage. The experimental hybrids were evaluated in unreplicated trials, two-row plots at 48 locations. The selection in the HYT1 stage was modelled as the selection on a yield phenotype with heritability of 0.98. The best performing eight experimental hybrids were advanced to the next trial.

Year 6

The eight experimental hybrids were evaluated in the hybrid yield trial 2 (HYT2) stage. The experimental hybrids were evaluated in on-farm strip tests of pre-commercial hybrids at 600 locations. The selection in the HYT2 stage was modelled as the selection on a yield

phenotype with heritability of 0.99. The best performing pre-commercial hybrids in this stage were released as commercial hybrids.

Figure 2 – Overview of hybrid breeding pipeline for the conventional program used in burn-in and phenotypic selection. The heterotic pools are represented by the colors blue and green, and the squares represents different stages of the breeding program. The "maize ears" represent testers from opposite heterotic pool used in the testcross yield trials and greenish-blue squares represent the hybrids generated by the crosses between the inbred lines from the different heterotic pools. The dashed lines show the stage in which inbred candidates are selected as parents for subsequent breeding

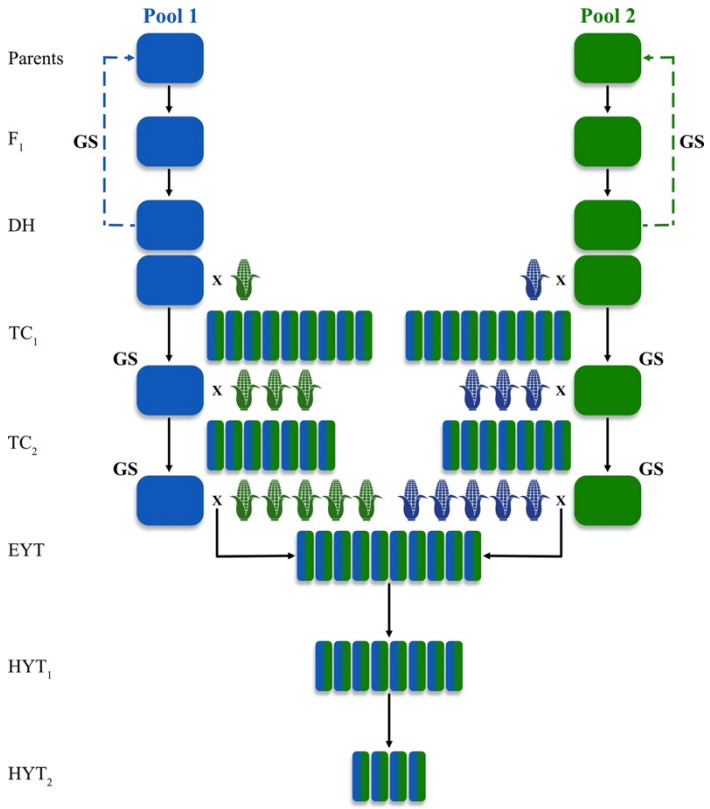


2.3. Future Breeding

After the 20 years of the recent breeding program (burn-in), based only on phenotypic selection, the future breeding phase was simulated. In the future breeding phase, alternative breeding programs were simulated with different marker scenarios and genomic prediction models. Each breeding program was simulated for an additional 20 years following the common burn-in breeding. Therefore, each alternative breeding program could be evaluated with an equivalent start point. The alternative breeding programs were modeled to compare the five genomic prediction models under different marker scenarios (low- and high-density SNP

markers or QTL), plus the phenotypic selection. All the alternative breeding programs were simulated based on the double haploid rapid cycle breeding program (Figure 3).

Figure 3 – Overview of hybrid breeding pipeline for the conventional program used in burn-in and phenotypic selection. The heterotic pools are represented by the colors blue and green, and the squares represents different stages of the breeding program. The "maize ears" represent testers from opposite heterotic pool used in the testcross yield trials and greenish-blue squares represent the hybrids generated by the crosses between the inbred lines from the different heterotic pools. The dashed lines show the stage in which inbred candidates are selected as parents for subsequent breeding cycles.



Double Haploid (DH) Rapid Cycle

The DH rapid cycle breeding program was designed based on a conventional breeding program as a template. The minimal changes were made to this template to keep the breeding program size similar to the standard breeding program, so they were more comparable. The main change in the DH rapid cycle breeding design compared to the conventional breeding design was the use of genomic prediction to select new parental lines for subsequent breeding cycle (parental recycling) and to advance candidate lines from the TC1 and TC2 stages. The DH rapid cycle reduced the minimum cycle time from the biparental cross to the selection of

new parents from 3 years in the conventional breeding program to 1 year. The operating costs of the breeding programs were not assessed in this present study.

Parental Recycling

For the selection of new parents for subsequent breeding cycles, the maximum avoidance of inbreeding was used as a crossing scheme. Maximum avoidance of inbreeding maintains uniform contributions and inbreeding coefficients across all crosses by ensuring that the crosses are made between the least related individuals (KIMURA; CROW, 1963; WRIGHT, 1921). In this study, the 2 best individuals from each family with the highest genomic prediction were selected, using the maximum avoidance scheme, as new parents for the subsequent breeding cycles (160 in total). Each individual produced 25 F_1 derived double haploid lines (4,000 in total). The 4,000 DH inbred lines were planted in separate plots and no selection was performed. Each DH inbred line was crossed to a single inbred tester from the opposite heterotic pool.

Testcross 1 and 2

The DH rapid cycle used genomic selection to advance candidate DH lines in the testcross 1 and testcross 2 stages. In the testcross 1, the 400 DH lines with the highest genomic prediction were selected to advance to the next trial. And in the testcross 2, the 40 DH lines with the highest genomic prediction were selected to advance to the next trial.

2.4. Genomic Prediction

Genomic Prediction Models

The genomic predictions were made by fitting models using the ridge regression best linear unbiased prediction (RRBLUP) framework with different effects(MEUWISSEN; HAYES; GODDARD, 2001; WHITTAKER; THOMPSON; DENHAM, 2000). The genotypes and phenotypes of the selection candidates were fitted differently using different genomic prediction models as indicated in the following, while the genetic background of the tester was accounted for by fitting a tester-by-stage fixed effect. Genomic predictions for each model were

calculated using the RRBLUP functions in AlphaSimR package (GAYNOR; GORJANC; HICKEY, 2020).

In year 1 of the alternative breeding programs, the prediction models were fitted using the EMMA algorithm (ZHANG et al., 2006). From year 2 onwards, the variance components estimated in the year 1 were used to fit the prediction model using the EM algorithm (DEMPSTER; LAIRD; RUBIN, 1977). We used this approach because the functions that use the EMMA algorithm involve a square matrix with dimensions equal to the number of phenotypic records and this approach uses a lot of memory when there are a large number of phenotypic records. Whereas the functions that use the EM approach involve a square matrix with dimensions equal to the number of fixed effects plus the number of loci (random effects) and the RRBLUP problem is solved by setting up and solving Henderson's mixed model equations. Consequently, this approach uses less memory compared to the functions that use the EMMA algorithm (GAYNOR; GORJANC; HICKEY, 2020).

In the following we describe the different genomic prediction models. The models differ in the way we fit genotype data and its association with phenotype data. Specifically, we are interested in the average effect of an allele substitution, or allele substitution effect for short, (α) at each locus, which is estimated by regressing phenotypes onto allele dosages (see more details in FALCONER, 1985; FISHER, 1941; PRICE, 1972). In a randomly mating population, allele substitution effect is a function of additive effect a, dominance effect d, and allele frequencies p and q at a locus: $\alpha = \mathrm{a} + \mathrm{d}(q - p)$ ((FALCONER, 1985; FALCONER; MACKAY; FRANKHAM, 1996; HIVERT; WRAY; VISSCHER, 2021; LYNCH; WALSH, 1998). The allele substitution is a "statistical" effect inferred for a population at hand and hence population dependent; due to being a function of the genotype composition of a population. Additive and dominance effects, sometimes referred to as "biological" effects, are not population dependent, assuming no interaction between additive and dominance effects, population's origin or environment. We have fitted the models using QTL or SNP marker genotypes. In the case of SNP marker genotypes the resulting estimates are not QTL effects, but only a proxy for QTL effects mediated by linkage-disequilibrium between QTL and SNP markers.

Testcross Model

The Testcross model fits allele substitution effects specific to each heterotic pool, which are fitted separately:

$$y = X\beta + Z\alpha + e, \quad (1)$$

where y is a vector $(n \times 1)$ of inbred line's average testcross performance; X is an incidence matrix $(n \times t)$ for fixed effects (tester-by-stage effects); Z is a genotype matrix $(n \times m)$ for inbred lines, this matrix has entries 0 and 2 respectively for inbred's genotypes AA and aa; β is a vector $(t \times 1)$ of fixed effects; α is a vector $(m \times 1)$ of allele substitution effects; and e is a vector $(n \times 1)$ of residuals. Genomic predictions from the Testcross model were calculated using the AlphaSimR function *RRBLUP* in the year 1 and from the year 2 onwards using the function *RRBLUP2*.

Hybrid Additive Model

The Hybrid additive model fits allele substitution effects for both heterotic pools, which are fitted simultaneously:

$$y = X\beta + Z\tau + e, \quad (2)$$

where y is a vector $(n \times 1)$ of hybrid's average performance; X is an incidence matrix $(n \times t)$ for fixed effects (tester-by-stage effects); Z is a genotype matrix $(n \times m)$ for hybrids, which contains 0, 1, and 2 respectively for hybrid's genotypes AA, Aa, and aa; β is a vector $(t \times 1)$ of fixed effects; τ is a vector $(m \times 1)$ of allele substitution effects; and e is a vector $(n \times 1)$ of residuals. Genomic predictions from the hybrid additive model were calculated using the AlphaSimR function *RRBLUP* in the year 1 and from the year 2 onwards using the function *RRBLUP2*.

Hybrid Pool Specific Additive Model

The Hybrid Pool Specific Additive model fits allele substitution effects for each heterotic pool, which are fitted simultaneously:

$$y = X\beta + Z_1\alpha_1 + Z_2\alpha_2 + e, \quad (3)$$

where y is a vector $(n \times 1)$ of hybrid's average performance; X is an incidence matrix $(n \times t)$ for fixed effects (tester-by-stage effects); Z_1 is a haplotype matrix $(n \times m)$ for inbred lines from heterotic pool 1, this matrix has entries 0 and 1 respectively for inbred's genotypes AA and aa; Z_2 is a haplotype matrix $(n \times m)$ for inbred lines from heterotic pool 2, this matrix has entries 0 and 1 respectively for inbred's genotypes AA and aa; β is a vector $(t \times 1)$ of fixed

effects; α_1 is a vector $(m \times 1)$ of allele substitution effects for heterotic pool 1; α_2 is a vector $(m \times 1)$ of allele substitution effects for heterotic pool 2; and e is the vector $(n \times 1)$ of residuals. Genomic predictions from the Hybrid Pool Specific Additive model were calculated using the AlphaSimR function *RRBLUP_GCA* in the year 1 and from the year 2 onwards using the function *RRBLUP_GCA2*.

Hybrid Additive + Dominance Model

The Hybrid Additive + Dominance model fits additive and dominance effects for both heterotic pools, which are fitted simultaneously:

$$y = X\beta + Z\mathrm{a} + W\mathrm{d} + e, \quad (4)$$

where y is a vector $(n \times 1)$ of hybrid's average performance; X is an incidence matrix $(n \times t)$ for fixed effects (tester-by-stage effects); Z is a genotype matrix $(n \times m)$ for hybrids, this matrix has entries 0, 1, and 2 respectively for hybrid's genotypes AA, Aa and aa; W is a heterozygous matrix $(n \times m)$, this matrix has entries 0, 1, and 0 respectively for hybrid's genotypes AA, Aa, and aa; β is a vector $(t \times 1)$ of fixed effects; a is a vector $(m \times 1)$ of additive effects; d is a vector $(m \times 1)$ of dominance effects; and e is a vector $(n \times 1)$ of residuals.

Considering the mixed model theory, the random effects a and d are assumed to have zero means. However, this assumption is not correct when the directional dominance is considered (Varona et al., 2018a). An alternative way to include the directional dominance in the Equation (4) was shown by Xiang et al. (2016) using:

$$y = X\beta + Z\mathrm{a} + W\mathrm{d}^* + W1\mu_d + e, \quad (5)$$

where $W1\mu_d$ is a vector $(n \times 1)$ of average dominance effects for each individual, which indicates the number of heterozygous loci for each individual by summing the rows of W (which has entries 0, 1, and 0 respectively for genotypes AA, Aa, and aa) multiplied by the average dominance effect μ_d. The deviation of dominance effects from the average (μ_d) is modeled by d^*, which is a vector $(m \times 1)$; and e is a vector $(n \times 1)$ of residuals. Genomic predictions from the Hybrid Additive + Dominance model were calculated using the AlphaSimR function *RRBLUP_D* in the year 1 and from the year 2 onwards using the function *RRBLUP_D2*.

Hybrid Pool Specific Additive Effects + Dominance Model

The Hybrid Pool Specific Additive Effects + Dominance model fits additive effects for each heterotic pool and dominance effects for both heterotic pools, which are fitted simultaneously:

$$y = X\beta + Z_1 a_1 + Z_2 a_2 + W d^* + W 1 \mu_d + e, \quad (6)$$

where y is a vector $(n \times 1)$ of hybrid's average performance; X is an incidence matrix $(n \times t)$ for fixed effects (tester-by-stage effects); Z_1 is a haplotype matrix $(n \times m)$ for inbred lines from heterotic pool 1, this matrix has entries 0 and 1 respectively for inbred's genotypes AA and aa; Z_2 is a haplotype matrix $(n \times m)$ for inbred lines from heterotic pool 2, this matrix has entries 0 and 1 respectively for inbred's genotypes AA and aa; W is a heterozygous matrix $(n \times m)$, this matrix has entries 0, 1, and 0 respectively for hybrid's genotypes AA, Aa, and aa; $W 1 \mu_d$ is a vector $(n \times 1)$ of fixed effects for the average heterozygosity for each hybrid and it contains the row-sums of W; β is a vector $(t \times 1)$ of fixed effects; a_1 is a vector $(m \times 1)$ of additive effects for heterotic pool 1; a_2 is a vector $(m \times 1)$ of additive effects for heterotic pool 2; d^* is a vector $(m \times 1)$ of dominance effects; and e is a vector $(n \times 1)$ of residuals. Genomic predictions from the Hybrid Pool Specific Additive Effects + Dominance model were calculated using the AlphaSimR function *RRBLUP_SCA* in the year 1 and from the year 2 onwards using the function *RRBLUP_SCA2*.

Training Population

The initial training population for genomic prediction considered the last four years of testcross 1 and 2 and, from the elite yield trial data from the recent burn-in phase. Separate training populations were developed for each of the heterotic pools for the Testcross model and the data compromised phenotypic records of the average performance of the inbred lines in the yield trials. The hybrid models (Hybrid Additive, Hybrid Pool Specific Additive, Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance) used a common training population and the data from the single-cross hybrids generated in the yield trials.

