MARCO ANTONIO PEIXOTO

APPLICATION OF QUANTITATIVE GENETICS TOOLS TO BREEDING PROGRAM OPTIMIZATION

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Leonardo Lopes Bhering

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
P379a
2023

Peixoto, Marco Antonio de Amorim, 1990-
Application of quantitative genetics tools to breeding program optimization / Marco Antonio de Amorim Peixoto. – Viçosa, MG, 2023.
1 tese eletrônica (179 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndices.
Orientador: Leonardo Lopes Bhering.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.135
Modo de acesso: World Wide Web.

1. Plantas - Melhoramento genético. 2. Genética quantitativa. I. Bhering, Leonardo Lopes, 1980-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 631.52

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

MARCO ANTONIO PEIXOTO

APPLINCATION OF QUANTITATIVE GENETICS TOOLS TO BREEDING PROGRAM OPTIMIZATION

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: March 8st, 2023.

Assent:

Documento assinado digitalmente
gov.br
MARCO ANTONIO DE AMORIM PEIXOTO
Data: 23/03/2023 11:51:21-0300
Verifique em https://validar.iti.gov.br

_____

Marco Antonio Peixoto
Author

Documento assinado digitalmente
gov.br
LEONARDO LOPES BHERING
Data: 23/03/2023 14:03:58-0300
Verifique em https://validar.iti.gov.br

_____

Leonardo Lopes Bhering
Advisor

To my parents, thank you for all the love and support you have given me throughout my life. I especially want to thank my mom for being a wonderful person and human being.

To my two brothers, for sharing existence.

DEDICATION

ACKNOWLEDGMENTS

This dissertation is a result of hard work and dedication of numerous individuals, who provided both direct and indirect support throughout the entire process.

Firstly, to **God** for guiding me in all steps I take, for giving me the opportunity to be born and reborn. The image of someone who guides me and watches over me.

To CNPq and PrInt-CAPES UFV for the scholarships that allows me to develop this project.

To my mother, Dona **Jacira**, I just want to take a moment to express my gratitude for everything you have done for me. Your help, support, and love have been invaluable to me and have made a tremendous impact on my life.

To my father **José** and my two brothers (**Marcílio** and **Mário**), thank you for all the support during difficult times that we have faced together. I love you all.

To the people from Laboratório de Biometria: **Suellen**, **Emannuel**, **Jeniffer**, **Felipe**, **Igor**, **Renan**, **Arthur**, **Andreia**, **Rodrigo**, and **Michelle**. Thank you for the friendship and all the learning during my PhD.

To the people in the Sweet Corn and Potato Genomics Lab: **Kristen**, **Sue**, **Mariana**, **Audrey**, **Rafaela**, **Rakshia**, **Juan**, **Vincent**, **Christina**, **Cleyssim**, **Leo**, **Gabi**, **Vinícius**, **Evandro**, and PI **Márcio**. Thank you for receiving me and helping me through this amazing experience.

To my fraternity in Ouro Preto: **República Taranoia**, the best place in the best city on earth. The friendship and love from all taranoicos mean a lot to me. I hope we can face the challenges together and move towards a glorious future. 'Aqui não, neném!'

To my good friends who have helped me along the way: **Cometa** and **JP Tosko (**for being around and being good friends), **Laiza (**for the talks and for being a dear and sweet person), **Igor Lorito** (thank you for all talking, time, and science that we shared), and **Renan** and **Jeniffer** (for all the friendship that came together with the PhD). You all have made a difference in my life, and I am grateful to have you all. Also, to the place where I spent good times in Viçosa, our República Valhalla. Thank you for the nonsense talks and friendship.

To my advisor Dr. **Leonardo L**. **Bhering**, I am so happy to have been part of your lab for the last four years. It has changed my life, and I am thankful for your support, professionalism, and passion.

To my co-advisor and friend Dr. **Márcio Resende**, thank you for the last few months that I have been part of your lab. I have learned a lot about quantitative genetics, leadership, and science from you. I hope I can share the knowledge you have imparted to me with a broad audience.

Also, I would like to say thank you to the professors who spent some time reading this dissertation and being a part of my committee: Dr. **Alencar Xavier**, Dra. **Camila Azevedo**, Dr. **Kaio Dias**, and Dr. **Márcio Resende**.

Thank you.

MUITO OBRIGADO!

The present work was financially supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001.

‘Stay hungry.

Stay foolish’

(Stewart Brand)

ABSTRACT

PEIXOTO, Marco Antonio, D.Sc., Universidade Federal de Viçosa, March, 2023. **Application of quantitative genetics tools to breeding program optimization.** Advisor: Leonardo Lopes Bhering.

Overall, the application of quantitative genetics theory has greatly influenced plant breeding programs over the last few decades. The aim of this study was to use and develop quantitative genetics tools to improve breeding programs. In the first chapter we simulated a hybrid crop breeding program and compared breeding pipelines with two strategies for parental updates, and we compare the gain and costs to implements genomic selection and high-throughput phenotyping into the pipeline. Our results suggest that early parental selection performs better and that high-throughput phenotyping together with genomic selection delivers the highest hybrid gain in the long-term. In the second chapter we evaluated the potential of implementing genomic selection in a sweet corn breeding program through hybrid prediction. We evaluated 506 hybrids in six environments and measured 21 traits. We considered eight statistical models for prediction and three cross-validation schemes CV1, CV0, and CV00. The results indicated RKHS model outperforming GBLUP models, and models with additive plus dominance kernels presented slight improvement for some traits. Therefore, we recommend using the RKHS model as a standard model for sweet corn hybrid prediction, and to implement genomic selection in sweet corn breeding programs to optimize the testcross stage and the candidates that reach the field stage. In the third chapter we describe SMate, a flexible R package for cross prediction and optimization, which represents a tool for breeding programs to balance genetic gains and inbreeding rate levels. The package builds a valid mate plan based on two core aspects: (i) prediction of usefulness for potential cross, and (ii) optimization of the set of crosses. In conclusion, SMate package enables to optimize cross selection in breeding programs targeting long term genetic gains while balancing genetic diversity and inbreeding rate levels. In summary, quantitative genetics tools have been largely applied in breeding programs and has evolved with it. Our study demonstrated potential to contribute to the quantitative genetics field and direct impact in breeding programs.

Keywords: Mate allocation. Inbreeding. Genomic hybrid prediction. Stochastic simulation.

RESUMO

PEIXOTO, Marco Antonio, D.Sc., Universidade Federal de Viçosa, março de 2023. **Aplicação de ferramentas em genética quantitativa na otimização de programas de melhoramento.** Orientador: Leonardo Lopes Bhering.

A aplicação da teoria da genética quantitativa influenciou positivamente programas de melhoramento de plantas nas últimas décadas. O objetivo deste estudo foi utilizar e desenvolver ferramentas em genética quantitativa para programas de melhoramento. No primeiro capítulo, simulamos um programa de melhoramento de híbridos e comparamos pipelines de melhoramento com duas estratégias para atualizar os parentais, e o ganho e custos para implementar a seleção genômica e de fenotipagem de alto rendimento no pipeline. Nossos resultados sugerem que a seleção parental precoce tem melhor desempenho e que a fenotipagem de alto rendimento, juntamente com a seleção genômica, oferece o maior ganho a longo prazo. No segundo capítulo avaliamos o potencial de implementação da seleção genômica em um programa de melhoramento de milho doce por meio da predição de híbridos. Avaliamos 506 híbridos em seis ambientes e medimos 21 características. Foram considerados oito modelos estatísticos para predição e três esquemas de validação cruzada. Os resultados indicaram que o modelo RKHS superou os modelos GBLUP. Portanto, recomendamos o uso do modelo RKHS como modelo padrão para predição de híbridos de milho doce e a implementação da seleção genômica em programas de melhoramento de milho doce. No terceiro capítulo descrevemos o SMate, um pacote flexível, em R, para predição de cruzamentos e otimização, que representa uma ferramenta para programas de melhoramento balancear ganhos genéticos e níveis de taxa de endogamia. O pacote constrói um plano de cruzamento com base em dois passos: (i) predição de utilidade do cruzamento e (ii) otimização da utilidade juntamente com restrição da taxa de endogamia. Em conclusão, o pacote SMate permite otimizar a seleção de cruzamentos em programas de melhoramento visando ganhos genéticos de longo prazo. Concluindo, as ferramentas de genética quantitativa têm sido amplamente aplicadas em programas de melhoramento e evoluído com ele. Nosso estudo demonstrou potencial para contribuir com o campo da genética quantitativa e impactar diretamente programas de melhoramento.