The initial training population for the testcross model considered phenotypic records on 17,760 inbred lines for each heterotic pool. The initial training population for the hybrid models compromised phenotypic records on 43,200 testcross genotypes. The training populations were

updated in subsequent years adding up new data from the yield trails. For the testcross model, every year 4,400 new phenotypes were added in the training population, separately for each heterotic pool. For the hybrid models, every year 10,800 new phenotypes were added in the training population, jointly for both heterotic pools.

2.5. General Combining Ability (GCA)

Genomic predictions of general combining ability were obtained in different ways, depending on the model. For the Testcross and Hybrid Pool Specific Additive models, GCA was predicted directly from allele substitution effects. But on the other hand, for the Hybrid Additive model, GCA was not truly predicted, and selection was based on predictions of breeding values across both heterotic pools. And for the Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance models, GCA was predicted indirectly using additive and dominance effects.

Genomic predictions of GCA from the Testcross, Hybrid Additive, and Hybrid Pool Specific Additive models were calculated as individual's genotypes multiplied by estimated allele substitution effects ($Z\hat{\alpha}$ for the Testcross model, $Z\hat{\tau}$ for the Hybrid model, and $Z\hat{\alpha}_1$ or $Z\hat{\alpha}_2$ for the Hybrid Pool Specific Additive model). Using these three models, we estimated "statistical" effects, hence these genomic predictions of GCA are confounded with a training population. For the Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance models, we first estimated "biological" additive and dominance effects ($\hat{a}$ and $\hat{d}$ for the Hybrid Additive + Dominance model or $\hat{a}_1$, $\hat{a}_2$ and $\hat{d}$ for the Hybrid Pool Specific Additive + Dominance model; VITEZICA; VARONA; LEGARRA, 2013), and then indirectly predicted GCA as individual's genotypes multiplied by estimated allele substitution effects ($Z\hat{\alpha}_1$ or $Z\hat{\alpha}_2$) assuming random mating between heterotic pools. We estimated allele substitution effects for these two models from allele frequencies in the other heterotic pool and "biological" additive and dominance effects as follows for the Hybrid Additive + Dominance model:

$$\hat{\alpha}_1 = \frac{1}{2}\left[\hat{a} + \hat{d}\left(q_2 - p_2\right)\right], \qquad (7)$$

$$\hat{\alpha}_2 = \frac{1}{2}\left[\hat{a} + \hat{d}\left(q_1 - p_1\right)\right],$$

where 1 denotes one heterotic pool and 2 denotes the other heterotic pool. And as follows for the Hybrid Pool Specific Additive + Dominance model:

$$\widehat{\alpha}_1 = \frac{1}{2}\left[\hat{a}_1 + \hat{d}\left(q_2 - p_2\right)\right], \quad (8)$$

$$\widehat{\alpha}_2 = \frac{1}{2}\left[\hat{a}_2 + \hat{d}\left(q_1 - p_1\right)\right].$$

2.6. Comparison of Breeding Programs

To evaluate the performance of the genomic prediction of GCA and their use in breeding, we tracked genetic gain and heterosis of all potential hybrids that could be formed at the DH stage. The aim here was to evaluate the power of the genomic predictions for early selection of inbred lines on their GCA. A breeding program would not be able to form all the possible hybrids so early in the breeding pipeline. Even in the context of a simulation study generating all possible hybrids so early in the breeding pipeline is computationally demanding (4,000 inbreds from each pool gives rise to 4,000 x 4,000 =16 million possible hybrids). We bypassed this issue by estimating the genetic gain and heterosis algebraically in addition to the simulated individuals of the breeding programme. We estimated the genetic gain as the mean genetic value of all possible hybrids in every year as follows (FALCONER; MACKAY, 1996):

$$M_{F_1} = \sum \mathrm{a}\left(p_1 - q_1 - y\right) + \mathrm{d}\left[2p_1q_1 + y\left(p_1 - q_1\right)\right] \quad (9)$$

where a and d are respectively additive and dominance effects; y is the difference of allele frequency between the two heterotic pools ($y = p_1 - p_2 = q_2 - q_1$); p_1 and q_1 are allele frequency in the heterotic pool 1 and; p_2 and q_2 are allele frequency in the heterotic pool 2. The heterosis of all possible hybrids in every year was estimated as follows (LAMKEY; EDWARDS, 1999):

$$H_{F_1} = \sum \left(2\bar{p}\,\bar{q}\right)\mathrm{d} + \left(1/2\,\mathrm{y}^2\right)\mathrm{d} \quad (10)$$

where $\bar{p} = \frac{\left(p_1 + p_2\right)}{2}$ and $\bar{q} = \frac{\left(q_1 + q_2\right)}{2}$; y^2 is the squared difference of allele frequency between the two heterotic pools; and d are dominance effects. Allele frequencies were calculated for every year.

The mean values of genetic gain and heterosis were centered at the mean values for hybrids in year 0 (last year of the recent breeding program in the burn-in phase), to aid in

visualization. The comparison between breeding programs were reported as ratios with 95% confidence intervals. These were calculated by performing paired Welch's test on log transformed values of the 100 simulation replicates. The log transformed differences and 95% confidence intervals from t-tests were then back-transformed to obtain the ratios (RAMSEY; SCHAFER, 2002).

3. RESULTS

Overall, the results show that QTL as expected provide higher accuracy than SNP markers, and that higher SNP maker density increases accuracy compared to lower density. The relative performance of genomic prediction models differs across marker scenarios. The results are structured in four main parts:

i. Effect of marker scenarios on genetic gain;
ii. Effect of marker scenarios on heterosis;
iii. Performance of genomic prediction models across markers scenarios; and
iv. Performance of heterosis across markers scenarios.

3.1. Effect of marker scenarios on genetic gain

SNP Markers

Higher SNP density scenario increases the genetic gain of the prediction models in a simulated maize breeding program. This is shown in Figure 4, which plots the hybrid genetic gain as the increase in mean genetic value under different marker scenarios over time for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The first graph shows the mean for each genomic prediction model evaluated in the future breeding program in the marker scenario with low density SNPs. The second graph shows the mean of genomic prediction models in the marker scenario with high density SNPs. The graphs show that increasing the marker density, all genomic prediction models presented an increasing in the genetic gain. On average, the high marker density generated 73.41% more genetic gain compared to the low-density markers.

QTL genotypes

QTL genotypes provide higher accuracy than SNPs markers in a hybrid breeding program simulation. This is shown in Figure 5, which plots the hybrid genetic gain as the increase in mean genetic value using the QTL genotypes overtime for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. Compared to the low-density marker scenario, the QTL genotypes scenario produced on average 100.46%

more genetic gain. Similarly, the QTL scenario was also superior to the high-density scenario and on average produced 15.57% more genetic gain.

Figure 4 – Hybrid genetic gain of different genomic prediction models over time of a simulated maize breeding program using two different marker densities. Hybrid genetic gain as the mean genetic value under two marker densities over time for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The lines in each plot represent the mean genetic value for the 100 simulated replicates and the shadings represent the 95% confidence interval.

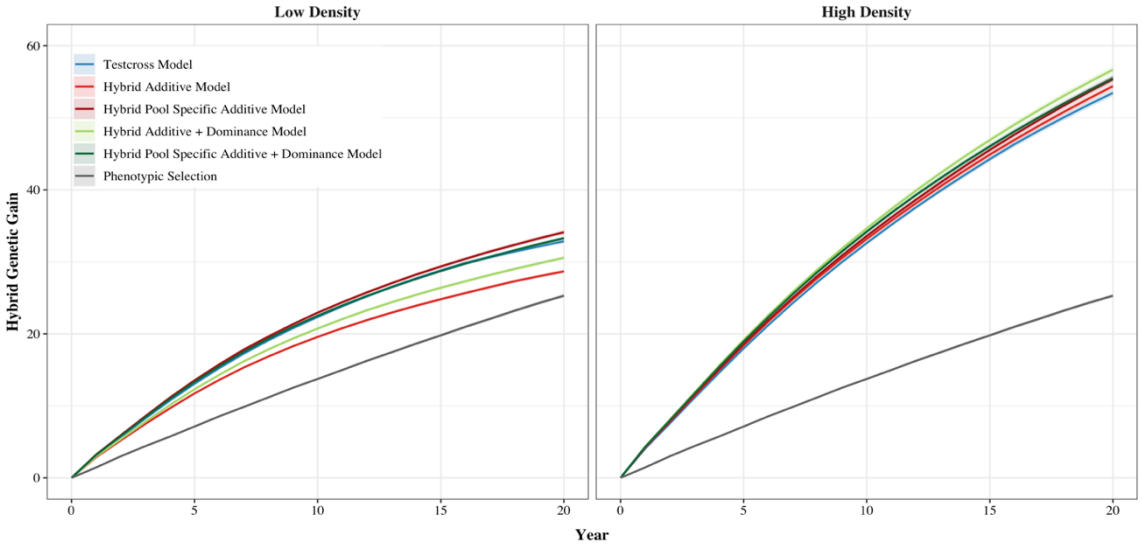
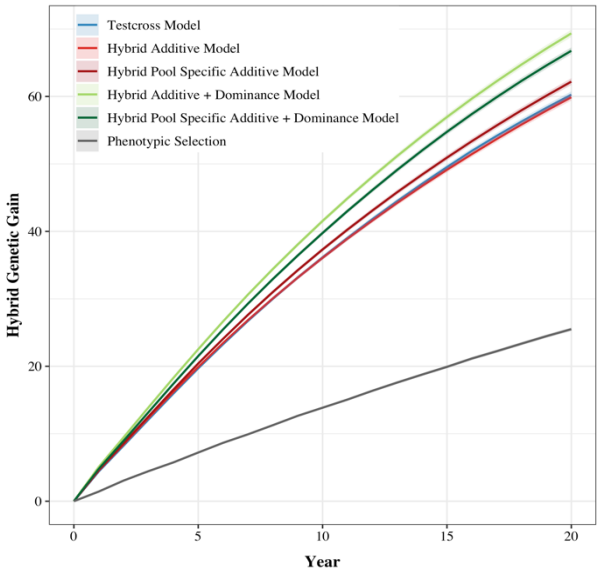


Figure 5 – Hybrid genetic gain of different genomic prediction models over time of a simulated maize breeding program using true QTL genotypes. Hybrid genetic gain as the mean genetic value using true QTL genotypes over time for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The lines in each plot represent the mean genetic value for the 100 simulated replicates and the shadings represent the 95% confidence interval.

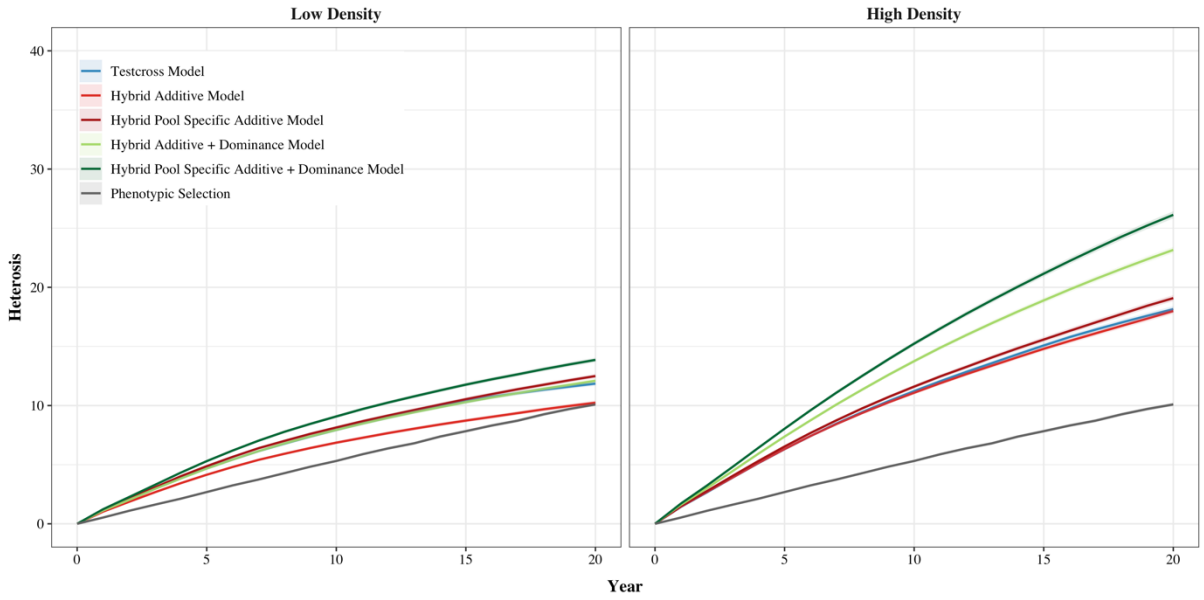


3.2. Effect of marker scenarios on heterosis

SNP Markers

Higher marker density increases the heterosis of the genomic prediction models in a simulated maize breeding program. This is shown in Figure 6, which plots the heterosis for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The first graph shows the mean heterosis for each genomic prediction model evaluated in the future breeding program under low-density coverage. The second graph shows the heterosis prediction models under high-density coverage. The figures show that increasing the marker density increased the heterosis for all genomic prediction models. On average, increasing the marker density produced 72.31% more heterosis for all genomic prediction models. Figure 6 also shows that the ranking of genomic prediction models was consistent across different maker densities. The rank of the models from the highest to lowest heterosis was: Hybrid Pool Specific Additive + Dominance, Hybrid Additive + Dominance, Hybrid Pool Specific Additive, Testcross and Hybrid Additive model. The only change in the raking was between Hybrid Pool Specific Additive and Hybrid Additive + Dominance models, but these models did not present statistical differences in the low-density scenario.

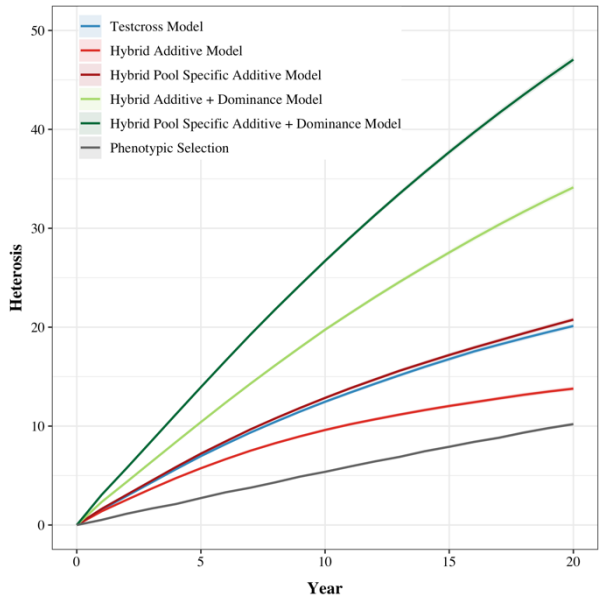
Figure 6 - Heterosis of different genomic prediction models over time of a simulated maize breeding program using two different marker densities. Heterosis means for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The lines in each plot represent the mean genetic value for the 100 simulated replicates and the shadings represent the 95% confidence interval.