Palavras-chave: Alocação de cruzamentos. Endogamia. Predição de híbridos. **S**imulação estocástica.

LIST OF ILLUSTRATIONS

Introduction

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Table 1 - Genetic mean for both scenarios implemented in the simulations. The values referred to the year 40. DHGS: doubled haploid scenario with genomic selection and random crosses; OCS: doubled haploid scenario with genomic selection and optimal cross selection using **SMate** package. ΔF is the rate of inbreeding based on the observed heterozygosity for the year 40. Gain (%) represents the gain in percentage comparing DHGS and OCS scenarios.157

SUMMARY

Introduction

Intelligent breeding tools aiming to increase performance of plant breeding programs are mostly based on quantitative genetics theory. Over the last few decades, quantitative genetics has developed and provided much of knowledge and information for breeding programs, from experimental designs to applied selection tools (Lynch & Walsh, 1988; Falconer & Mackay, 1996; Isik et al., 2017). In a simple and impactful example, the understanding of inheritance and variation of quantitative traits, that plays a significant role in plant breeding performance improvement, came from estimates of traits genetic variance and heritabilities (Bernardo, 2020).

Two technological advancements have deeply impacted how quantitative genetics influences breeding. The first one is how computers have enhanced the power to process data and simulations. This improvement make feasible the implementation of simulations (*e.g.*, stochastic simulations), to handle large datasets (like markers and high-throughput data on phenotypic records), and to implement artificial intelligence with neurons with thousands of hundreds of hidden layers (Chen et al., 2009; Moreira et al., 2019; Zingaretti et al., 2020; Beyene et al., 2021). The second was the invention of molecular markers. When the genomic BLUP (GBLUP) was first proposed (Bernardo, 1994; Meuwissen et al., 2001), molecular datasets were neither abundant nor cheaper. However, over the years it became feasible to plant and animal breeding programs to implement genomic selection into their breeding pipelines. After a few decades of efforts in this subject, genomic selection has been proven to work inside breeding programs. The two most impactful features of genomic selection are the capability to shorten the breeding cycle by decreasing time for selection and to increase the accuracy in the selection process, especially for traits with lower heritability (Atanda et al., 2021). As a result, breeding programs started to modified their pipeline in a way to include genomic selection to maximize the benefits of the tool (Gaynor et al., 2017; Powell et al., 2020).

Thinking of a breeding program as sequential process, we can break it down into several stages or phases, from creating a new cross to the release of a new variety (Figure 1). Here, we divided it into three stages or phases. In the first phase we allocate and create crosses and assess several genotypes in a few environments together with initial selections. The second phase is the conduction of populations and selections. The third phase is the later yield trials, with few genotypes being assessed in many environments, largely impacted by genotype-by-environment interaction. Furthermore, in all of them, quantitative genetics tools have been applied and contributed to its improvement.

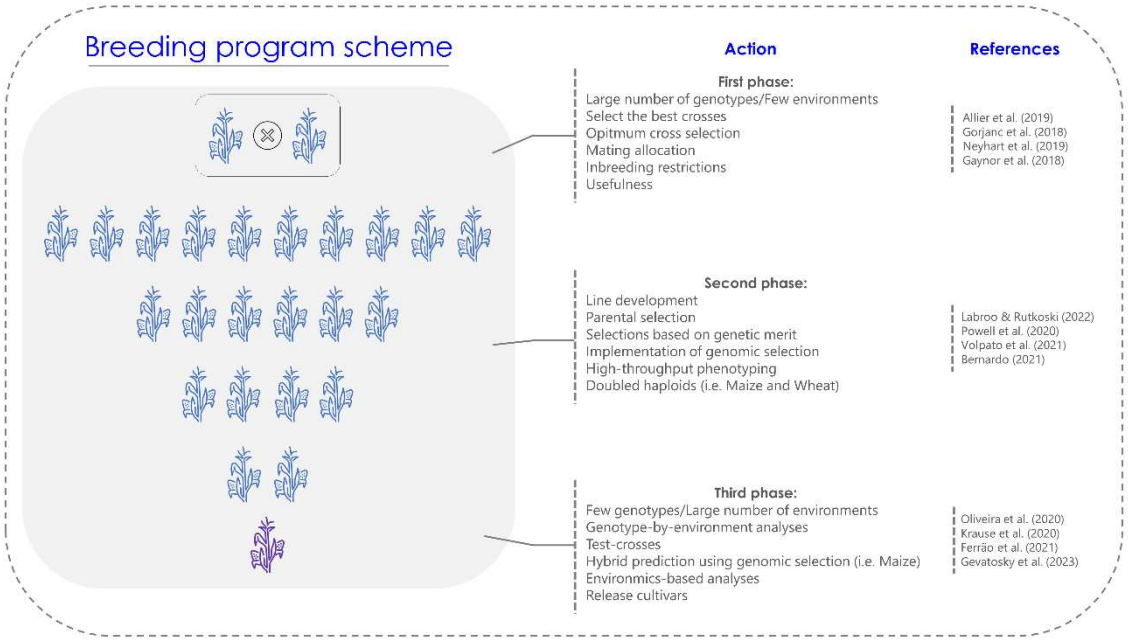


Figure 1 - Toy representation of a breeding program. Three phases are described, and the action lists some important characteristics and quantitative tools used in each phase. The references describe some studies that applied quantitative genetics to the respective phase of the breeding program.

For instance, in the beginning of the breeding program, we should select the best candidates for parents and set the new crosses for returning the best individuals in the next generation. This choice could ensure that the F_1 generation presents higher mean, compared to the parental population. Several endeavors over the last few years have improved this stage. One relevant proposal to optimize the mate plan selection came from Meuwissen and Sonesson

(1998) (Table 1). Moreover, in long-term, the selection process indirectly increases the population' inbreeding rates, which decreases the genetic variance and the possibility of improving genetic gains. Then, those authors proposed the optimum contribution selection, a strategy which maximizes the genetic value while minimize the inbreeding rates of the population. With the advance of molecular markers, a few ideas have been proposed toward the same goal, but using molecular markers. Besides genomic selection increases genetic gain in breeding programs, it fixes some alleles with large effects and loses some rare alleles in the process (Jannink, 2010). Ultimately, it increases the inbreeding rates and decreases genetic variability.

Table 1 - Description of proposed strategies which balance rates of genetic gain and loss of genetic diversity/inbreeding rates.

Techniques	Description	Reference
Optimum contribution selection	Genetic value is maximized while pedigree-based inbreeding is constrained to give the optimal contributions of parents to the next generation	(Meuwissen & Sonesson, 1998)
Weighting of rare alleles	Allelic effects are weighted by their frequency such that rare favorable alleles are preserved	(Goddard, 2009)
Weighted genomic selection	Allelic effects are weighted by their frequency and magnitude of their effect, preserving rare favorable alleles with large effects on EBV	(Jannink, 2010)
Genotype building	A subpopulation is selected to segregate for maximal haplotype values and intermated such that the two best segments segregate with equal frequency	(Kemper et al., 2012)
Genomic optimum contribution selection	Estimated breeding values are maximized while genomic-based inbreeding is constrained to give the optimal contributions of parents to the next generation	(Sonesson et al., 2012)
Optimal haploid value selection	Outbred individuals are selected for their predicted value to produce the best DH lines	(Daetwyler et al., 2015)
Genomic mating	Genetic value, inbreeding, and risk are optimized	(Akdemir & Sánchez, 2016)
IND-HE	Genetic gain and expected heterozygosity are balanced in selection	(De Beukelaer et al., 2017)
Optimal population value selection	Sets of individuals are selected for their collective maximum possible haploid value	(Goiffon et al., 2017)
Optimal contribution selection to update reference population	Selection of training set candidates balances genetic gain and inbreeding for updating the reference population	(Eynard et al., 2018)
Optimal cross selection	Selection intensity, inbreeding, and cross allocation are optimized	(Gorjanc et al., 2018)
Expected maximum haploid breeding value selection	Individuals are selected for their maximum possible haploid value	(Müller et al., 2018)
Usefulness criterion parental contribution	Overall and within-family selection intensity, inbreeding, and cross allocation are optimized	(Allier et al., 2019b)
Look-ahead selection	Sets of individuals are selected for their collective maximum possible haploid value in a user-specified target generation	(Moeinizade et al., 2019)
Optimal contribution selection with branching	The population mating scheme is branched to maintain diversity and maximize gain	(Santantonio & Robbins, 2020)

Table modified from Labroo et al., (2021).