QTL genotypes

QTL genotypes provide higher heterosis than SNPs markers scenarios in a hybrid breeding program simulation. This is shown in Figure 7, which plots the heterosis using the QTL genotypes overtime for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. A considerable difference was observed when the QTL scenario was compared to the low-density scenario. The QTL scenario yielded 118.45% more heterosis. A superiority of the QTL was also observed compared to the high-density scenario and on average yielded 24.75% more heterosis.

Figure 7 – Heterosis of different genomic prediction models over time of a simulated maize breeding program using true QTL genotypes. Heterosis means for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. The lines in each plot represent the mean genetic value for the 100 simulated replicates and the shadings represent the 95% confidence interval.



3.3. Performance of genomic prediction models across scenarios

SNP markers

The relative performance of the prediction models differs across marker scenarios. Figure 4 also shows that the performance of the genomic prediction models depends on the maker density. The first graph shows that under low-density, the genomic prediction models showed different performance for genetic gain. The second graph shows that under high-density the genomic prediction models have a similar performance for genetic gain.

Under low-density marker coverage, the genomic selection models showed a different performance for hybrid genetic gain. The models that fitted the average or additive effects specific for each heterotic pool (Testcross, Hybrid Pool Specific Additive and Hybrid Pool Specific Additive + Dominance models) showed higher genetic gain compared to the models that have common effects for both pools. These models produced a 12.82% more genetic gain than the models that modelled common average and additive effects for both heterotic pools (Hybrid Additive and Hybrid Additive + Dominance models).

In contrast, under high-density marker coverage, the genomic prediction models showed similar performance for genetic gain. The models the included dominance effects (Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance) and the Hybrid Pool Specific Additive showed higher genetic gain compared to the other models, but there was no statistical difference between some of models. The rank of the models showed that there is no correlation between the genomic prediction models and the marker densities.

QTL genotypes

The relative performance of the prediction models that fit dominance effects is superior to the models that fit average or additive effects only. This is shown in Figure 5, which plots the hybrid genetic gain as the mean genetic value using the QTL genotypes over time for the hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage. Modelling the dominance effects in the genomic prediction models produced on average 11.98% more genetic gain compared to the models that fitted only average or additive effects.

3.4. Performance of heterosis across scenarios

SNP Markers

The heterosis of genomic prediction models differs across marker scenarios. The first graph shows that models that fitted the average or additive effects for each pool specific for each heterotic pool showed higher heterosis compared to the models that modelled common effects for both heterotic pools in the low-density. Modelling the average or additive effects for each pool separately increased the heterosis in 22.04 and 14.60% for the Hybrid Additive compared to the Hybrid Pool Specific Additive model and Hybrid Additive + Dominance compared to the Hybrid Pool Specific Additive + Dominance model, respectively. The second

graph shows that models that fitted dominance effects presented higher heterosis compared to models that fitted only average or additive effects. With high-density markers the models that modelled dominance effects (Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance models) showed on average 33.87% higher heterosis compared to the other models (Testcross, Hybrid Additive and Hybrid Pool Specific Additive models).

QTL genotypes

Genomic prediction models that fitted dominance effects showed higher heterosis compared to the models that fitted average or additive effects only. This is shown in Figure 7, which plots the heterosis for hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage using the QTL genotypes. Models that included dominance effects produced on average 122.73% more heterosis compared to the other models.

4. DISCUSSION

The results study highlights four main topics for discussion:

i. The impact of marker scenarios in the genomic prediction;
ii. The relative performance of the genomic prediction models across marker scenarios;
iii. The impact of marker scenarios and genomic prediction models on heterosis;
iv. Implications in hybrid breeding programs; and
v. Limitations of stochastic simulations of hybrid breeding programs.

4.1. The impact of marker scenarios in the genomic prediction

Different marker scenarios were used to evaluate the impact of different marker densities on the genetic gain of prediction models. The marker scenarios used were low and high-density SNP markers and the QTL genotypes. As expected, the QTL genotypes outperform SNP markers and high-density markers outperform low-density markers.

SNP markers

Hybrid breeding programs using high-density markers produced more genetic gain for hybrids formed by crossing inbred lines from different heterotic pools at the double haploid stage, as expected. The increase in genetic gain is related to the selection accuracy. The expected accuracy is determined by the genetic variance, the number of phenotypic individuals, the effective number of chromosome segments, and the number of markers (DAETWYLER; VILLANUEVA; WOOLLIAMS, 2008; LEE et al., 2017). These authors explored the main factors that impact on the accuracy of genomic prediction. The number of observed phenotypes is an important factor that may limit the increase of accuracy, irrespective of the number of markers. Furthermore, accuracy may reduce due to the use of markers with an imperfect linkage disequilibrium with the QTL and the inclusion of markers with zero effects.

The results in the present study support this, showing that in breeding programs with low-density coverage, the markers may be in imperfect linkage disequilibrium with QTL. Mainly because with fewer markers, long-range linkage disequilibrium blocks are formed and there is less chance of the markers being in strong linkage disequilibrium with the QTL. Thus, the selection accuracy of the best parents is low and consequently, the genetic gain of the hybrids is also low. On the other hand, with high-density markers, the genome is saturated, and

it is ensured that the markers are in strong linkage disequilibrium with the QTL. Therefore, a great proportion of the QTL is captured by the markers. So, higher prediction accuracies ensure that the best individuals are selected based on their GEBV, and, consequently, the more genetic gain will be produced over time.

In this study, the ridge regression BLUP (RR-BLUP) models were fitted for the genomic prediction. RR-BLUP is a penalized regression model that imposes a penalty term, which is estimated as the ratio of the residual variance and the variance of marker effects. These models assume that all markers contribute equally to the genomic variance, with each marker effect following a normal distribution with a common variance (MEUWISSEN; HAYES; GODDARD, 2001). Although computationally efficient, these models may have some limitations depending on the genetic architecture of the trait. An alternative would be to use models that differentially weight markers. These models use variable selection and/or different prior (for example, LASSO (LI; SILLANPÄÄ, 2012; TIBSHIRANI, 1996)) or a mixture prior, such as BayesR (ERBE et al., 2012; MOLLANDIN; RAU; CROISEAU, 2021) may perform better in high-density scenarios. However, a further investigation would need to confirm this.

QTL genotypes

The hybrid breeding programs that used QTL genotypes showed higher genetic gain when compared to the breeding programs that used SNP markers. This was expected because the causal genotypes that, represent the QTL itself, were used by the prediction models and consequently, the prediction models captured all the QTL variance that drives the trait in the simulation. The genomic prediction models that used the QTL genotypes produced 15.57 % more genetic gain then the models that used high density markers.

The results obtained with the QTL genotypes also show that even when high-density markers were used, would not be possible to capture all the variance of the QTL that explain the trait. The main reason for that is the linkage disequilibrium between the markers and QTL. When the markers are used, they are considered the causal loci and to have a good estimation of the QTL variance these markers need to have a strong linkage disequilibrium with the QTL. Although high densities can be used, it is not possible to ensure that markers are in perfect linkage disequilibrium with all QTL.

4.2. The relative performance of the genomic prediction models across marker scenarios

In the simulations, we tried to mimic the evolutionary history of the two maize heterotic pools. The split between the heterotic pools was simulated to be 100 generations ago. So, it would be possible for the heterotic pools to have a difference in the linkage disequilibrium between markers and QTL and also differences in the allele frequency. The difference in the performance of the models can be attributed to the ability of prediction models to capture the genetic architecture of the trait and the difference in linkage disequilibrium pattern between the pools.

The low-density coverage was the scenario in which the genomic prediction models showed a substantial difference in the performance. When low-density coverage was used the markers are expected to be relatively more physically distant from QTL resulting in less opportunities for strong LD between markers and QTL. Consequently, the markers are not expected to estimate QTL effects as accurately. The hybrid additive model showed the worst performance relative to the genetic gain. The model, in this case, estimates the average effects based on the hybrid performance and not from the average performance of the inbred lines. This leads to a poor estimation of the average effects and hence a poor performance for genetic gain. Additionally, the hybrid additive model does not exploit the heterosis appropriately because it is driving the heterotic pools together. The hybrid additive and dominance model has the second-worst performance compared to others models. The dominance effects are more difficult to estimate than the average or additive effects. Specifically, under low-density the dominance effects are not accurately estimated and, this leads to poor performance. The heterosis performance of the model is similar to the better performance models, so it is clearly driving the heterotic pools apart. The other models, Testcross, Hybrid Pool Specific Additive, and Hybrid Pool Specific Additive + Dominance, are about equal for genetic gain. These models shared a common characteristic they modeled the average or the additive effects specific to each heterotic pool. The superiority of these models may be explained by the fact that these models may account for pool-specific LD patterns between QTL and markers for each heterotic pool. Regarding heterosis, the Hybrid Pool Specific Additive + Dominance model showed the best heterosis performance, and the other models being about equal. This shows that this model is acting differently than the other models.

On the other hand, in the high-density scenario, the genomic prediction models have similar performance for genetic gain. The performance of the prediction models has a modest

difference and was not statistically significant in most of the cases. Under high-density, there is a chance of more markers being in close physical proximity to QTL, providing more opportunity for markers to be in tight LD with the QTL. In this marker scenario, the single additive model (Hybrid Additive + Dominance model) has a slightly better performance than the other models. Probably this model has a better match with the genetic architecture of the trait simulated. The ranking of the models is likely to be sensitive to the genetic architecture (amount of dominance and divergence between heterotic pools) assumed in this study. Thus, an empirical validation would be necessary to choose between these models. A substantial difference between the models was observed for heterosis. The results show that although important for hybrid breeding programs, heterosis alone is not the only way to improve the performance of the hybrids. As expected, explicitly modeling dominance resulted in more heterosis. Furthermore, having separate additive effects increased heterosis the most. This is likely due to the ability of this model to account for pool-specific LD patterns.

Lastly, in the QTL genotypes scenario, models that included the dominance effects yielded more genetic gain compared to the models with only the average or additive effects. In this scenario, all pairwise comparisons between the prediction models were statistically different, except for the comparison between the Testcross and the Hybrid Additive models. The superiority of the Hybrid Additive + Dominance and Hybrid Pool Specific Additive + Dominance models was expected once these models match the genetic architecture of the trait, that includes additive and dominance effects. In contrast, the models that fitted only the average or additive effects are not able to capture only the dominance of the QTL and, hence fail to explain all the genetic variance of the trait. Another important thing to highlight is that the ranking for the QTL genotypes did not match the rankings for marker high-density scenario. There is the false idea that SNP markers, usually under high densities, will outcome the same results obtained from QTL. The SNP model depends on linkage disequilibrium as a surrogate for the QTL. Usually, not all markers are in tight linkage disequilibrium with all causal loci and, hence fail to explain all the genetic variance of the trait.

4.3. The impact of marker scenarios and genomic prediction models on heterosis

We also examined how the prediction models would impact on the heterosis of all possible hybrids formed by crossing inbred lines from different heterotic pools. In other words, to assess how the selection of the parents based on GCA affects the heterosis using different prediction models and how breeders can explore the heterosis in breeding programs. Heterosis

is present only when dominance is present and also if there is difference in allele frequency between the populations or inbreds (FALCONER; MACKAY, 1996).

In the low-density scenario, the prediction models resulted in different levels of heterosis at end of 20 years of breeding. In this scenario, probably the markers were not in good LD with the QTL that drives the trait and, hence the QTL effects were captured less accurately. Moreover, the dominance effects are one of the main effects that guide the heterosis are more difficult to be estimated, especially under low-density, compared to the average or additive effects. Another factor to be considered is the fact of the prediction models that fitted only average or additive effects did not estimate appropriately the dominance effects and hence assigned incorrectly the QTL with dominance effects. So, the heterosis at the end of breeding cycle for these models is lower. The Hybrid Pool Specific Additive + Dominance was the model that showed more gain in heterosis and, it was statistically different from the other models. The superiority of this model can be explained by the fact that it estimated better the effects of the QTL with dominance effects and also captured the difference in allele frequency between the heterotic pools.

The Hybrid Additive + Dominance model showed similar performance to the models that fitted only the average or additive effects; although it was expected it would yield more heterosis because the dominance effects are included in the prediction model. In this model, the additive effects of both heterotic pools were considered together and, the difference in the allele frequency was the mean of allele frequency of both pools. So, the dominance effects, that are the same for both heterotic pools, decrease because difference of allele frequency between the pools (^{y2}) is small and hence the heterosis was the same of the models that did not fit the dominance effects.

In the high-density scenario, the prediction models that fitted dominance effects showed more heterosis than the other prediction models. In general, the increase in the heterosis is related to the number of markers in linkage disequilibrium with QTL. Moreover, the higher increase of the models that fitted the dominance effects is related to the ability of these models the captured the QTL with dominance effects, what do not happen with the models that not fit the dominance effects.

In the QTL genotypes scenario, as was showed in the high-density, the prediction models that fitted dominance effects showed more heterosis than the other prediction models. With the QTL genotypes, we fitted the prediction models using the causal genotypes of the trait and show how the model perform.

4.4. Implications in hybrid breeding programs

The present study showed that there are some factors that can impact on genomic prediction models for general combining ability in early stages of hybrid breeding programs. These study highlights how the marker density, the genomic prediction models performance and heterosis can implicate in breeding programs.

We showed how the marker density affects the genetic gain in a hybrid breeding program. The increasing the marker density implicated in a direct increase of genetic gain for all prediction models. Although, there was an increase in the genetic gain, apparently the genetic gain reached a plateau with the high-density marker coverage. There are many factors that affect genomic prediction, such as marker density, training population size, relatedness between training and selection populations (DE LOS CAMPOS et al., 2013; HICKEY et al., 2014; LIU et al., 2018; ZHAO; METTE; REIF, 2015). In the present study the hybrid breeding programs in which high density markers were used, produced 72.31% more genetic gain then the breeding programs that used low density markers, but no significant increase in genetic gain was observed when the number of markers increased from 10,00 to 20,000 (results not showed). This is consistent with previous studies that verify the impact of marker density and type and size of the training population (HICKEY et al., 2014; LIU et al., 2018). The most important fact to be considered with genomic selection is to use a number of markers that ensure a good LD between marker and QTL. So, the marker can explain a good proportion of the QTL variance.

The final aim of hybrid breeding programs is to produce a hybrid that had a superior performance compared to the hybrids available in the market. To achieve this is necessary to develop and select good inbred lines and, identify the best combinations that produce the hybrids with superior performance (SHULL, 1909). In this study, we showed that the performance of the prediction models that estimated GCA depended on the marker density. The prediction models showed a different performance under low-density coverage, but under high-density coverage, the performance was similar between the models. The models that showed a better performance in both low- and high-density coverage were the models that fitted the average and additive effects specific to each heterotic pool and/or fitted the dominance effects. Therefore, these models would provide a better fit compared to the simpler models.

Another factor important in hybrid breeding programs is heterosis, because it ensures the development of hybrid cultivars. The concept of heterosis was coined by SHULL (1908) and is attributed to the superior performance of the hybrid compared to their parental inbred lines. The high performance of a hybrid is a result from a large amount of heterosis between two parents that have modest per se performance or a modest amount of heterosis between two parents that have higher per se performance (BERNARDO, 2020). In this study, we assessed the effects of the prediction models on the heterosis of the hybrids to understand what was associated with the improved performance of the hybrids. The results showed that hybrid genetic gain was increasing over the breeding cycles. One of the reasons was the increase in the performance per se of the inbred lines. Apart from this, the improvement in the performance of the hybrids was due to heterosis, mainly for the prediction models that fitted the dominance effects. These models that fit the dominance effects have an advantage because they can capture the dominance effects of the QTL.

4.5. Limitations of stochastic simulations of hybrid breeding programs

Breeding programs are very complex, and many factors can affect them in different ways. The simulations performed in the present study did not model all the complexity existent in a real hybrid breeding program. The limitations and impact of these assumptions used on the simulated hybrid breeding programs are discussed in the following.

The genetic architecture of the trait used in the simulation was chosen to match long-term rates of genetic gain observed for inbred mid-parent and hybrid yield observed in real data (TROYER; WELLIN, 2009). The trait simulated in this study was controlled by 3,000 QTL, with a dominance degree of 0.92 and dominance variance of 0.2 and, the split of heterotic pools that occurred 100 generations ago. The results showed in this study are limited to the genetic architecture of the trait, in this case, yield, used in the simulations. The results may give an idea of how genomic prediction models perform in different marker scenarios. It is important to highlight that a trait with different genetic architecture may show different results. Although, no major changes are expected in the performance of the genomic prediction models.

The genomic selection accuracy directly affects the genetic gain in simulated breeding programs. In the present study the genomic selection accuracies observed were higher than those observed in real breeding programs conditions. As indicated by GAYNOR et al. (2017) this happens because in the simulations the molecular markers do not have genotyping errors,

the genetic control of the trait does not involve epistasis and, the simulation used a closed breeding program. As a consequence, the simulations favored high genomic selection accuracy and hence the genetic gain. These effects should affect all hybrid crop breeding programs using genomic selection equally, suggesting that the difference observed between the prediction models is exclusively due to the difference in the prediction models.

5. CONCLUSIONS

We evaluated the performance of genomic prediction models to estimate GCA at the early stages of hybrid breeding programs. The genomic prediction models delivered higher genetic gain and higher heterosis under high-density coverage. The prediction models showed a difference in the performance under low-density coverage. Although, under high-density, the performance between the models was similar. Accordingly, three main conclusions can be drawn from these results:

i) increasing the marker density increases the genetic gain;
ii) the relative performance of the prediction models depends on the marker density and;
iii) there is a significant difference in the genetic gain when markers and true QTL genotypes are used.

Our results suggest that more complex models, such as models that fit effects specific to heterotic pools and dominance effects, can be more beneficial to estimate GCA in hybrid breeding programs.

REFERENCES

ALBRECHT, T. et al. Genome-based prediction of testcross values in maize. **Theoretical and Applied Genetics**, v. 123, n. 2, p. 339–350, 1 jul. 2011.

ALBRECHT, T. et al. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. **Theoretical and Applied Genetics**, v. 127, n. 6, p. 1375–1386, 1 jun. 2014.

BERNARDO, R. N. **Breeding for quantitative traits in plants**. 3rd ed ed. Woodbury, Minn: Stemma Press, 2020.

BERNARDO, R.; YU, J. Prospects for Genomewide Selection for Quantitative Traits in Maize. **Crop Science**, v. 47, n. 3, p. 1082–1090, maio 2007.

CHEN, G. K.; MARJORAM, P.; WALL, J. D. Fast and flexible simulation of DNA sequence data. **Genome Research**, v. 19, n. 1, p. 136–142, 2009.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. **PLoS ONE**, v. 3, n. 10, p. e3395, 14 out. 2008.

DE LOS CAMPOS, G. et al. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, v. 193, n. 2, p. 327–345, fev. 2013.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 1–22, set. 1977.

EATHINGTON, S. R. et al. Molecular Markers in a Commercial Breeding Program. **Crop Science**, v. 47, p. S-154-S-163, dez. 2007.

ERBE, M. et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. **Journal of Dairy Science**, v. 95, n. 7, p. 4114–4129, jul. 2012.

FALCONER, D. S. A note on Fisher's 'average effect' and 'average excess'. **Genetical Research**, v. 46, n. 3, p. 337–347, dez. 1985.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. Harlow, UK: Longman, 1996.

FALCONER, D. S.; MACKAY, T. F.; FRANKHAM, R. Introduction to Quantitative Genetics (4th edn). **Trends in Genetics**, v. 12, n. 7, p. 280, 1996.

FISHER, R. A. AVERAGE EXCESS AND AVERAGE EFFECT OF A GENE SUBSTITUTION. **Annals of Eugenics**, v. 11, n. 1, p. 53–63, jan. 1941.

GAYNOR, R. C. et al. A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. **Crop Science**, v. 57, n. 5, p. 2372–2386, set. 2017.

GAYNOR, R. C.; GORJANC, G.; HICKEY, J. M. **AlphaSimR: An R-package for Breeding Program Simulations**. [s.l.] Genetics, 11 ago. 2020. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2020.08.10.245167>. Acesso em: 9 nov. 2020.

GEIGER, H. H.; GORDILLO, G. Doubled Haploids in Hybrid maize breeding. **Maydica**, v. 54, p. 485–499, 2009.

HALLAUER, A. R.; CARENA, M. J.; MIRANDA FILHO, J. B. **Quantitative genetics in maize breeding**. 3rd ed. ed. New York ; London: Springer, 2010.

HEFFNER, E. L. et al. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. **Crop Science**, v. 50, n. 5, p. 1681–1690, set. 2010.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J.-L. Genomic Selection for Crop Improvement. **Crop Science**, v. 49, n. 1, p. 1–12, jan. 2009.

HICKEY, J. M. et al. Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. **Crop Science**, v. 54, n. 4, p. 1476–1488, jul. 2014.

HILL, W. G.; GODDARD, M. E.; VISSCHER, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. **PLoS Genetics**, v. 4, n. 2, p. e1000008, 29 fev. 2008.

HIVERT, V.; WRAY, N. R.; VISSCHER, P. M. Gene action, genetic variation, and GWAS: A user-friendly web tool. **PLOS Genetics**, v. 17, n. 5, p. e1009548, 20 maio 2021.

KADAM, D. C. et al. Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. **G3: Genes|Genomes|Genetics**, v. 6, n. 11, p. 3443–3453, nov. 2016.

KIMURA, M.; CROW, J. F. On the maximum avoidance of inbreeding. **Genetics Research**, v. 4, n. 3, p. 399–415, 1963.

LAMKEY, K. R.; EDWARDS, J. W. Quantitative Genetics of Heterosis. In: **The Genetics and Exploitation of Heterosis in Crops**. Proceedings of the International Symposium on the Genetics and Exploitation of Heterosis in Crops. Mexico City, Mexico: [s.n.]. p. 31–48.

LEE, S. H. et al. Using information of relatives in genomic prediction to apply effective stratified medicine. **Scientific Reports**, v. 7, n. 1, p. 42091, fev. 2017.

LI, Z.; SILLANPÄÄ, M. J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and Applied Genetics**, v. 125, n. 3, p. 419–435, ago. 2012.

LIU, X. et al. Factors affecting genomic selection revealed by empirical evidence in maize. **The Crop Journal**, v. 6, n. 4, p. 341–352, ago. 2018.

LYNCH, M.; WALSH, B. Genetics and analysis of quantitative traits. 1998.

MASSMAN, J. M. et al. Genomewide predictions from maize single-cross data. **Theoretical and Applied Genetics**, v. 126, n. 1, p. 13–22, jan. 2013.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, abr. 2001.

MOLLANDIN, F.; RAU, A.; CROISEAU, P. An evaluation of the predictive performance and mapping power of the BayesR model for genomic prediction. **G3 Genes|Genomes|Genetics**, p. jkab225, 7 jul. 2021.

POWELL, O. et al. **A Two-Part Strategy using Genomic Selection in Hybrid Crop Breeding Programs**. [s.l.] Genetics, 25 maio 2020. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2020.05.24.113258>. Acesso em: 4 jun. 2020.

PRICE, G. R. Fisher's "fundamental theorem" made clear. **Annals of Human Genetics**, v. 36, n. 2, p. 129–140, nov. 1972.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2019.

RAMSEY, F. L.; SCHAFER, D. W. **The statistical sleuth: a course in methods of data analysis**. 2nd ed ed. Australia ; Pacific Grove, CA: Duxbury/Thomson Learning, 2002.

SHULL, G. H. The Composition of a Field of Maize. **Journal of Heredity**, v. os-4, n. 1, p. 296–301, 1 jan. 1908.

SHULL, G. H. A Pure-Line Method in Corn Breeding. **Journal of Heredity**, v. os-5, n. 1, p. 51–58, 1 jan. 1909.

SU, G. et al. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS ONE**, v. 7, n. 9, p. e45293, 13 set. 2012.

TECHNOW, F. et al. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. **Theoretical and Applied Genetics**, v. 125, n. 6, p. 1181–1194, out. 2012.

TECHNOW, F. et al. Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. **Genetics**, v. 197, n. 4, p. 1343–1355, ago. 2014.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

TORO, M. A.; VARONA, L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, n. 1, p. 33, dez. 2010.

TROYER, A. F.; WELLIN, E. J. Heterosis Decreasing in Hybrids: Yield Test Inbreds. **Crop Science**, v. 49, n. 6, p. 1969–1976, nov. 2009.

VARONA, L. et al. Non-additive Effects in Genomic Selection. **Frontiers in Genetics**, v. 9, p. 78, 6 mar. 2018.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. **Genetics**, v. 195, n. 4, p. 1223–1230, dez. 2013.

WELLMANN, R.; BENNEWITZ, J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics Research**, v. 94, n. 1, p. 21–37, fev. 2012.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetical Research**, v. 75, n. 2, p. 249–252, abr. 2000.

WRIGHT, S. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. **Genetics**, v. 6, n. 2, p. 124, 1921.

ZHANG, X. et al. EMMA: An Efficient Massive Mapping Algorithm Using Improved Approximate Mapping Filtering. **Acta Biochimica et Biophysica Sinica**, v. 38, n. 12, p. 857–864, dez. 2006.

ZHAO, Y.; METTE, M. F.; REIF, J. C. Genomic selection in hybrid breeding. **Plant Breeding**, v. 134, n. 1, p. 1–10, fev. 2015.

ARTIGO 2 – MULTIVARIATE GWAS FOR THE RESISTANCE TO DIPLODIA AND FUSARIUM EAR ROT IN MAIZE

ABSTRACT

Maize is one of the most important cultivated crops and, suffers attack by several diseases. Ear rots, caused by the fungi *Stenocarpella maydis* and *Fusarium verticillioides*, are the most limiting factor in maize production, reducing yield and grain quality. Additionally, some fungi produce secondary metabolites, as known as mycotoxins, that are harmful to humans and animals. Traditionally, multiple traits are collected on the same individuals. Therefore, multivariate methods have been suggested to exploit the increasing quantity and complexity of the traits. The objectives of this study were to identify loci controlling the resistance to ear rot caused by *Stenocarpella maydis* and caused by *Fusarium verticillioides* and to identify markers linked to causal loci that contribute to the resistance to both Diplodia and Fusarium ear rot. We evaluated the univariate and multivariate approaches using a maize diverse panel for three different traits, percentage of rotten kernels, ear rot incidence score, and yield. Our analyses revealed two markers associated with the traits in the univariate analyses. In the multivariate analyses, two markers were associated with the traits, with one marker in common to the univariate analysis. The markers associated with the traits in both analyses were located inside or close to candidate genes. The candidate genes were related to biological functions that may be associated to the ear rot resistance. The resistance to ear rots is complex and most of the genetic and biochemical pathways of its resistance to the ear rots are still unknown. The results obtained by the multivariate were encouraging once multiple traits are available, but further studies are necessary.

Keywords: Ear rots; Association mapping; *Zea mays* L.

1. INTRODUCTION

Maize is one of the most cultivated crops in the world and suffers the attack of numerous pathogens, which can infect all parts of the plant. Among the diseases that infect maize plants, the ear rots are the most limiting factor in maize production, mainly because of the yield losses and reduction of grain quality, and some of these fungi produce mycotoxins (LANUBILE et al., 2017).

Fusarium ear rot (FER) is caused by several *Fusarium* species, but the most common causal agent is *Fusarium verticillioides* (previously known as *Fusarium moniliforme*) (MESTERHAZY; LEMMENS; REID, 2012; MUNKVOLD; WHITE, 2016). The causal agent overwinters on corn residues or other plants that serve as hosts. The infection occurs mainly when the fungal spores germinate on the silks and grow in the silk channel, consequently infecting the kernels. Moreover, the infection occurs through wounds caused by hail or insects and systemic infection of the ear through infected stalks that generate infected seeds(MUNKVOLD; MCGEE; CARLTON, 1997). The optimal environmental conditions for the disease development are dry and warm weather, which can cause severe infections (REID et al., 1999). The most common FER symptoms are the development of a white-pinkish mold on the ear and the starburst pattern, which is a light-colored streak radiating from the top of kernels (LANUBILE et al., 2017; MUNKVOLD; WHITE, 2016). In addition to the reduction of yield and grain quality, *Fusarium verticillioides* produces mycotoxins, known as fumonisins, which can be carcinogenic and are associated with several diseases in animals (LOGRIECO et al., 2002). Many countries have established regulations for fumonisins concentrations for livestock and human consumption (LANUBILE et al., 2017).

Diplodia ear rot (DER) is caused by *Stenocarpella maydis* (previously known as *Diplodia maydis*) (MUNKVOLD; WHITE, 2016). Maize is the only known host and the disease severity depends on the infected and unburied corn debris (MUNKVOLD; WHITE, 2016; PINTO et al., 2003). The optimal environmental conditions for disease development are wet weather and moderate temperatures. The DER symptoms are mainly tan spots on husks or decolored husks. On the ear, there is the development of a white mycelial infection from the base of the ear that progresses to the top (MUNKVOLD; WHITE, 2016; ROSSOUW et al., 2009). As FER, the main impact on the crop is due to the reduction of grain quality and yield. The mycotoxin produced by *S. maydis* is known as diplodiatoxin, which presents a risk to domestic animals such as cattle, sheep, and poultry. This mycotoxin is the cause of diplodiosis

and has been related to neurological disorders in cattle in South Africa and Argentina (KELLERMAN et al., 1991; ODRIOZOLA et al., 2005; WICKLOW et al., 2011).