Further, the inclusion of the variance of each cross can aid to explore the potential of a mate pair and may help handle genetic variance performance. Following Schnell & Utz (1976), the combination of genetic mean and genetic gain was derived from the a cross mean, selection intensity, and the square root of a cross' genetic variance altogether in a term known as usefulness.

The prediction of genetic mean is easily calculated from the genetic values of the parental candidates whilst genetic variance was just available preceded by coancestry information (Zhong & Jannink, 2007). With the development and advancement of genotyping and its consequently reduction in costs, the use of molecular markers improves the estimation of genetic variance, especially when linkage disequilibrium between quantitative trait loci (QTL) is included (Lehermeier et al., 2017). This leads to the development of tools and application of the concept of usefulness cross prediction. Therefore, few propositions have been trying to optimize a system that contains usefulness and inbreeding rates (Akdemir & Sánchez, 2016; Gorjanc et al., 2018; Allier et al., 2019a).

Regarding the second phase of a breeding program, several advancements has been pointed out in the last decades. With the advance of stochastic processes and power of computer processing capabilities, simulations became more popular (Chen et al., 2009; Faux et al., 2016; Gaynor et al., 2021; Villiers et al., 2022). One strength of simulations is related to the feasibility of its implementation, being simple to disentangling question involving genetic, statistical, and breeding. In addition, the cost to implement simulations in a breeding program is quite low. Moreover, simulations' results can guide the decision-making process, choosing the most likely approach to be implemented in the field and rejecting the ones with lower performance. Simple questions, such as the impact of increase the number of individuals selected inside each family, the impact of implementation of tools such as high-throughput phenotyping and doubled haploid, the number of crosses in the crossing block can be checked (Figure 2). More complex

questions such as the impacts of genomic selection and changes in breeding pipelines have been successfully applied (Bančič et al., 2021; Batista et al., 2021a; Sabadin et al., 2022).

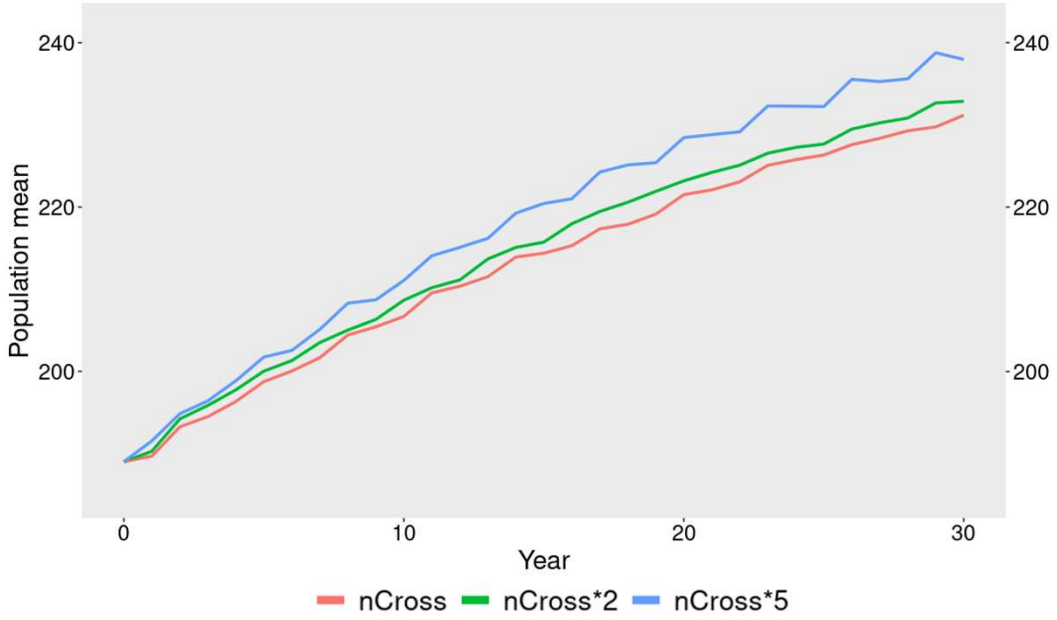


Figure 2 - Breeding program simulation using the package AlphaSimR. A sweet corn breeding program was simulated accounting for recurrent selection in a 5-year pipeline with phenotypic selection, and random crosses. Three scenarios were simulated. In the first scenario 50 crosses were made in the crossing block at the beginning of each cycle (nCross). For the second (nCross*2) and third (nCross*5) scenarios, we increased the number of crosses two and five times, respectively. The lines represent the means of 50 replications over 30 years.

In addition, simulations became a quantitative genetics tool largely applied to breeding purposes, and implemented in several different crosses (*e.g.* maize (Cowling et al., 2020; Powell et al., 2020), soybean (Silva et al., 2021), rice (Sabadin et al., 2022), wheat (Tessema et al., 2020)), crop systems (Bančič et al., 2021), breeding schemes (Batista et al., 2021a; Covarrubias-Pazaran et al., 2022), and theoretical aspects of genetics and breeding (Batista et al., 2021b; Lara et al., 2022).

In addition to the several tools applied in the stage, the choice for the parents of the next generation is also a question. Recently, Labroo & Rutkoski, (2022) reinforced that to mix the

parents with new individuals in the beginning of each cycle decrease the long-term genetic gain of a breeding program. Further, early parental selection or rapid cycling seems to return more gain over time, especially in breeding programs where advanced tools such as genomic selection and doubled haploid are implemented (Gaynor et al., 2017; Powell et al., 2020; Volpato et al., 2021). However, for several crops, questions regarding the impact of early parental selection remains unsolved and further investigation crop-specific is needed.

The last part of a breeding is the later trial analyses. The number of genotypes decrease with selection over cycles while the number of environments increases in a way to access the potential candidates in the target population of environments. The genotype-by-environment interaction plays a key role in this stage. In maize breeding programs where the hybrid composition represents the cultivar to be released, the breeding process evaluates two heterotic groups, and the best cultivars (hybrids) are those with good combination with the other heterotic group. In this way, several combinations should be tested to choose the best hybrid between lines of distinct heterotic groups. However, the creation and assessment of all possible combinations between pairs of lines and many environments is cumbersome. For example, with 50 lines (*n*), the total number of crosses (*nc*) is represented by: $nc = (n*(n-1))/2$, which in this case is 1225 combinations. This number represents how difficult it is to assess all potential candidates in the field. Some alternatives to decrease the number of hybrids to test while increasing the accuracy of selection were proposed and implemented. One of them is the implementation of test crosses. In this case, the potential candidates are crossed with only a few individuals from the other heterotic group, individuals that we called testers. It decreases the number of crosses, even though it allows to measure how good the combinations of the lines from distinct heterotic groups are.

A technological development that impacts hybrids prediction was how cheap and dense in covering the genome molecular markers became. Thenceforth, in silico prediction of hybrid

performance may decrease the field labor and increase the certainty in assessing hybrid combinations. For a genomic prediction of hybrids, a small number of hybrids are planted in the field and have their phenotypic performance measured. After, the lines that potentially can became a hybrid parent are genotyped. With the molecular marker dataset, we combine both marker information (*e.g.*, markers from line A and line B) and generate a jointly marker panel for each potential hybrid. With phenotypic information and genotypic information of the hybrids, we create a training population. Altogether, we can predict the performance of non-tested hybrids, for the same environment or even for non-tested environment (Krause et al., 2020). So, we can advance to field stage only a hybrid set that represents the best candidates, once the identification of superior hybrids crosses early on hybrid breeding pipeline is important for the rapid development of commercial hybrids.

In summary, quantitative genetics tools have been largely applied in breeding programs and have evolved together. Several tools are daily used to inform breeding decisions and every new contribution in quantitative genetics field has a direct impact in breeding programs.