Studies have been carried out to understand the genetic control of resistance to ear rots and mycotoxins. Many of them showed that it is highly polygenic and complex trait (KING; SCOTT, 1981; NANKAM; PATAKY, 1996; PÉREZ BRITO et al., 2001; ROSSOUW; VAN RENSBURG; VAN DEVENTER, 2002a, 2002b; VAN RENSBURG; FERREIRA, 1997). Studies for FER and fumonisin content found many minor QTL and markers associated with or fumonisins (CHEN et al., 2012, 2016; DE JONG et al., 2018; DING et al., 2008; MASCHIETTO et al., 2017; PÉREZ BRITO et al., 2001; ROBERTSON-HOYT et al., 2006; SAMAYOA et al., 2019; ZILA et al., 2013, 2014). Most of the QTL and markers explained a small portion of the phenotypic variance, suggesting that many genes with small effects underline the inheritance of FER. Moreover, the disease is mainly affected by environmental conditions. Despite the importance of DER, only a few studies of QTL mapping were performed, and some QTL were identified for resistance to ear rot (GUTIÉRREZ, 2008; ROMERO LUNA, 2012; ROSSOUW et al., 2009). No Genome-Wide Association Study (GWAS) to understand the resistance to DER has been reported so far.

Traditionally GWAS are performed using single traits or multiple traits independently. Although, in practice, multiple traits are collected on the same individuals, most of the analyses are performed on single traits. To exploit the increasing quantity and complexity of the traits, multivariate methods have been suggested and developed to analyze multiple traits. Multivariate procedures have some advantages when compared to univariate methods. Some studies have shown that multivariate analysis increased the statistical power and precision of the parameter estimation of the analysis, considering the correlation between traits (JIANG; ZENG, 1995; KORTE et al., 2012; LIU et al., 2009; ZHOU; STEPHENS, 2014). Another advantage is that multivariate analysis performs a single test for association for a set of traits. Therefore, there is a reduction in the number of performed tests and eases the multiple testing burden (GALESLOOT et al., 2014). Lastly, multivariate analysis can be used to detect the presence of pleiotropy when a causal region is associated with multiple traits (FERNANDES et al., 2021; RICE; FERNANDES; LIPKA, 2020).

Considering the importance of ear rots for maize, the objectives of this study were: i) to identify loci controlling the resistance to ear rot caused by *Stenocarpella maydis*; ii) to identify loci controlling the resistance to ear rot caused by *Fusarium verticillioides*; and iii) to identify

markers linked causal loci that contribute to the resistance to both Diplodia and Fusarium ear rot.

2. MATERIALS AND METHODS

2.1. Plant material and field experiments

A maize diverse panel of 228 inbred lines that represent the diversity of maize germplasm around the world, were evaluated in the crop season 2012/2013 under inoculation with two ear rot fungi, *Fusarium verticillioides* and *Stenocarpella maydis*, in two locations. The maize inbred lines represent a sample of the diverse panel previously described by CANTELMO; VON PINHO; BALESTRE (2017).

The field experiments were conducted in two locations in the Minas Gerais State, Brazil. One experiment was conducted in Lavras (910 m above sea level, 21˚ 14'S 45˚ 00" W), which has a humid subtropical climate (Cfa), and the other experiment was conducted in Uberlândia (863 m above sea level, 18˚ 55'S 48˚ 16'W), which has a tropical savanna climate (Aw), according to Köppen-Geiger classification, respectively.

The inbred lines were evaluated in two separate experiments in each location. In one experiment, the inbred lines were inoculated with an inoculum of *F. verticilliodes* and, in the second experiment, the inbred lines were inoculated with an inoculum of *S. maydis*. Each experiment was evaluated in a complete random block design with three replicates. The experimental plots consisted of a row with a length of 3 meters and a width of 0.7 meters between rows.

The isolates of *F. verticillioides* and *S. maydis* were obtained from rotten ears collected in the experimental fields in Lavras and Uberlândia and replicated at the Seed Pathology Laboratory at the Federal University of Lavras (UFLA). The isolates of *S. maydis* were grown in a complete medium and stored in glass tubes for 30 days. The isolates of *F. verticillioides* were grown in a complete medium over a period of seven days before inoculation. The conidial suspension for both fungi was adjusted to 1.0×10^6conidia.mL^{-1}, by counting spores in a Neubauer chamber. The inoculation was performed 15 days after 100% of the plants had developed styles-stigmata. Using the methodology proposed by CLEMENTS et al. (2003), 1 mL of the conidial suspension was inoculated into each ear.

The incidence of ear rot was evaluated using two different methods (dos Santos et al., 2016):

i) percentage of rotten kernels (PRK); and

ii) ear rot incidence score (SCO).

The percentage of rotten kernels was measured according to the procedure proposed in decree no. 11 of 12/96 (BRASIL, 1996). For this evaluation, a sample of 230 grams of kernels per plot was collected and determined the percentage of kernels showing discoloration in more than 25% of the total kernel surface. For the ear rot incidence score, a diagrammatic scale proposed by REID et al. (2002) was used. The evaluation was based on 7-class rating scale where 1 = 0%, 2 = 1–3%, 3 = 4–10%, 4 = 11–25%, 5 = 26–50%, 6 = 51–75%, and 7 = >75% of the kernels exhibiting visible symptoms of infection such as rot and mycelial growth. In addition, the yield (YLD) of each inbred line. was evaluated based on the husked ear weight in $kg.plot^{-1}$, which was converted to $kg.ha^{-1}$.

2.2. Genotypic data

The maize inbred lines were genotyped with the high-throughput DArTseq technology. This technology is based on the PstI-based complexity reduction method (WENZL et al., 2004). All the procedures to obtain the genomic representations were described by KILIAN et al. (2012). Markers were scored “1” for presence and “0” for absence and “-” for failure to score. A total of 23,153 polymorphic silicoDArT markers were generated.

The genotypic data quality control was tested for call rate (%), minor allele frequency (MAF) with the *dartR* package GRUBER et al. (2018) in the R software (R CORE TEAM, 2020). Markers with a call rate below 90% and MAF lower than 0.05 were filtered out from the genotypic dataset. Additionally, all the monomorphic markers were removed. After the quality control, 12,936 silicoDArT were selected for the study.

2.3. Statistical analysis

2.3.1. Adjusted means and heritabilities

The ear rot measurement traits, percentage and score, and yield were transformed using a Box-Cox transformation (*boxcox* function in MASS package (VENABLES; RIPLEY, 2002)). The lambda (“optimal value”) was calculated for each trait and the one that resulted in the best approximation of a normal distribution was used in the transformation.

Adjusted means were estimated for each inbred line by fitting the linear mixed model:

$$y_{ijklm} = \mu + g_i + d_j + l_k + gd_{ij} + gl_{ik} + dl_{jk} + b(rl)_{klm} + e_{ijklm} \quad (1)$$

where y_{ijklm} is the observed trait value; μ is the intercept; g_i is the fixed effect of genotype *i*; d_j is the fixed effect of disease *j;* l_k is the fixed effect of location *k;* gd_{ij} is the fixed effects of interaction between genotype *i* and disease *j*; gl_{ik} is the fixed effect of the interaction between genotype *i* and location *k*; dl_{jk} is the fixed effects between disease *j* and location *k*; gdl_{ijk} is the fixed effect of the interaction between genotype *i*, disease *j* and location *k*; $b(rl)_{klm}$ is the random effect of block *m* nested with replication *l* and location *k;* e_{ijklm} is the error. Unstructured covariance structure was used, allowing to estimate the adjusted means for each inbred line separately for each ear rot disease.

To estimate the heritability of the inbred lines for each ear rot disease, a similar model was used but all the effects were considered as random. The heritability was calculated using the following model:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \left(\frac{\sigma_{gl}^2}{l}\right) + \left(\frac{\sigma_e^2}{rl}\right)\right)} \quad (2)$$

where σ_g^2 refers to the genetic variance, where σ_{gl}^2 refers to the interaction between genotype and location variance, σ_e^2 refers to the residual error variance and r and l refer to the number of replications and environments, respectively.

2.3.2. Relationship matrix and Population structure

The genomic relationship matrix was estimated using the additive relationship matrix (A) from markers using the function A.mat function in rrBLUP package (ENDELMAN, 2011). The relationship matrix was estimated as proposed by (VANRADEN, 2008):

$$A = \frac{WW'}{2\sum pq} \quad (3)$$

where W is the centered matrix on $2p$ (average of the favorable alleles for a given locus); p is the frequency of the favorable allele; q is the frequency of the unfavorable allele; and $2\sum pq$ is the sum of the locus variances.

To detect and correct for population structure in GWAS (PRICE et al., 2006), a principal component analysis (PCA) was performed using the function snpgdsPCA in SNPRelate package (ZHENG et al., 2012). To account for the number of subgroups in the population, initially, a visual inspection was done plotting the first component versus the second component. Another procedure based on BIC model selection was used to determine how many

principal components (PCs) were necessary to control for population structure using the software GAPIT (LIPKA et al., 2012).

2.3.3. Genetic Correlation

The genetic, phenotypic, genotype by location interaction and residual correlations were estimated by fitting a multivariate version of the model used to estimate heritability. The genotype, genotype by location interaction, residual and block nested within replication effects were fit as random effects with separate variance for each disease ear rot and with covariance between effects on each disease. The other effects, location, disease, and location by disease interaction, were fitted as fixed effects. All mixed model analyses were performed with ASReml-R version 4 (BUTLER et al., 2017).

The genetic correlations were estimated using two different methods. For traits measured on the same disease experiment (DER or FER), genetic correlations were calculated using the following model:

$$r_g = \frac{\sigma_{g(xy)}}{\sqrt{\left[\sigma^2_{g(x)} \times \sigma^2_{g(y)}\right]}} \tag{4}$$

where $\sigma^2_{g(x)}$ and $\sigma^2_{g(y)}$ are the genetic variance for traits x and y, respectively; and $\sigma_{g(xy)}$ is the covariance between traits x and y.

For traits measured on different disease experiments, genetic correlations were calculated according to the model proposed by BURDON (1977):

$$r^*_g = \frac{r_p}{\sqrt{\left[h^2_x \times h^2_y\right]}} \tag{5}$$

where r_p is the phenotypic correlation between traits across disease experiments; and h^2_x and h^2_y are the heritabilities of each trait within each disease experiment.

2.4. Univariate and Multivariate GWAS

Univariate and multivariate GWAS analyses were conducted using the software GEMMA (ZHOU; STEPHENS, 2012, 2014). The univariate GWAS model for each ear rot measurement trait (PERC and SCO) and yield individually, using the following linear mixed model:

$$y = \mathrm{W}\alpha + x\beta + \mathrm{Z}u + \varepsilon \tag{6}$$

where y is a $n \times 1$ vector of traits for n individuals; W is a $n \times c$ matrix of fixed effects including a column vector of 1s; α is a $c \times 1$ vector of corresponding coefficients including the intercept; x is a $n \times 1$ vector of marker genotypes; β is the effect size of the marker; Z is an $n \times m$ loading matrix; u is an $m \times 1$ vector of random effects; $u \sim MVN_m(0, \lambda\tau^{-1}K)$, where τ^{-1} is the variance of residual errors; λ is the ratio between the two variance components; K is a m$\times$ m kinship matrix; ε is an $n \times 1$ vector of errors; $\epsilon \sim MVN_n(0, \tau^{-1}I_n)$, where I_n is an $n \times n$ identity matrix and MVN denotes multivariate normal distribution.

The multivariate GWAS was fitted using the following linear mixed model:

$$\tilde{Y} = \mathrm{A}\tilde{W} + \beta\tilde{\mathrm{x}}^{T} + \tilde{G} + \tilde{E} \tag{7}$$

where $\tilde{Y}$ is a $d \times n$ matrix of d phenotypes for n individuals; $\tilde{W}$ is a $c \times n$ matrix of fixed effects including a column of 1s; A is $d \times c$ matrix of the corresponding coefficients; $\tilde{x}^T$ is a $n \times 1$ vector of genotype for a particular marker; β is a $d \times 1$ vector of its effect sizes for the d phenotypes; $\tilde{G}$ is a $d \times n$ matrix of random effects; $\tilde{G} \sim MN_{d\times n}(0,\ V_g, K)$ where V_g is a $d \times d$ symmetric matrix of genetic variance component; K is a known $n \times n$ relatedness matrix; $\tilde{E}$ is a $d \times n$ matrix of residual errors; $\tilde{E} \sim MN_{d\times n}(0,\ V_e, I_{n\times n})$ where V_e is a $d \times d$ symmetric matrix of environmental variance component; $I_{n\times n}$ is a $n \times n$ identity matrix; and $MN_{d\times n}(0,\ V_1,\ V_2)$ denotes the $d \times n$ matrix normal distribution with mean 0, row covariance matrix V_1 ($d \times d$), and column covariance matrix V_2 (n$\times$ n).

3. RESULTS

3.1. Heritabilities

The variance components are presented in Table 1 to provide an overview of the genetic variance, genotype by location interaction variance, residual variance, and the broad-sense heritability for each trait in Diplodia and Fusarium ear rot experiments. For DER experiments, the disease measurements traits, percentage and score, did not present significant genotypic variation among the inbred lines. However, there was a significant genotypic variation among inbred lines ($p > 0.01$) for yield. For genotype by location interaction, all traits presented a significant variation. For FER experiments, significant genotypic variation among inbred lines ($p > 0.01$) and for genotype by location interaction ($p > 0.01$) were observed in the analysis for all traits.

Table 1 – Variance components and heritability for percentage of rotten kernels (PRK), incidence score (SCO) and yield (YLD) evaluated for Diplodia ear rot (DER) and Fusarium ear rot (FER).

	trait	**V_g**	**V_{gxl}**	**V_e**	**h^2**
Diplodia Ear Rot	PRK	0.0000	0.1789**	0.00472	0.0000
	SCO	0.0524	1.6838**	0.6636	0.0517
	YLD	0.2045**	0.4893**	0.4216	0.3871
Fusarium Ear Rot	PRK	0.0524**	0.1478**	0.0634	0.3417
	SCO	1.2917**	1.7447**	0.5356	0.5697
	YLD	0.7041**	0.2356**	0.4653	0.7720

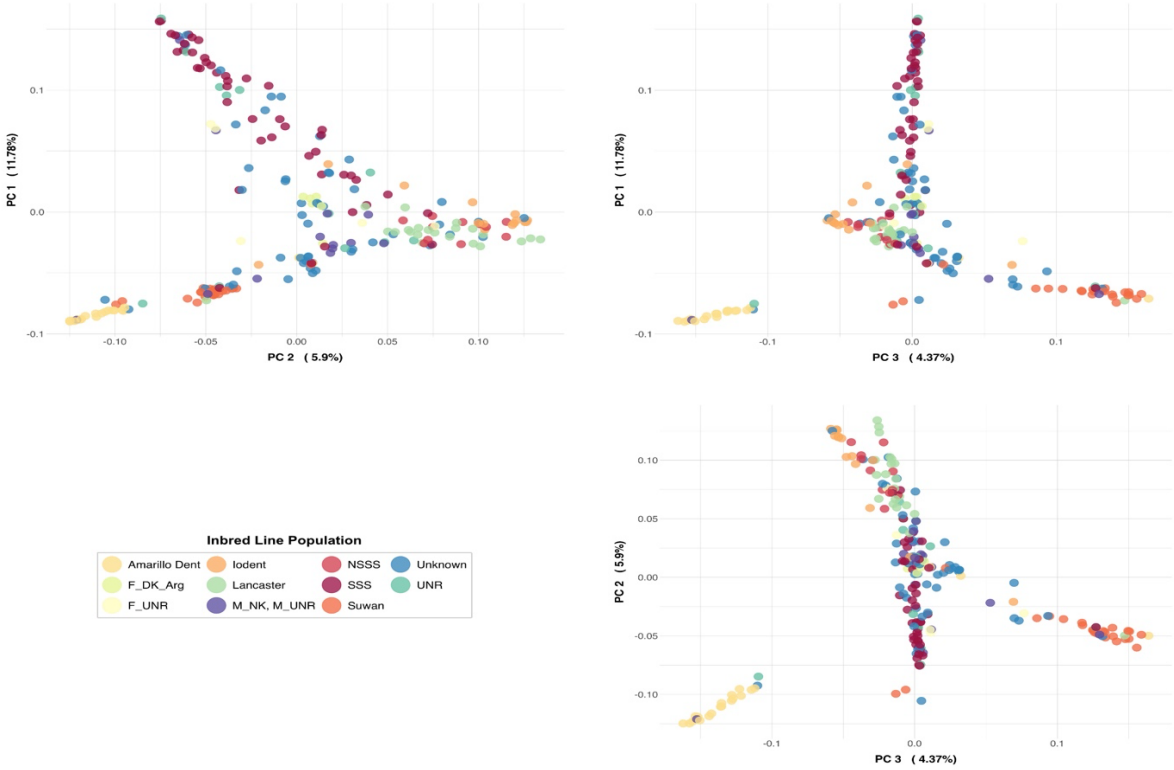
** Significant at 0.01 probability level (p > 0.01).

The broad-sense heritabilities were estimated separately for each ear rot experiment. The heritabilities for the traits evaluated in DER experiments were low to moderate, ranging from zero to 0.3871. On the other hand, the traits evaluated in the FER experiments showed moderate to high heritabilities, ranging from 0.3417 to 0.7720.

3.2. Population Structure

The PCA clustering method showed that most of the variance was explained by the first three principal components. This is shown in Figure 1 which plots the principal components (PCs) in a pair plot of the first three PCs. The graph shows the first three PCs explained 22.05% of the variation.

Figure 1 – Genetic relationships between the 228 lines of the diverse panel visualized using a principal component analysis of the relatedness matrix. In the first two graphs the horizontal axes are the first component and second and third components, respectively. In the third graph the horizontal axis is the second and the third components. Each point in the graphs represent an inbred line and the colors represent the different subpopulations.

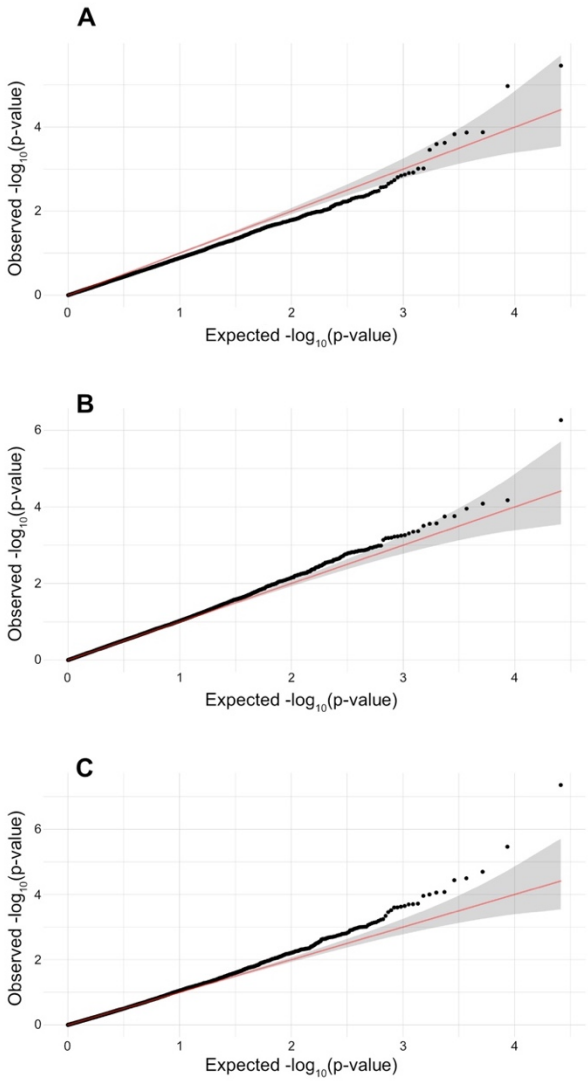


The plot shows a distinction between the inbred lines populations and the genetic characterization is similar to germplasm characterization reported by DOS SANTOS et al (2016). The plot has a triangle shape, which shows a distinction of the three main heterotic groups used in breeding programs. On the left-hand side of the plots, it is possible to observe the tropical subgroups, which includes Amarillo Dent and Suwan groups. The temperate groups are clearly grouped apart from the tropical groups. In the upper side of the plot, the temperate groups observed are the Stiff Stalk Synthetic (SSS) and UNR. On the right-hand side of the plot, there are the other temperate groups Lancaster, Iodent, Non-Stiff Stalk Synthetic (NSSS). In the center of the plot, it is possible to observe the F_DK_Arg and M_NK, M_UNR groups.

These two groups are in between the two temperate groups so, it is possible to suggest that inbred lines were derived from temperate germplasms.

In addition to the visual inspection of the population structure, the Bayesian information criteria were used to define the number of PCs to account for the population structure. The results of the analysis showed that no fixed effect covariates were necessary to account for the population structure in the GWAS analyses. This shows that the subgroups of the population do not have a clear influence on the variability of the traits.

Figure 2 – Quantile-quantile plots of observed versus expected -log10(p-values) for association analysis for A) Score evaluated in DER - univariate; B) Score evaluated in FER - univariate and; C) for the traits evaluated in FER –



The quantile-quantile plots show that the models had a good fit of GWAS p-values to the null hypothesis for score evaluated for DER and FER in the univariate analyses and the

traits evaluated for FER in the multivariate analysis. This is shown in Figure 2, which plots the expected versus observed -$\log_{10}$(p-values) for association analysis. In Figure 2A, most of the observed p-values for the trait score evaluated in DER experiments were distributed below the line of the expected p-values. Although this may indicate an absence of marker significant associated with the trait, there were few markers that were above the expected p-value line. Additionally, the plot shows that the GWAS model is adequate for population structure control.

3.3. Univariate GWAS

Univariate GWAS was performed on each trait separately for DER and FER, independently to distinguish markers associated specifically with trait and disease. We used the version of AGPv5 of the B73 genome reference with a 100 Kbp window for each marker that was significantly associated with the traits to identify gene models. In the univariate GWAS, two markers were significantly associated with ear rot resistance (Table 2). One marker was significantly associated with score evaluated for DER and the second marker was significantly associated with score evaluated for FER.

For the trait score evaluated for DER, one marker was significantly associated with the trait in the univariate analysis. This is shown in Figure 3, which plots the -$\log_{10}$(p-values) of the markers for each chromosome at 5% false discovery rate threshold (FDR; red dashed line) and at 5% Bonferroni threshold (gray dashed line). The marker was identified on chromosome 7 (marker 4581294; position 134,906,163 – 134,919,295 bp).

For the trait score evaluated for FER, one marker was significantly associated with the trait in the univariate analysis. This is shown in Figure 4, which plots the -$\log_{10}$(p-values) of the markers for each chromosome at 5% false discovery rate threshold (FDR; red dashed line) and at 5% Bonferroni threshold (gray dashed line). The marker was identified on chromosome 3 (marker 4768360; position 218,972,244 – 218,978,776 bp).

For the trait score evaluated for FER, one marker was significantly associated with the trait in the univariate analysis. This is shown in Figure 4, which plots the -$\log_{10}$(p-values) of the markers for each chromosome at 5% false discovery rate threshold (FDR; red dashed line) and at 5% Bonferroni threshold (gray dashed line). The marker was identified on chromosome 3 (marker 4768360; position 218,972,244 – 218,978,776 bp).

Table 2 – Significant markers associated with Diplodia ear rot and Fusarium ear rot in the univariate GWAS analyses. Chromosome, marker identification, marker position (AGPv5), minor allele frequency (MAF), estimated marker effect, -log10(p-value), closest gene and gene annotation.

Analysis	**chr**	**Marker ID**	**Position**	**MAF**	**Effect**	**-log$_{10}$(p-value)**	**Closest candidate gene**	**Annotation**
Diplodia – Score	7	4581294	134,912,61 - 134,912,680	0.802	0.3888	5.4616	Zm00001eb315780	S-adenosyl-L-methionine-dependent methyltransferase superfamily protein gpm 647
Fusarium – Score	3	4768360	218,972,244 - 218,978,776	0.661	0.5951	6.2647	Zm00001eb158100	CTP synthase; umi8

Figure 3 – Manhattan plots showing significant association in the univariate GWAS analysis for Diplodia ear rot (DER) for the trait score (SCO). The horizontal axes indicate chromosomes and the physical position of markers, and the vertical axes indicate the -log10(p-value) scores. The grey dashed line shows the significance threshold of Bonferroni correction, nominal significance of 5% ($-\log_{10}(0.05/12{,}936)$ = p-value > 5.41283. The red dashed line shows the significance threshold of FDR correction with a nominal significance of 5%.

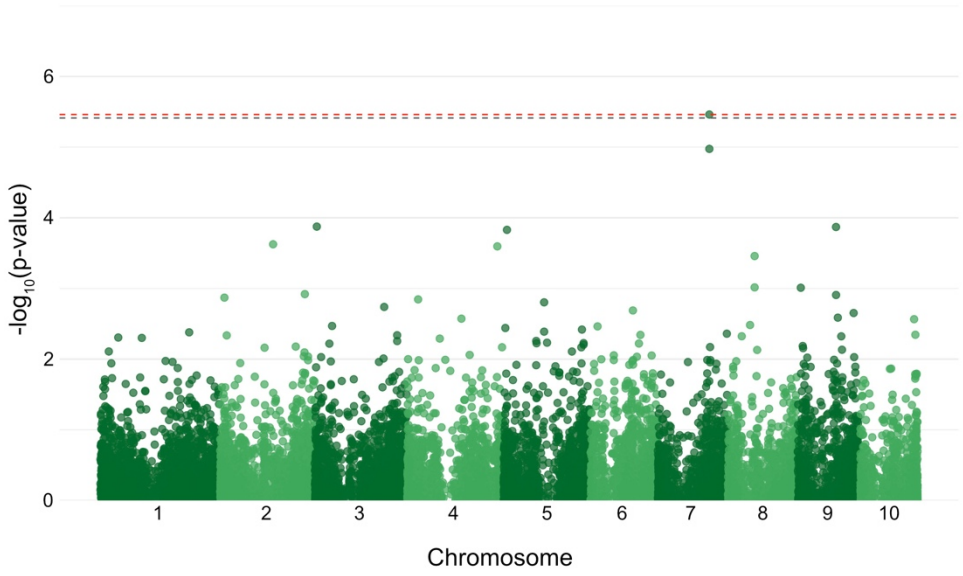
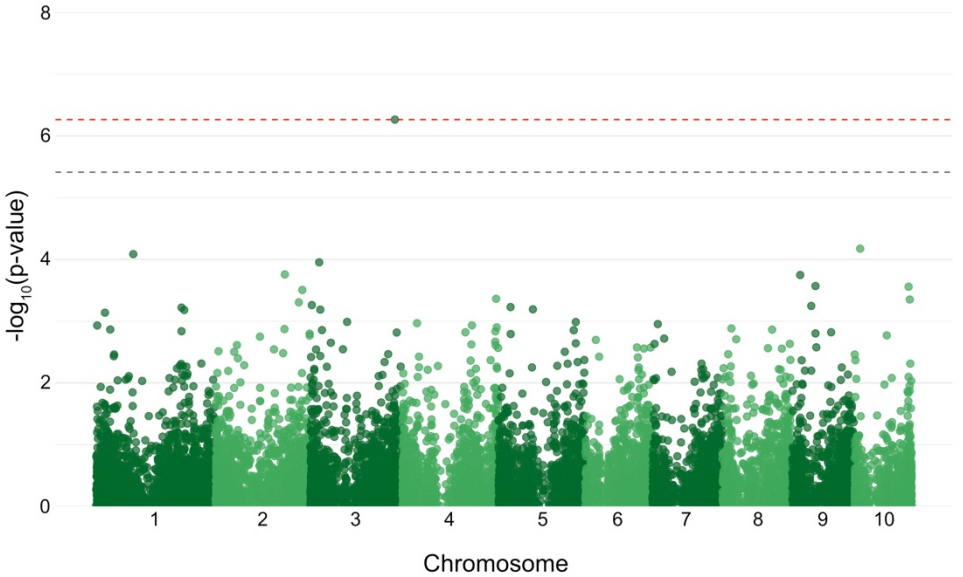


Figure 4 – Manhattan plots showing significant association in the univariate GWAS analysis for Fusarium ear rot (FER) for the trait score (SCO). The horizontal axes indicate chromosomes and the physical position of markers, and the vertical axes indicate the -log10(p-value) scores. The grey dashed line shows the significance threshold of Bonferroni correction, nominal significance of 5% ($-\log_{10}(0.05/12{,}936)$ = p-value > 5.41283. The red dashed line shows the significance threshold of FDR correction with a nominal significance of 5%.



3.4. Multivariate GWAS

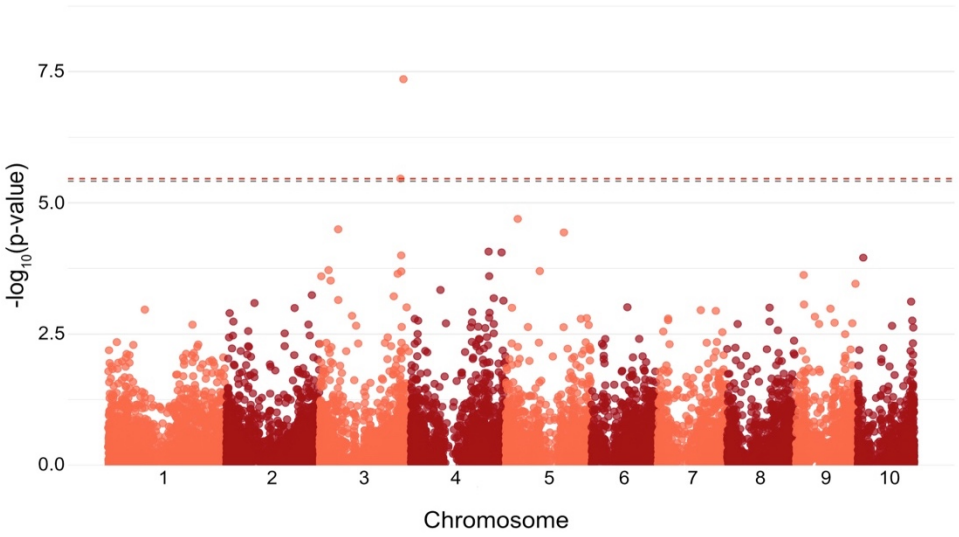
The multivariate GWAS was performed using three datasets. The analysis was performed using the data of traits evaluated for DER, using the data of traits evaluated for FER and, using the data of both DER and FER. The idea to perform the multivariate analysis separately for each ear rot was to assess if the traits could improve the analysis for each disease independently. We used the version of AGPv5 of the B73 genome reference with a 100 Kbp window for each marker that was significantly associated with the traits to identify gene models.