References

Akdemir, D., & Sánchez, J.I. (2016). Efficient breeding by genomic mating. *Frontiers in Genetics*, *7*, 1–12. https://doi.org/10.3389/fgene.2016.00210

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., & Teyssèdre, S. (2019)(a). Improving short-and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Frontiers in Genetics*, *10*, 1–15. https://doi.org/10.3389/fgene.2019.01006

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., & Lehermeier, C. (2019)(b). Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. *G3: Genes, Genomes, Genetics*, *9*, 1469–1479. https://doi.org/10.1534/g3.119.400129

Atanda, S.A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., Gowda, M., Dreher, K., Zhang, X., Prasanna, B.M., Tongoona, P., Danquah, E.Y., Olaoye, G., & Robbins, K.R. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theoretical and Applied Genetics*, *134*, 279–294. https://doi.org/10.1007/s00122-020-03696-9

Bančič, J., Werner, C.R., Gaynor, R.C., Gorjanc, G., Odeny, D.A., Ojulong, H.F., Dawson, I.K., Hoad, S.P., & Hickey, J.M. (2021). Modeling Illustrates That Genomic Selection Provides New Opportunities for Intercrop Breeding. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.605172

Batista, L.G., Gaynor, R.C., Margarido, G.R.A., Byrne, T., Amer, P., Gorjanc, G., & Hickey, J.M. (2021)(a). Long-term comparison between index selection and optimal independent culling in plant breeding programs with genomic prediction. *PLoS ONE*, *16*, 1–5. https://doi.org/10.1371/journal.pone.0235554

Batista, L.G., Mello, V.H., Souza, A.P., & Margarido, G.R.A. (2021)(b). Genomic prediction with allele dosage information in highly polyploid species. *Theoretical and Applied Climatology*, *135*, 723–739. https://doi.org/https://doi.org/10.1007/s00122-021-03994-w

Bernardo, R. (1994). Prediction of maize single-cross performance using RLFPs and information from related hybrids. *Crop Science, Madison*, *1*, 20–25

Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*, *125*, 375–385. https://doi.org/10.1038/s41437-020-0312-1

De Beukelaer, H., Badke, Y., Fack, V., & Meyer, G. De. (2017). Moving Beyond Managing Realized Genomic Relationship in Long-Term Genomic Selection. *Genetics*, *206*, 1127–1138. https://doi.org/10.1534/genetics.116.194449/-/DC1.1

Beyene, Y., Gowda, M., Pérez-Rodríguez, P., Olsen, M., Robbins, K.R., Burgueño, J., Prasanna, B.M., & Crossa, J. (2021). Application of Genomic Selection at the Early Stage of Breeding Pipeline in Tropical Maize. *Frontiers in Plant Science*, *12*, 1–11. https://doi.org/10.3389/fpls.2021.685488

Chen, G.K., Marjoram, P., & Wall, J.D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Research*, *19*, 136–142. https://doi.org/10.1101/gr.083634.108

Covarrubias-Pazaran, G., Gebeyehu, Z., Gemenet, D., Werner, C., Labroo, M., Sirak, S., Coaldrake, P., Rabbi, I., Kayondo, S.I., Parkes, E., Kanju, E., Mbanjo, E.G.N., Agbona, A., Kulakow, P., Quinn, M., & Debaene, J. (2022). Breeding Schemes: What Are They, How to Formalize Them, and How to Improve Them?. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.791859

Cowling, W.A., Gaynor, R.C., Antolín, R., Gorjanc, G., Edwards, S.M., Powell, O., & Hickey, J.M. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Scientific Reports*, *10*, 1–8. https://doi.org/10.1038/s41598-020-61031-0

Daetwyler, H.D., Hayden, M.J., Spangenberg, G.C., & Hayes, B.J. (2015). Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity.

Genetics, *200*, 1341–1348. https://doi.org/10.1534/genetics.115.178038

Eynard, S.E., Calus, M.P.L., Hulsegge, I., & Hiemstra, S. (2018). The impact of using old germplasm on genetic merit and diversity — A cattle breed case study 311–322. https://doi.org/10.1111/jbg.12333

Falconer, D.S., & Mackay, T.F.C. (1996). Introduction to quantitative genetics. *Harlow, Essex, UK: Longmans Green*, *3*, 280

Faux, A., Gorjanc, G., Gaynor, R.C., Battagin, M., Edwards, S.M., Wilson, D.L., Hearne, S.J., Gonen, S., & Hickey, J.M. (2016). AlphaSim: Software for Breeding Program Simulation. *The Plant Genome*, *9*, plantgenome2016-02. https://doi.org/10.3835/plantgenome2016.02.0013

Gaynor, R.C., Gorjanc, G., Bentley, A.R., Ober, E.S., Howell, P., Jackson, R., Mackay, I.J., & Hickey, J.M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, *57*, 2372–2386. https://doi.org/10.2135/cropsci2016.09.0742

Gaynor, R.C., Gorjanc, G., & Hickey, J.M. (2021). AlphaSimR: An R package for breeding program simulations. *G3: Genes, Genomes, Genetics*, *11*, 1–21. https://doi.org/10.1093/G3JOURNAL/JKAA017

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, *136*, 245–257. https://doi.org/10.1007/s10709-008-9308-0

Goiffon, M., Kusmec, A., Wang, L., Hu, G., & Schnable, P.S. (2017). Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection. *Genetics*, *206*, 1675–1682. https://doi.org/https://doi.org/10.1534/genetics.116.197103

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

Isik, F., Holland, J., & Maltecca, C. (2017). *Genetic Data Analysis for Plant and Animal Breeding*. Springer.

Jannink, J. (2010). Dynamics of long-term genomic selection. *Genetics, Selection and Evolution*, *42*, 1–11

Kemper, K.E., Bowman, P.J., Pryce, J.E., Hayes, B.J., & Goddard, M.E. (2012). Long-term selection strategies for complex traits using high-density genetic markers. *Journal of Dairy Science*, *95*, 4646–4656. https://doi.org/10.3168/jds.2011-5289

Krause, M.D., Dias, K.O.G., Pedroso Rigal dos Santos, J., Oliveira, A.A., Guimarães, L.J.M., Pastina, M.M., Margarido, G.R.A., & Garcia, A.A.F. (2020). Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Science*, *60*, 3049–3065. https://doi.org/10.1002/csc2.20253

Labroo, M.R., & Rutkoski, J.E. (2022). New cycle , same old mistakes ? Overlapping vs . discrete generations in long-term recurrent selection. *BMC Genomics*, *22*, 1:15. https://doi.org/10.1101/2021.10.12.464059

Labroo, M.R., Studer, A.J., & Rutkoski, J.E. (2021). Heterosis and hybrid crop breeding : A multidisciplinary review. *Frontiers in Genetics*, *12*, 1–19. https://doi.org/10.3389/fgene.2021.643761

Lara, L.A.C., Pocrnic, I., Oliveira, T.P., Gaynor, R.C., & Gorjanc, G. (2022). Temporal and genomic analysis of additive genetic variance in breeding programmes. *Heredity*, *128*, 21–32. https://doi.org/10.1038/s41437-021-00485-y

Lehermeier, C., de los Campos, G., Wimmer, V., & Schon, C.-C. (2017). Genomic variance estimates: With or without disequilibrium covariances?. *Journal of Animal Breeding and Genetics*, *134*, 232–241. https://doi.org/10.1111/jbg.12268

Lynch, M., & Walsh, B. (1988). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Meuwissen, T.H.. E., Hayes, B.J., & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829. https://doi.org/11290733

Meuwissen, T.H.E., & Sonesson, A.K. (1998). Maximizing the response of selection with a predefined rate of inbreeding: overlapping generations. *Journal of Animal Science*, *76*, 2575–2583

Moeinizade, S., Hu, G., Wang, L., & Schnable, P.S. (2019). Optimizing Selection and Mating in Genomic Selection with a Look-Ahead Approach: An Operations Research Framework. *G3 Genes|Genomes|Genetics*, *9*, 2123–2133. https://doi.org/10.1534/g3.118.200842

Moreira, F.F., Hearst, A.A., Cherkauer, K.A., & Rainey, K.M. (2019). Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. *Plant Methods*, *15*, 1–9. https://doi.org/10.1186/s13007-019-0519-4

Müller, D., Schopp, P., & Melchinger, A.E. (2018). Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection. *G3 (Bethesda, Md.)*, *8*, 1173–1181. https://doi.org/10.1534/g3.118.200091