Table 3 – Significant markers associated with Diplodia ear rot and Fusarium ear rot in the multivariate GWAS analyses. Chromosome, marker identification, marker position (AGPv5), minor allele frequency (MAF), estimated marker effect, -log10(p-value), closest gene and gene annotation.

Analysis	chr	Marker ID	Position	MAF	-log$_{10}$(p-value)	Closest Candidate gene	Annotation
Fusarium	3	4768360	218,972,244 - 218,978,776	0.659	7.3566	Zm00001eb158100	CTP synthase; umi8
Fusarium	3	4767263	211,321,271 - 211,344,685	0.332	5.4619	Zm00001eb155870	uncharacterized protein LOC100274395

In the multivariate analyses, two markers were significantly associated with FER analysis (Table 3). This is shown in Figure 5, which plots the -log$_{10}$(p-values) of the markers for each chromosome at 5% FDR threshold (red dashed line) and at 5% Bonferroni threshold (gray dashed line). For the FER dataset, the peaks association were identified both on chromosome 3 (marker 4768360; position 265,503,622 bp and marker 4767263; position 211,321,271 – 211,344,685 bp). The marker 4768360, the most significant marker of these, was also significant in the univariate GWAS for score in FER experiment.

Figure 5 – Manhattan plots showing significant association in the multivariate GWAS analysis with the traits evaluated for Fusarium ear rot (FER). The horizontal axes indicate chromosomes and the physical position of markers, and the vertical axes indicate the -log10(p-value) scores. The grey dashed line shows the significance threshold of Bonferroni correction, nominal significance of 5% ($-\log_{10}(0.05/12{,}936)$ = p-value > 5.41283. The red dashed line shows the significance threshold of FDR correction with a nominal significance of 5%.



3.5. Correlations

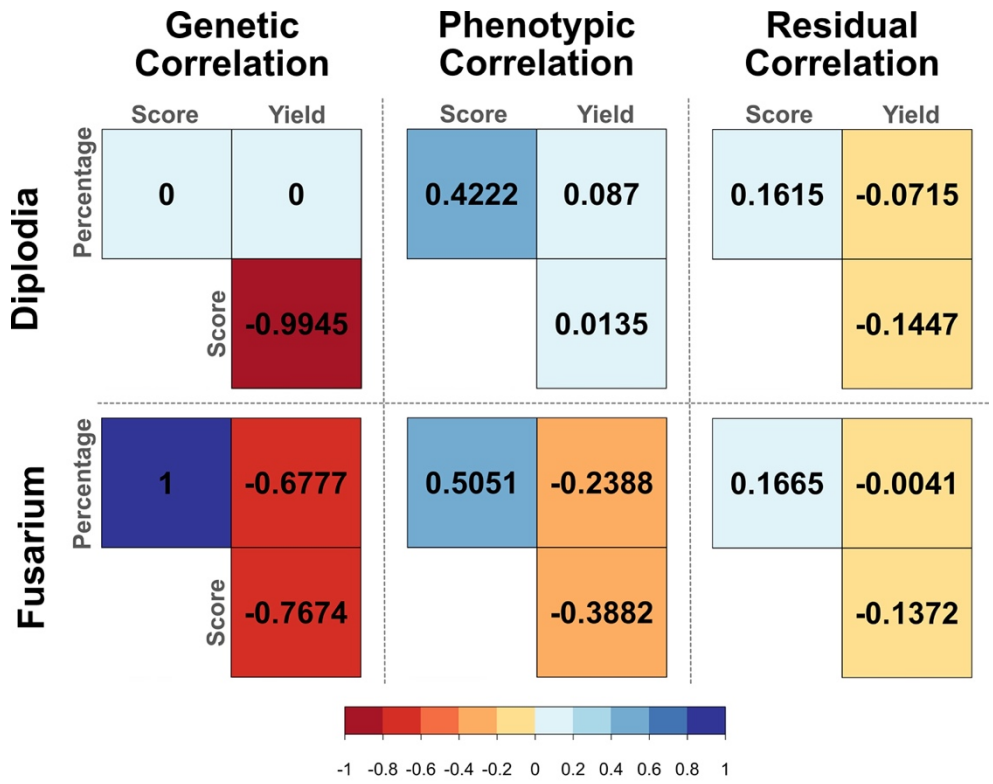
The genetic correlations were positively high between disease measurement traits (percentage and score) and negatively high for yield and disease measurement traits for DER and FER. For DER, score and yield were negatively high correlated with each other, this was expected once a high score of DER reduces the yield. For the other traits, it was not possible to calculate the genetic correlations because the genetic variance for percentage was zero for DER. For FER, percentage and score were positively high correlated with each other. For yield, both percentage and score were negatively high correlated.

Phenotypic correlations were weak to moderate between all traits for both DER and FER. Percentage and score were positively moderately correlated in DER and FER. For DER, the disease measurement traits showed a low correlation, close to zero, when correlated to yield. In contrast, the correlations between yield and disease measurement traits were negatively moderate in FER.

Residual correlations were low between traits for both DER and FER. The residual correlations were like the genotype by location interaction correlations. The correlation

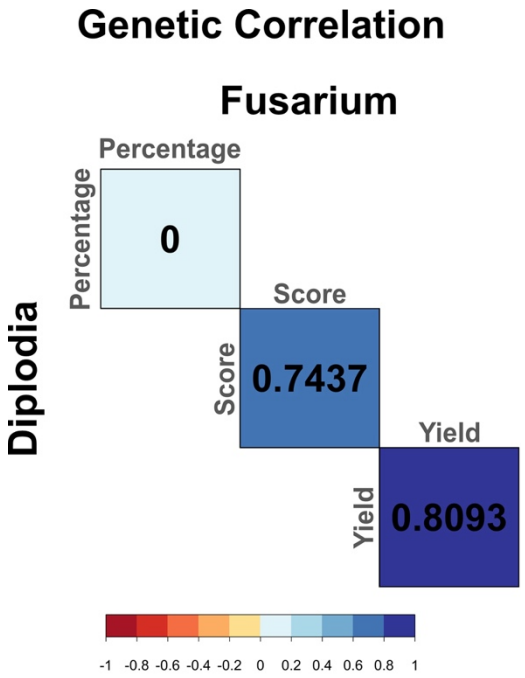
between the disease measurement traits was positive but presented a negative magnitude when correlated to yield.

Figure 6 – Correlations between traits evaluated in the same disease experiment (DER or FER).



The genetic correlation for the same traits evaluated for DER and FER was positively high for score and yield. This is shown in Figure 7, which plots the genetic correlations for percentage, score, and yield between DER and FER. For yield, the genetic correlation between diseases was 0.8093 and for score, the correlation was 0.7347. For percentage, it was not possible to calculate the correlation because the heritability for this trait was zero for DER.

Figure 7 – Correlations for same traits evaluated in the different disease experiments.



4. DISCUSSION

To assess the ability of Genome-Wide Association Studies to identify the loci controlling the resistance to Diplodia and Fusarium ear rots, we analyzed the performance of univariate and multivariate GWAS. The results in the present study are highlighted in three main topics for discussion:

i) Univariate and Multivariate GWAS;

ii) Factors that affect the analysis; and

iii) Benefits and drawbacks of multivariate GWAS.

4.1. Univariate and multivariate GWAS

The results obtained in this study show that both univariate and multivariate GWAS can identify potential loci controlling the resistance for ear rot in maize. These approaches may help to understand the mechanisms underlying resistance to Diplodia and Fusarium ear rot. Additionally, multivariate approaches could contribute to the understanding of pleiotropy.

In the present study, in the univariate analyses, two markers were significantly associated with ear rot resistance. The first marker significantly associated was found in DER analysis for the trait score. The marker was located in the position 134,912,616 - 134,912,680 bp (marker 4581294) in chromosome 7. This maker was located inside the gene model Zm00001eb315780 (position 134,906,163 - 134,919,295 bin 7.03), which is responsible for methyltransferase activity. This gene belongs to the S-adenosyl-L-methionine-dependent methyltransferase superfamily protein, and it is related to many biochemical reactions in plants. Methylated secondary metabolites can be synthesized and accumulated in response to development, stress, or other metabolic signals (MOFFATT; WERETILNYK, 2001).

The second marker significantly associated with ear rot resistance in univariate analysis was found in FER analysis for the trait score. The marker was located in the position 218,981,973 - 218,982,021 bp (marker 4768360) in chromosome 3. This marker was located 3Kbp from the gene model Zm00001eb158100 (umi8; position 218,972,244 - 218,978,776; bin 3.08), which encodes proteins with putative functions as CTP synthase. This gene is responsible for the synthesis of cytidine triphosphate (CTP) that is essential for gene transcription and nucleic acid synthesis during cell division (DAUMANN et al., 2018). Usually, cell division rates are high during germination and embryo development but also increases during biotic or abiotic stresses, there is a high demand for CTP. BASSE (2005) conducted a study to

understand the defense-related and developmental transcriptional response during Ustilago maydis infection in maize. A strong up-regulation of Umi8 and other Umi genes was found during *U. maydis* infection.

In the multivariate analysis, two markers were significantly associated with ear rots. Both markers were found in the analysis with the traits evaluated in FER experiments. One marker was the same associated with the trait score for FER in the univariate analysis (marker 4768360). The second marker was located in the position 211,320,3745 - 211,320,439 bp (marker 4767263) in chromosome 3. This marker was located 832 bp from the gene model Zm00001eb155870 (uncharacterized LOC100274395; position 211,321,271 - 211,344,685), but there is no predicted function for the gene for maize.

The resistance to ear rots is complex and most of genetic and biochemical pathways of the resistance to the ear rots still unknown. The results obtained by the univariate and multivariate GWAS are promising. The significant markers were located inside or close to the gene models that may be important for metabolic processes and may be related to resistance to ear rots. The candidate genes found in this study may be used in further studies to understand better and validate the mechanisms of resistance to ear rots in maize.

4.2. Factors that affect the analysis

Multivariate GWAS demonstrated to be an effective approach to understand the genetic architecture of a trait when multiple traits are available. This is the first study using a multivariate GWAS approach for ear rot diseases in maize, and we would like to highlight some factors that may affect the GWAS analyses for ear rot resistance.

Although the genetic correlations between traits, either for DER and FER, presented high correlations, the phenotypic correlations were only moderate between traits. This may be an indication that the genotype by environment interactions may have influenced the analysis. The main cause of the strong genotype by environment or genotype by disease interaction is the different climates between the locations. The climate in Lavras is characterized as humid subtropical and Uberlândia has a tropical savanna climate. The incidence of DER usually is restricted to humid and moderate to low temperature climates, which is observed in Lavras. On the other hand, the incidence of FER is restricted to dry and hot temperature climates, which is observed in Uberlandia. Evaluating the same diverse panel, DOS SANTOS et al. (2016) e PEREIRA et al. (2015) observed strong interaction genotype by environment interaction for all traits evaluated.

The architecture for disease resistance and yield are complex and highly polygenic and, hence become challenging to identify causal loci with small effects in GWAS. The few markers significantly associated with the traits analyzed for DER and FER might be an indication of the complexity of the resistance. The genetic architecture of DER or FER is highly polygenic, and the variants explain small effects that usually are not detected in GWAS (COAN et al., 2018; DE JONG et al., 2018; HOLLAND et al., 2020; KUKI et al., 2020; ZILA et al., 2013, 2014). Additionally, rare allele variants with small effects may also control the resistance to ear rots. Usually, in GWAS analyses, markers with minor allele frequency below the 0.05 threshold are removed. If these rare alleles are components of the genetic architecture, they will be missed. Consequently, the power of analysis to detect these variants will be reduced.

The constitution of the diverse panel used in a GWAS is important to identify casual regions related to ear rot resistance. Using a wider diverse panel or germplasm, historical recombination and natural genetic diversity are exploited and result in a higher mapping resolution. It is well known that tropical inbred lines are the resistance source to ear rots. In a study to select inbred lines resistant to DER, DOS SANTOS et al. (2016) found that almost 80% of the resistant lines were clustered in the tropical subgroup and the Iodent and Stiff Stalk Synthetic from the temperate group were the most susceptible lines. Most of the lines utilized in this study were clustered into temperate groups and only Suwan and Amarillo Dent were clustered in the tropical group. This fact may be the reason why not many markers were significantly associated with traits evaluated.

The size of the population may be another factor that can influence the identification of causal loci. Usually, only the main QTL are detected when small populations are used in QTL mapping. With small sample sizes, there is a limitation of the detection power, which is not sufficient to identify small and medium QTL effects (STANGE et al., 2013). WANG; XU, (2019) demonstrated that 500 individuals are adequate to detect a QTL that explains 5% of the phenotypic variance. Although, more individuals are necessary to detect rare variants, dominance, and epistatic effects. In this study, the markers significantly associated with the traits presented small effects, despite the fact of the small population size. This demonstrates that studies with small sample sizes tend to identify loci with large effects, being necessary a bigger sample to detect loci with small to medium effects.

4.3. Benefits and drawbacks of multivariate GWAS

In plant breeding programs routinely, several traits are collected on the same plant, and these traits may share common genetic or environmental factors. For this reason, the multivariate GWAS is the recommended approach to consider all the traits simultaneously. The main advantages of the multivariate approach compared to the univariate approach are the increase of the power of the analysis, the reduction of the number of tests, and the possibility of detecting the presence of pleiotropy (FERNANDES et al., 2021; GALESLOOT et al., 2014; JIANG; ZENG, 1995; LIU et al., 2009; RICE; FERNANDES; LIPKA, 2020; ZHOU; STEPHENS, 2012, 2014; ZHU et al., 2016).

Multivariate GWAS increased the power of the analysis when compared to the univariate GWAS. The results showed that the marker 4768360 which presented a $-\log_{10}$(p-value) of 6.2647 in the univariate analysis, increased the $-\log_{10}$ (p-value) to 7.3566 in the multivariate analysis. The extra information from cross-trait covariance provides more power to the multivariate approach (GALESLOOT et al., 2014). The power of the multivariate analysis is affected by the degree and the direction of the residual correlations (JIANG; ZENG, 1995; LIU et al., 2009). In this study, we observed that score was the trait that presented a higher degree of correlation with percentage and yield. Particularly for FER, the genetic correlation was perfect between score and percentage and high negatively between score and yield. The other correlations presented similar tendencies, although the degrees were more moderate. The increase in the power of the multivariate approach can also be observed in Figure 5. Some markers that presented low $-\log_{10}$(p-values) in the univariate analysis, even not being significantly associated with the trait, increased in the multivariate analyses.

Multivariate and univariate GWAS need to be used to complement each other, especially to detect causal loci with pleiotropic effects. The results showed that the multivariate approach was not only capable to detect the same marker associated with the trait found in the univariate analysis (marker 4768360), but also a new marker significantly associated with trait (marker 4767263) that did not appear in the univariate analyses. Moreover, studies showed that multivariate GWAS can facilitate the identification of loci that are pleiotropic loci or in linkage disequilibrium. The use of both multivariate and univariate GWAS may complement each other and hence boost the information available of multiple traits and help to elucidate the genetic architecture of the traits (FERNANDES et al., 2021; RICE; FERNANDES; LIPKA, 2020).

The multivariate GWAS performs a single test for the association for multiple traits. When a trait is tested separately, adjustments are necessary for type I error using single-trait-based tests, usually Bonferroni correction or Benjamini-Hochberg procedure. The correction

procedures, especially the Bonferroni that tends to be more conservative and, may lead to a loss of power. In a univariate GWAS, the statistical significance of a possible causal locus may reduce if adjusted to the number of loci. On the other hand, in a multivariate GWAS is likely that the potential causal locus remains significant (LIU et al., 2009). The number of tests is reduced because multiple traits are tested at once and hence alleviates the burden of multiple testing compared to analyzing all traits separately (GALESLOOT et al., 2014). In a simulation study, (ZHU; ZHANG, 2009) testing different causal relationships observed that the estimated type I for the multiple traits analyses were close to the nominal significance level than those of the univariate analyses.

Multivariate GWAS has many advantages when compared to the univariate approach but there are also some limitations. The multivariate approach has some computational barriers, the analysis becomes extremely computationally expensive when the number of traits increases. This is because the number of iterations required to converge increases as the number of traits also increase (ZHOU; STEPHENS, 2014). In this study, we included in the multivariate analyses only three traits, but it was possible to realize that the multivariate approach is more computationally demanding than the univariate approach. Analyzing 10 or more traits would not be feasible since the computational time required to run the analyses is overly demanding (RICE; FERNANDES; LIPKA, 2020).

A further drawback of the multivariate approach is the difficulty to interpret the effect of the causal locus. Additionally, the causal locus identified in the multivariate GWAS does not necessarily mean that all the traits evaluated in the analysis are related to this locus(RICE; FERNANDES; LIPKA, 2020). In this case, the use of univariate GWAS as posteriori analysis to complement the multivariate and provide more information about the genetic architecture of the trait and if the causal locus has pleiotropic effects (FERNANDES et al., 2021).

5. CONCLUSIONS

This is the first report of multivariate GWAS to understand the resistance to ear rot caused by Diplodia ear rot and Fusarium ear rot. The few markers associated with the traits evaluated reinforce the complexity of the genetic architecture of ear rots. The multivariate is a promising tool to use when multiple traits are available. Although, both approaches should be used as a complement to each other. So, it is possible to maximize the amount of information obtained from multiple traits.

REFERENCES

BASSE, C. W. Dissecting Defense-Related and Developmental Transcriptional Responses of Maize during *Ustilago maydis* Infection and Subsequent Tumor Formation. **Plant Physiology**, v. 138, n. 3, p. 1774–1784, 1 jul. 2005.

BRASIL. Portaria n. 11 de 12 de abril de 1996: Estabelece critérios complementares para classificação do milho. **Diário oficial da União**, n. 72, 1996.

BURDON, R. Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. **Silvae Genet**, v. 26, n. 5–6, p. 168–175, 1977.

BUTLER, D. G. et al. **ASReml-R Reference Manual Version 4**. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd, 2017.

CANTELMO, N. F.; VON PINHO, R. G.; BALESTRE, M. Genomic analysis of maize lines introduced in the early stages of a breeding programme. **Plant Breeding**, v. 136, n. 6, p. 845–860, dez. 2017.

CHEN, J. et al. Detection and verification of quantitative trait loci for resistance to Fusarium ear rot in maize. **Molecular Breeding**, v. 30, n. 4, p. 1649–1656, dez. 2012.

CHEN, J. et al. Genome-Wide Association Study and QTL Mapping Reveal Genomic Loci Associated with *Fusarium* Ear Rot Resistance in Tropical Maize Germplasm. **G3 Genes|Genomes|Genetics**, v. 6, n. 12, p. 3803–3815, 1 dez. 2016.

CLEMENTS, M. J. et al. Evaluation of Inoculation Techniques for Fusarium Ear Rot and Fumonisin Contamination of Corn. **Plant Disease**, v. 87, n. 2, p. 147–153, fev. 2003.

COAN, M. M. D. et al. Genome-Wide Association Study of Resistance to Ear Rot by *Fusarium verticillioides* in a Tropical Field Maize and Popcorn Core Collection. **Crop Science**, v. 58, n. 2, p. 564–578, mar. 2018.

DE JONG, G. et al. Genome-wide association analysis of ear rot resistance caused by Fusarium verticillioides in maize. **Genomics**, v. 110, n. 5, p. 291–303, set. 2018.

DING, J.-Q. et al. QTL mapping of resistance to Fusarium ear rot using a RIL population in maize. **Molecular Breeding**, v. 22, n. 3, p. 395–403, out. 2008.

DOS SANTOS, J. P. R. et al. Genomic selection to resistance to Stenocarpella maydis in maize lines using DArTseq markers. **BMC Genetics**, v. 17, n. 1, p. 86, dez. 2016.

ENDELMAN, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. **The Plant Genome**, v. 4, n. 3, p. 250–255, nov. 2011.

FERNANDES, S. B. et al. How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? **Frontiers in Genetics**, v. 11, p. 602526, 8 jan. 2021.

GALESLOOT, T. E. et al. A Comparison of Multivariate Genome-Wide Association Methods. **PLoS ONE**, v. 9, n. 4, p. e95923, 24 abr. 2014.

GRUBER, B. et al. DARTR : An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. **Molecular Ecology Resources**, v. 18, n. 3, p. 691–699, maio 2018.

GUTIÉRREZ, H. I. **Mapeamento de QTLs para resistência a grãos ardidos causados por diplodia (Stenocarpella Sp.) em milho (Zea Mays L.)**. Master's Thesis—[s.l.] Universidade Federal de Uberlândia, 28 fev. 2008.

HOLLAND, J. B. et al. Genomic prediction for resistance to Fusarium ear rot and fumonisin contamination in maize. **Crop Science**, v. 60, n. 4, p. 1863–1875, jul. 2020.

JIANG, C.; ZENG, Z. B. Multiple trait analysis of genetic mapping for quantitative trait loci. **Genetics**, v. 140, n. 3, p. 1111–1127, 1 jul. 1995.

KELLERMAN, T. S. et al. Perinatal mortality in lambs of ewes exposed to cultures of Diplodia maydis (= Stenocarpella maydis) during gestation. **The Onderstepoort journal of veterinary research**, v. 58, n. 4, p. 297–308, dez. 1991.

KILIAN, A. et al. Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms. In: POMPANON, F.; BONIN, A. (Eds.). . **Data Production and Analysis in Population Genomics**. Methods in Molecular Biology™. Totowa, NJ: Humana Press, 2012. v. 888p. 67–89.

KING, S.; SCOTT, G. Genotypic differences in maize to kernel infection by Fusarium moniliforme. **Phytopathology**, v. 71, n. 12, p. 1245–1247, 1981.

KORTE, A. et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. **Nature Genetics**, v. 44, n. 9, p. 1066–1071, set. 2012.

KUKI, M. C. et al. Association mapping and genomic prediction for ear rot disease caused by *Fusarium verticillioides* in a tropical maize germplasm. **Crop Science**, v. 60, n. 6, p. 2867–2881, nov. 2020.

LANUBILE, A. et al. Molecular Basis of Resistance to Fusarium Ear Rot in Maize. **Frontiers in Plant Science**, v. 8, p. 1774, 12 out. 2017.

LIPKA, A. E. et al. GAPIT: genome association and prediction integrated tool. **Bioinformatics**, v. 28, n. 18, p. 2397–2399, 15 set. 2012.

LIU, J. et al. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. **Genetic Epidemiology**, v. 33, n. 3, p. 217–227, abr. 2009.

LOGRIECO, A. et al. Toxigenic Fusarium Species and Mycotoxins Associated with Maize Ear Rot in Europe. **European Journal of Plant Pathology**, v. 108, n. 7, p. 597–609, 2002.

MASCHIETTO, V. et al. QTL mapping and candidate genes for resistance to Fusarium ear rot and fumonisin contamination in maize. **BMC Plant Biology**, v. 17, n. 1, p. 20, dez. 2017.

MESTERHAZY, A.; LEMMENS, M.; REID, L. M. Breeding for resistance to ear rots caused by Fusarium spp. in maize–a review. **Plant Breeding**, v. 131, n. 1, p. 1–19, 2012.

MOFFATT, B. A.; WERETILNYK, E. A. Sustaining *S* -adenosyl- L -methionine-dependent methyltransferase activity in plant cells. **Physiologia Plantarum**, v. 113, n. 4, p. 435–442, dez. 2001.

MUNKVOLD, G. P.; MCGEE, D. C.; CARLTON, W. M. Importance of Different Pathways for Maize Kernel Infection by *Fusarium moniliforme*. **Phytopathology®**, v. 87, n. 2, p. 209–217, fev. 1997.

MUNKVOLD, G. P.; WHITE, D. G. **Compendium of corn diseases**. [s.l.] Am Phytopath Society, 2016. v. 165

NANKAM, C.; PATAKY, J. Resistance to kernel infection by Fusarium moniliforme in the sweet corn inbred IL 125b. **Plant disease (USA)**, 1996.

ODRIOZOLA, E. et al. Diplodia maydis: a cause of death of cattle in Argentina. **New Zealand Veterinary Journal**, v. 53, n. 2, p. 160–161, abr. 2005.

PEREIRA, G. S. et al. Indirect selection for resistance to ear rot and leaf diseases in maize lines using biplots. **Genetics and Molecular Research**, v. 14, n. 3, p. 11052–11062, 2015.

PÉREZ BRITO, D. et al. QTL mapping of Fusarium moniliforme ear rot resistance in highland maize, Mexico. 2001.

PINTO, L. R. et al. Reciprocal recurrent selection effects on the genetic structure of tropical maize populations assessed at microsatellite loci. **Genetics and Molecular Biology**, v. 26, n. 3, p. 355–364, 2003.

PRICE, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**, v. 38, n. 8, p. 904–909, ago. 2006.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2020.

REID, L. M. et al. Interaction of Fusarium graminearum and F. moniliforme in Maize Ears: Disease Progress, Fungal Biomass, and Mycotoxin Accumulation. **Phytopathology®**, v. 89, n. 11, p. 1028–1037, nov. 1999.

REID, L. M. et al. Effect of inoculation time and point of entry on disease severity in *Fusarium graminearum* , *Fusarium verticillioides* , or *Fusarium subglutinans* inoculated maize ears [1]. **Canadian Journal of Plant Pathology**, v. 24, n. 2, p. 162–167, jun. 2002.

RICE, B. R.; FERNANDES, S. B.; LIPKA, A. E. Multi-Trait Genome-Wide Association Studies Reveal Loci Associated with Maize Inflorescence and Leaf Architecture. **Plant and Cell Physiology**, v. 61, n. 8, p. 1427–1437, 1 ago. 2020.

ROBERTSON-HOYT, L. A. et al. QTL Mapping for Fusarium Ear Rot and Fumonisin Contamination Resistance in Two Maize Populations. **Crop Science**, v. 46, n. 4, p. 1734–1743, jul. 2006.

ROMERO LUNA, M. P. **Managing Diplodia ear rot in corn: Short and long-term solutions**. Doctoral dissertation—[s.l.] Purdue University, 2012.

ROSSOUW, J. D. et al. Breeding for Resistance to *Stenocarpella* Ear Rot in Maize. In: JANICK, J. (Ed.). . **Plant Breeding Reviews**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2009. p. 223–245.

ROSSOUW, J. D.; VAN RENSBURG, J. B. J.; VAN DEVENTER, C. S. Breeding for resistance to ear rot of maize, caused by *Stenocarpella maydis* (Berk) Sutton. 2. Inheritance of resistance. **South African Journal of Plant and Soil**, v. 19, n. 4, p. 188–194, jan. 2002a.

ROSSOUW, J. D.; VAN RENSBURG, J. B. J.; VAN DEVENTER, C. S. Breeding for resistance to ear rot of maize, caused by *Stenocarpella maydis* (Berk) Sutton. 1. Evaluation of selection criteria. **South African Journal of Plant and Soil**, v. 19, n. 4, p. 182–187, jan. 2002b.

SAMAYOA, L. F. et al. Genome-wide association analysis for fumonisin content in maize kernels. **BMC Plant Biology**, v. 19, n. 1, p. 166, dez. 2019.

STANGE, M. et al. High-density genotyping: an overkill for QTL mapping? Lessons learned from a case study in maize and simulations. **Theoretical and Applied Genetics**, v. 126, n. 10, p. 2563–2574, out. 2013.

VAN RENSBURG, J. B. J.; FERREIRA, M. J. Resistance of elite maize inbred lines to isolates of *Stenocarpella maydis* (Berk.) Sutton. **South African Journal of Plant and Soil**, v. 14, n. 2, p. 89–92, jan. 1997.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, nov. 2008.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. New York, NY: Springer New York, 2002.

WANG, M.; XU, S. Statistical power in genome-wide association studies and quantitative trait locus mapping. **Heredity**, v. 123, n. 3, p. 287–306, set. 2019.

WENZL, P. et al. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. **Proceedings of the National Academy of Sciences**, v. 101, n. 26, p. 9915–9920, 29 jun. 2004.

WICKLOW, D. T. et al. Bioactive metabolites from Stenocarpella maydis, a stalk and ear rot pathogen of maize. **Fungal Biology**, v. 115, n. 2, p. 133–142, fev. 2011.

ZHENG, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. **Bioinformatics**, v. 28, n. 24, p. 3326–3328, 1 dez. 2012.

ZHOU, X.; STEPHENS, M. Genome-wide efficient mixed-model analysis for association studies. **Nature Genetics**, v. 44, n. 7, p. 821–824, jul. 2012.

ZHOU, X.; STEPHENS, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. **Nature Methods**, v. 11, n. 4, p. 407–409, abr. 2014.

ZHU, W.; ZHANG, H. Why do we test multiple traits in genetic association studies? **Journal of the Korean Statistical Society**, v. 38, n. 1, p. 1–10, mar. 2009.

ZHU, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. **Nature Genetics**, v. 48, n. 5, p. 481–487, 28 mar. 2016.

ZILA, C. T. et al. A Genome-Wide Association Study Reveals Genes Associated with Fusarium Ear Rot Resistance in a Maize Core Diversity Panel. **G3 Genes|Genomes|Genetics**, v. 3, n. 11, p. 2095–2104, 1 nov. 2013.

ZILA, C. T. et al. Genome-wide association study of Fusarium ear rot disease in the U.S.A. maize inbred line collection. **BMC Plant Biology**, v. 14, n. 1, p. 372, dez. 2014.