Powell, O., Gaynor, R.C., Gorjanc, G., Werner, C.R., & Hickey, J.M. (2020). A two-part strategy using genomic selection in hybrid crop breeding programs. *bioRxiv*, 1–46. https://doi.org/10.1101/2020.05.24.113258

Sabadin, F., DoVale, J.C., Platten, J.D., & Fritsche-Neto, R. (2022). Optimizing self-pollinated crop breeding employing genomic selection: From schemes to updating training sets. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.935885

Santantonio, N., & Robbins, K. (2020). A hybrid optimal contribution approach to drive short-term gains while maintaining long-term sustainability in a modern plant breeding program. *bioRxiv*, . https://doi.org/10.1101/2020.01.08.899039

Schnell, F.W., & Utz, H.F. (1976). F1 Leistung und Elternwahl in der Zuchtung von Selbstbefruchtern. *Ber Arbeitstag Arbeitsgem Saatzuchtleiter*,

Silva, É.D.B., Xavier, A., & Faria, M.V. (2021). Impact of Genomic Prediction Model,

Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding. *Frontiers in Genetics*, *12*, 1–12. https://doi.org/10.3389/fgene.2021.637133

Sonesson, A.K., Woolliams, J.A., & Meuwissen, T.H.E. (2012). Genomic selection requires genomic control of inbreeding. *Genetics Selection Evolution*, *44*, 1–10

Tessema, B.B., Liu, H., Sørensen, A.C., Andersen, J.R., & Jensen, J. (2020). Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat. *Frontiers in Genetics*, *11*, 1–12. https://doi.org/10.3389/fgene.2020.578123

Villiers, K., Dinglasan, E., Hayes, B.J., & Voss-Fels, K.P. (2022). genomicSimulation: fast R functions for stochastic simulation of breeding programs. *G3: Genes, Genomes, Genetics*, *12*. https://doi.org/10.1093/g3journal/jkac216

Volpato, L., Bernardeli, A., & Gomez, F. (2021). Genomic selection with rapid cyclingcycling: Current insights and future prospects. *Crop Breeding and Applied Biotechnology*, *21*, 1–8. https://doi.org/10.1590/1984-70332021v21sa27

Zhong, S., & Jannink, J. (2007). Using Quantitative Trait Loci Results to Discriminate Among Crosses on the Basis of Their Progeny Mean and Variance. *Genetics*, *576*, 567–576. https://doi.org/10.1534/genetics.107.075358

Zingaretti, L.M., Gezan, S.A., Ferrão, L.F. V., Osorio, L.F., Monfort, A., Muñoz, P.R., Whitaker, V.M., & Pérez-Enciso, M. (2020). Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. *Frontiers in Plant Science*, *11*, 1–14. https://doi.org/10.3389/fpls.2020.00025

Objective

General

To apply and propose quantitative genetics tools to optimize different phases of a breeding program.

Specific

To compare multiple breeding pipeline strategies accounting for genomic selection and high-throughput phenotyping by means of hybrid gain and cost effectiveness.

To explore the potential of implementing genomic selection in a sweet corn breeding program through hybrid prediction in an across-year and across site framework.

To describe and implement an R package to cross prediction and optimization.

Chapter 1

Simulation based decision making and implementation of tools in hybrid crop breeding pipelines

Submitted to Crop Science

Simulation based decision making and implementation of tools in hybrid crop breeding pipelines

Marco Antônio Peixoto[1,2], Igor Ferreira Coelho[1,2], Kristen A. Leach[2], Leonardo Lopes Bhering[1], Márcio F. R. Resende Jr.[2,*]

[1] Laboratório de Biometria, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil.

[2] Sweet Corn Breeding and Genomics Lab, University of Florida, Gainesville, Florida, United States

* **Corresponding author:** mresende@ufl.edu

ABSTRACT

New technologies have been developed over the last few years aiming to support breeding pipeline optimization for long-term genetic gains. However, the implementation of these new tools and their impact on any breeding program's budget are not well studied. Here, we compare multiple breeding pipeline strategies accounting for genomic selection and high-throughput phenotyping by means of hybrid gain and cost effectiveness. We simulated a hybrid crop breeding program through coalescent theory. We compared two strategies for parental updates and four breeding pipelines: conventional breeding pipeline; conventional breeding pipeline with high-throughput phenotyping; conventional breeding pipeline with genomic selection; conventional breeding pipeline with genomic selection and high-throughput phenotyping. All analyses were implemented under three different levels of genotype-by-environment interaction (G×E) and two traits heritabilities (0.3 and 0.7). Overall, the results show that scenarios with early parental selection perform better than the others. In addition, the implementation of high-throughput phenotyping (HTP) delivered the highest hybrid gain in the long-term, whereas the implementation of genomic selection seems to be more cost-effective. We suggest, considering breeding programs with complex trait inheritance and accounting for higher levels of G×E, to invest in breeding pipelines accounting for genomic selection as a strategy to create and maintain long-term hybrid gain. Moreover, considering an unconstrained budget, the investment in both, genomic selection and HTP, represents the best strategy. Hence, these results provide strategies that may aid breeders in optimizing self-pollination breeding programs.

Keywords: Cost-effectiveness, G×E interaction, hybrid prediction, pipeline optimization. predictive accuracy.

INTRODUCTION

Plant breeding programs are continually developing and evaluating methods and tools to increase efficiency that can lead to increased genetic gains, reduction in operational costs, and reduction of the breeding cycle time. Two of such tools with applications in hybrid crop breeding are genomic selection, and high-throughput phenotyping (HTP) (Wang et al., 2020; Bernardo, 2021; Persa et al., 2021). While these tools have demonstrated benefits and are used routinely in large hybrid breeding programs, a breeder implementing them for the first time need to decide the stage of the breeding pipeline where the tool will be implemented, and the scale of implementation. This taken together with the specificity of a given breeding pipeline creates a difficult task in selecting which tools to prioritize, especially in smaller breeding programs. In this manuscript, we aim to explore these questions applied to hybrid breeding programs and using as a model the vegetable sweet corn.

Sweet corn is a valuable vegetable for the United States and Canada (Lertrat and Pulam, 2007; Hu et al., 2021). Like field corn, sweet corn breeding programs focus on the development of productive hybrids that can be targeted to the fresh or processing markets. However, sweet corn breeding differs from that of field corn as it does not have well defined heterotic groups (Jha et al., 2016), and public breeding programs typically focus on the development of inbred lines that can be released to become parents of commercial hybrids. Breeders will typically rely on recurrent phenotypic selection to develop the inbred lines, and select the parent for the next generation based on the performance of the inbred line per se combined to good performance of the testcross hybrids (high general combining ability or GCA) where these lines are crossed to commercial inbred (Zystro et al., 2021a). The highest performing inbreds selected over different seasons (often three) of commercial testcrosses and in several environments, represents the final product of the breeding program. Along with the lack of well-defined heterotic groups, a second common challenge of sweet corn breeding is

the impact of genotype-by-environment (G×E) interaction in the selection of the best potential lines. This can typically take place because seed production environments are generally different than target hybrid growing environments. Furthermore, breeders use off sites nurseries to speed up the breeding program and the development of cultivars to a target environment. The incorporation of genomic selection, and/or high throughput phenotyping represent one alternative that may enhance the genetic gain in public sweet corn breeding programs, accelerate the development and release of new inbreds, and help minimize the effect of G×E interaction.

One of the advantages of genomic selection (Meuwissen et al., 2001) is the possibility of circumventing issues related to material evaluation in different environments (Ferrão et al., 2020). For such, genomic models must be calibrated in environments considered to be the target, which can allow for selection in off-season environments, reducing the cycle time and increasing genetic gains. Recently, several sources of information (*i.e*., dominance effects, G×E interaction effects, environmental covariates, high-throughput phenotypic measures) have been added to improve prediction models (Jarquín et al., 2014; Pérez-Rodríguez et al., 2017; Mir et al., 2019; Ferrão et al., 2020). On the other hand, high throughput phenotyping is a broad term indicating recent developments to reduce phenotyping costs, increase phenotyping precision or enable the phenotyping of traits that previously couldn't be phenotyped (Neupane and Baysal-Gurel, 2021; Volpato et al., 2021; Yassue et al., 2021)). While the specific choice of phenotyping method is dependent on the crop and trait of interest, in this paper we are broadly referring as high-throughput phenotyping as an improve in the efficiency of phenotyping relative to a traditional standard. The ability to collect more data can then be used to expand breeding programs, reduce the evaluation costs of a breeding program, or as an input to genomic selection models (Sun et al., 2017).

While there has been extensive research exploring the individual impact of each of these tools (Feng et al., 2021; Fritsche-Neto et al., 2021; Marinho et al., 2022), the implementation of them in a way to maximize the selection gains in a targeted breeding program must be tested to choose the best strategy for implementation. In this context, tools that allow us to ask multiple questions, test several different scenarios and then choose the one that best fits into a breeding program are helpful. Herein, we utilize the context of a public sweet corn breeding program to mimic a standard hybrid crop breeding program. We used stochastic simulation implemented in AlphaSimR (Gaynor et al. 2021) to evaluate the use of genomic selection and phenomics in a breeding pipeline aiming for long-term gain in hybrid production.

MATERIALS AND METHODS

Stochastic simulations were used to compare the implementation of HTP and genomic selection into an allogamous breeding program. We also tested different strategies for choosing parental lines that would compose the crossing block at the beginning of each cycle. These scenarios were compared using the mean of 50 replicates, each replicate consisting of: i. a burn-in phase, used in all scenarios as common starting point and representing the 20 years of breeding preceding the decision to incorporate genomic selection and HTP, and, ii. an advanced phase of evaluation comprising of 30 years of the continuous testing of each breeding scenario. Four scenarios were considered in the first phase on this analysis. After which, the better scenarios were used as a baseline model for the implementation of HTP.

Simulation and genetic structure of the founder population

The genome was simulated as a maize genome, accounting for the parameters set for the ‘MAIZE’ option in the package. Each genome sequence consisted of 10 chromosomes. The genetic length of each chromosome was set to 2 Morgans, and the physical length was set

to $2x10^8$ base pairs. The recombination and the mutation rates were $1.25x10^{-8}$ and $2.5x10^{-8}$ per base pair, respectively (Hickey et al., 2014; Powell et al., 2020).

The founder genotypes were generated using Markovian Coalescent Simulator (Chen et al., 2009). Two hundred genotypes were generated, in each program replicate. A gamma model was used to simulated the genetic recombination and meiosis, according to McPeek & Speed (1995) for a diploid species, which was deployed in AlphaSimR (Gaynor et al., 2021). A set of 300 biallelic quantitative trait nucleotides (QTN) and 3,000 single nucleotide polymorphisms (SNP) per chromosome were randomly simulated along each chromosome. In total, each trait structure was represented by 3,000 QTN and 30,000 SNP markers.

Simulation of genetic values

Genetic values were simulated considering the structure of two independent traits:

(i) A trait with a broad sense heritability of 0.3, to represent a quantitative trait with a large number of small effect genes, such as yield, and that is significantly impacted by the environment.

(ii) A trait with a broad sense heritability of 0.7 to represent a trait with a large number of small effect genes small with little impact by environment variations.

We simulated both traits using three effects at each QTN: additive, dominance, and genotype-by-environment (ADG trait, as referred to in the package) as described on classic quantitative genetics models.

Initial values of the additive effect (α) for each QTN were sampled from a normal distribution and were scaled to $\sigma_a^2 = 30$ and mean additive effect of 150 for both traits. In addition, the dominance effect (*d*) at a particular locus was calculated considering the degree of dominance (δ) and absolute value of its additive effect (α), as the following expression:

$$d = \delta|\alpha|$$

In the simulations, a dominance degree of 0 represents no dominance, in other words, an additive model. A dominance degree of 1 corresponds to a model with complete dominance. Dominance degrees values between 0 and 1 correspond to partial dominance, and values above 1 correspond to over-dominance. Degrees of dominance values were then sampled from a normal distribution, with an average degree of dominance equal to 0.92 and $\hat{\sigma}_{\delta}^{2}$ = 0.3, representing a trait with partial dominance. These mean and variance were picked to follow the historical levels that roughly represents the heterosis in field corn (Powell et al., 2020). The values were then used together with the additive effects to calculate the dominance effects at each locus (QTN) as follows:

$$d_i = \begin{cases} 0, & if\ QTN\ is\ homozygous, \\ \delta_i|\alpha_i|, & if\ QTN\ is\ heterozygous \end{cases}$$

A scale process was then performed to the dominance effects as described before for the additive effects.

In the end, 3 different levels of G×E interaction (σ_{gxe}^{2}) were used: 0, 1, and 4 times the additive variance (σ_{gxe}^{2} = 0, 30, and 120, respectively). The sweet corn breeding pipeline was simulated considering the same planting seasons utilized by the University of Florida sweet corn breeding program. Hence, three different seasons are considered: winter (January to April), spring (April to July), and fall (October to December). The spring season was selected for new crosses to be created and is not part of the cycling process in the pipeline. Phenotypic evaluation of the materials is completed in the winter season, in the target environment.

Aiming to set a different G×E interaction intensity for each environment (i.e. target environment – winter – and off season environment – fall), we utilized a p-value to calculate the environmental covariate value for any level of G×E interaction (Gaynor et al., 2021). This

choice of p-value is used by AlphaSimR to define the amount (higher or lower) of G×E interaction. It is worth mentioning that a p-value of 0.5 represents an environmental covariate equal to zero and is equivalent to the target environment (Gaynor et al. 2021), whereas numbers closer to 0.5 represents lower values of G×E interaction and, as much closer to the extremes we sample this covariate, meaning close to 0 or to 1, it returns higher G×E interaction values.

In our simulation, the winter season (target environment, lower G×E interaction), we considered a p-value of 0.5 plus a random deviation sampled from a normal distribution with mean 0 and variance of 0.03, whereas for the fall season (off-season environment, higher G×E interaction) the p-value was considered as 0.1 or 0.9 (randomly set for each repetition) plus a random deviation sampled from a normal distribution with mean 0 and variance of 0.03. We implemented a higher G×E interaction for the fall season once the target environment was presented to the materials in the winter season. Therefore, any phenotypic variation caused by this non-target environment should be considered as noise.

The phenotypic value of each genotype was calculated by adding a random sampled error for the genetic values for both traits. The random error was sampled from a normal distribution with mean zero and residual variance (σ_e^2), based on the trait's broad sense heritability (H^2 = 0.3 and 0.7). The heritability of each trait was calculated as follows (Bernardo, 2010):

$$H^2 = \frac{\sigma_a^2 + \sigma_d^2}{\sigma_a^2 + \sigma_d^2 + \sigma_{\text{gxe}}^2 + {\sigma_e^2}/{r}},$$

where *r* represents the number of repetitions.

Burn-in phase

The burn-in phase was performed through 20 years of phenotypic selection, simulating a conventional phenotypic selection program, and serving as a starting point to the

comparison among the scenarios. The phenotypic selection (conventional breeding program) was considered as follows (**Figure 1**):

- Year 1. A crossing block with 50 parents generated 50 families through 50 random crosses. The parents for each cross were randomly assigned, being the participation of each parent a matter of randomization. Populations advancement by selfing the lines and selection within families.
- Year 2. Population advancement by selfing the lines and selection of the top two individuals within each family.
- Year 3 . First testcross (TC1) for selection based on general combining ability (GCA) of each line in one environment and using two testers. The testers here used come from the private breeding program that runs in parallel to our simulated breeding program. Every year, the top two lines in genetic merit in the private program is chosen and used as a tester against the lines from the main breeding program. Population advancement, selection within families, and selection of the best individuals based on the GCA of each line in TC1. Also, the best 50 individuals, based on GCA, will compose the crossing block in the next cycle (hereafter, we called this parental update strategy of the TC1_par).
- Year 4. Second testcross (TC2) in three environments and three testers (same as described above, where the top three testers is chosen), population advancement, and individual selection based on the GCA of each line.
- Year 5. Third testcross (TC3) in 19 environments using five testers (same as described above, where the top three testers is chosen), population advancement, and individual selection of the best lines based on GCA. Release the two best lines based on hybrid production.

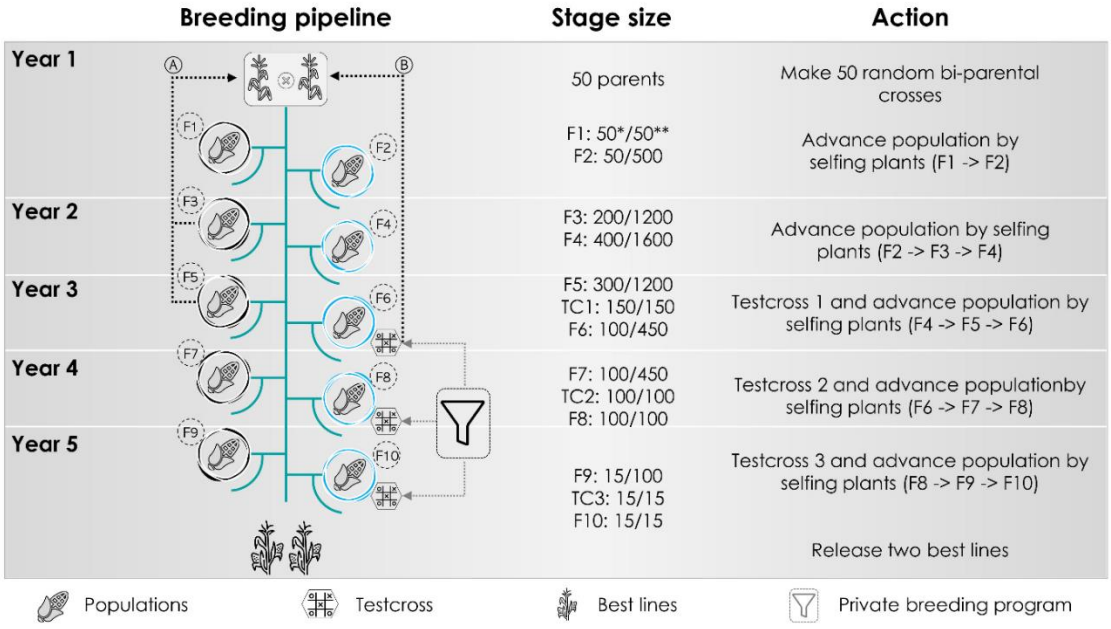


Figure 1 - The sweet corn breeding program pipeline. *Number of families. **Number of individuals. A: indicates the strategy with the selection of parents from the individuals with the best performance in F3 and F5 populations. B: indicates the strategy with the selection of parents from the individuals with the best performance in the first testcross. The colored circles around each ear indicated where the plants of that generation are grown, where blue represents the winter season (target environment) and black indicates the fall season (off-season environment).

A seven-year cycle long breeding program was also simulated in parallel with the main breeding program. It was built from the same founder population and, by using the argument 'split', we set a split equals to 30 between the main breeding program and the parallel one, representing a historic population separation in terms of generations ago to create distinct genetic groups. In our case, one group represented the sweet corn breeding program (were all the scenarios was applied) and the second one, the private breeding program. Here, we used this second breeding program (that comes from the other distinct group) to generate lines, latter used as testers at the principal breeding program. This strategy mimics the use of private testers by the program to produce lines with good GCA to be commercialized. We considered 100 individuals as parents and 200 crosses in the crossing block per cycle for the private

program. The best lines through the cycle were select based on phenotypic selection. Therefore, in the last year of the breeding cycle (year 7) the best five lines based on the genetic value were considered as the testers for the TC1 (two testers), TC2 (three testers), and TC3 (five testers) in the sweet corn program. In addition, both programs ran in parallel to make them equivalent in terms of genetic advance.

Future breeding phase

The future breeding phase was used to compare the implementation of genomic selection strategies, parental update strategies, and the implementation of HTP into the sweet corn breeding program. For such, 30 years of future breeding were considered in two sets of simulations as described in sequence.

Long-term breeding program considering genomic selection

The first one was to evaluate four scenarios in the sweet corn pipeline, developed from two parent selection strategies within the conventional breeding program with phenotypic selection (used as benchmark) and within the conventional breeding program with the deployment of genomic selection.

Baseline breeding program

The first scenario simulated was the conventional program (CONV) and is a continuation of the burn-in phase. In the second scenario (CONVe), we simulated what would happen to gain if parents were selected at earlier stages instead of waiting until the completion of the cycle for new parents to be selected. Parents of the crossing block in the CONVe strategy were replaced by 50 new individuals from the F3 and F5 stages. These selections are made based on the phenotypic performance of the F3s and F5s in the target environment.

Genomic selection strategy

Genomic selection deployment was used aiming at selecting superior individuals in F3 and F5 populations, which circumvented the higher level of G×E interaction present in the off-season environment. In this strategy, the F1 and F2 populations were both planted in the winter season and phenotypic measures were taken to calibrate the genomic selection model used in the selection of individuals in F3 and F5 populations. We used the Ridge Regression BLUP (RR-BLUP) model to obtain the estimated breeding values. In the third scenario (GS), the individuals selected at the first testcross, based on higher GCA, were them ranked on phenotypic values and the top 50 lines were selected to compose the crossing block in the next, replacing all old parents in the crossing block. In the fourth scenario (GSe), the parental update was made considering the 50 top individuals from F3-F5 populations, where the individuals were ranked based on the estimated breeding values. It is worth to mentioning that we did not predict hybrid crosses, aided by genomic selection, as standard in hybrid program, such as field corn and sweet corn (Ferrão et al., 2020; Zystro et al., 2021a; b). The reason is that because we assumed that the commercial testers were not available for genotyping, a generalization of the public sweet corn program here simulated, once the testers come from the private sector. Then, we just deployed GS in the beginning of the program (selection at off-season environment for populations F3 and F5).

Employing high throughput phenotyping into a breeding program

To examine the effect of implementing high throughput phenotyping (HTP) into the breeding program, we compared four scenarios (Table 1). As a benchmark for comparison, we used the scenarios CONVe and GSe. The first scenario with HTP (CONVe_HTP) represents the CONVe scenario where we implemented HTP for the evaluation and selection of phenotypes in the field. The genotypes were ranked, and selections were made using the output from the HTP assessments. The second scenario with HTP was the genomic selection scenario (GSe), and we simulated the genotype phenotypic assessment through HTP

(GSe_HTP), which includes the model calibration (F1 and F2 populations) and genotypes selection in all phases, but F3 and F5, as we described in genomic selection scenarios.

Table 1 - Summary of number of plots and annual total costs of the conventional breeding program, the implementation of genomic selection and the implementation of high throughput phenotyping into the sweet corn breeding pipeline. The costs to implement each scenario was also assumed to enable the comparison between scenarios. The per plot field cost was assumed to be $12 for the inbred advancement fields and $8 per plot for the testcross hybrid evaluation fields. Genotyping costs were assumed to be $10 per sample, whereas the cost for phenotypic evaluation was assumed to be $1 or $4 per plot with the use of HTP or conventional phenotyping, respectively. The total costs per scenario are presented in the **Table 1 and Supplementary material - Tables S1-S3.**

Year	Scenario							
	CONVe		CONVe_HTP		GSe		GSe_HTP	
	Plots	Cost ($)	Plots	Cost ($)	Plots	Cost ($)	Plots	Cost ($)
1	200	2800	300	3800	300	7900	500	9900
2	1000	15200	1800	23200	1000	23200	1800	31200
3	1320	17040	1620	17640	1320	26040	1620	26640
4	960	11120	1460	12120	960	11120	1460	12120
5	475	4740	775	5340	475	4740	775	5340
Total	3955	50900	5955	62100	4055	73000	6155	85200

CONV: conventional breeding program with parental update using individuals with best performance in the first testcross. CONVe: conventional breeding program with parental update using individuals with best performance in F3 and F5. CONVe_HTP: conventional breeding program with the imputation of high throughput phenotyping (HTP). GS: conventional breeding program with implementation of genomic selection and parental update considering the individuals with best performance in the first testcross. GSe: conventional breeding program with implementation of genomic selection using F1 and F2 populations as training sets and parental update considering the individuals with best performance in F3 and F5 populations. GSe_HTP: conventional breeding program with implementation of genomic selection using F1 and F2 populations as training sets and parental update considering the individuals with best performance in F3 and F5 populations and phenotypic assessment made though HTP.

In a way to simulate the application of HTP in CONVe_HTP and GSe_HTP scenarios, we assumed that HTP would be used to double the number of repetitions in the winter season (target environment) and increased the trait broad sense heritability, from 0.3 to 0.4 and from

0.7 to 0.8, as a proxy of the implementation of HTP. Those assumptions were made to represent strategies with a more precise phenotype, which may not always be the case with HTP but is expected to be the case when we are phenotyping, for example, ear traits using machine vision instead of individual ratings (Gonzalez et. al. 2022). To test the impact of those assumptions, we completed additional simulations implementing only one change at a time, *i.e*. only increasing the broad sense heritability or only increasing the number of repetitions.

In addition, we simulated the four scenarios described before (CONVe, CONVe_HTP, GSe, and GSe_HTP) but in a breeding program that was assumed to be 10 times larger. For such, we simulated 500 crosses in the crossing block, with 100 parents for each generation, and the individual's selection and genetic assessment were increased in 10 times throughout the whole pipeline. Here, our goal was to test the sensitivity of the outcome to breeding programs of different sizes, resembling larger breeding programs.

Performance evaluation

To analyze each scenario and compare their performance, we calculated: (i) genetic gain, (ii) genetic variance, (iii) hybrid mean at the TC3, and (iv) prediction accuracy. For the genetic gain and the genetic variance, we used the genetic mean and genetic variance of the parents for each cycle, respectively. Standard error (*se*) was estimated for the predictions, as follows:

$$se = \frac{standard\ deviation}{\sqrt{n}}$$

where n represents the number of repetitions for each scenario.

To calculate the prediction accuracy, we used the correlation between the genetic values and the estimated breeding values at F3 and F5 populations (scenarios with genomic selection). Therefore, scenarios that did not accounted for genomic selection, the accuracy

was represented by the correlation between the genetic value and the phenotypic value of the target population (F3 and F5).

The costs to implement each scenario was also assumed to enable the comparison between scenarios. The per plot field cost was assumed to be $12 for the inbred advancement fields and $8 per plot for the testcross hybrid evaluation fields. Genotyping costs were assumed to be $10 per sample, whereas the cost for phenotypic evaluation was assumed to be $1 or $4 per plot with the use of HTP or conventional phenotyping, respectively. The total costs per scenario are presented in the **Table 1 and Supplementary material - Tables S1-S3.**

Software implementation

All analyses were carried out in the R package AlphaSimR (Gaynor et al., 2021). We used R (version 4.1.2) for post-processing the outputs from the analyses and the package plyr (version 1.8.5) to generate the graphs. All codes used to implement the simulations and to creates the visual representation of the graphs are available at https://github.com/Resende-Lab/Peixoto_2023_Simulation_based_decision.

RESULTS

Long-term breeding program with genomic selection

The long-term genetic mean performance of the scenarios was impacted by both, the strategy used for selection and the amount of G×E interaction at each scenario (Figure 2 and Table 2). Scenarios using early parental recycling (both strategies, CONVe and GSe) produced greater and faster genetic gains than scenarios with the strategy where parental selection took place after TC1 evaluation (both strategies, CONV and GS). The performance of CONVe scenario in the absence of G×E interaction was greater than the CONV and GS. However, this difference in performance decreased with the presence of G×E interaction (G×E 30). The CONV and GS scenarios (strategy where the parents are individuals with

higher GCA, tested against lines from the private breeding program in the first testcross) only slightly outperformed the CONVe in the highest level of G×E interaction (G×E 120) for the trait with a heritability of 0.3.

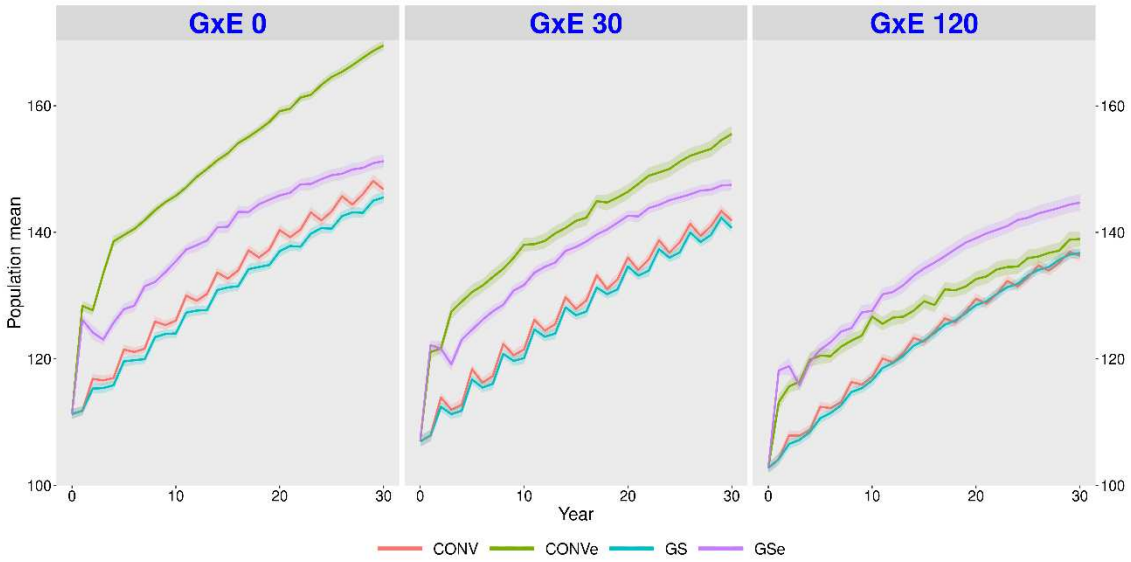


Figure 2 - Genetic gain overtime for the four simulated scenarios for a trait with heritability of 0.3. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 1 (G×E 30), and 4 (G×E 120) times the additive effect. The gain is plotted as the mean of the parents for each cycle with 50 replicates. The standard error of the mean is represented by the shading around the line. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Table 2 - Genetic gain per levels of G×E interaction (G×E 0, G×E 30, and G×E 120) of four scenarios considering different strategy for the update of parents for the two levels of heritability.

Scenario	Total plots	Total costs ($)	Genetic gain - (G×E)			Relative genetic gain (G×E)		
			0	30	120	0	30	120
Heritability equals to 0.3								
CONV	3955	50900	35.50	34.88	33.55	1	1	1
CONVe	3955	50900	58.26	48.60	36.21	1.64	1.39	1.08
GS	4055	73000	34.24	33.75	34.00	0.96	0.97	1.01
GSe	4055	73000	39.97	40.56	41.96	1.13	1.16	1.25
Heritability equals to 0.7								
CONV	3955	50900	38.02	37.84	37.21	1	1	1
CONVe	3955	50900	61.17	48.18	37.77	1.61	1.27	1.02
GS	4055	73000	35.56	35.63	37.08	0.94	0.94	1.00
GSe	4055	73000	43.21	45.21	43.22	1.14	1.19	1.16

CONV: conventional breeding program with parental update made by the lines that best perform in testcross 1. CONVe: conventional breeding program with parental update made by the lines that best perform at F3 and F5 populations. GS: breeding program with genomic selection with parental update made by the lines that best perform in testcross 1. GSe: breeding program with genomic selection with parental update made by the lines that best perform at F3 and F5 populations.

Overall, GS scenario presented lower levels of performance compared to CONV scenario. However, with a G×E of 120, the two methods presented similar performance. The increase of G×E interaction effect highly impacted CONVe, decreasing the gain by 38% (G×E 0 against G×E 120). Alternatively, GSe displayed gains that slightly increases from 39.97 to 41.96 when comparing no G×E interaction to the highest level of G×E. For the trait simulated considering 0.7 of heritability, we found results comparable with the results reported for the trait with 0.3 of heritability. Hence, the long-term behavior of CONVe and GSe (early parental selection) outperforms the scenarios CONV and GS (later parental selection) and it also was impact by the G×E interaction (Table 2 and Supplementary material – Figure S1).

Overall, for both trait heritabilities, the genomic prediction accuracy of scenario with early parental selection (GSe), was superior to late selection (GS). However, the accuracy decreased as the G×E level increased (Supplementary material – Figures S2 and S3). Overall, the accuracy of models that incorporated genomic selection (GS and GSe) did not presented a pronounced decrease in the presence of G×E interaction. Also, for all scenarios 30 years in the future, the genetic mean seems to continue to increase without the indication of a plateau, suggesting in this population, and under these selection schemes, there is potential to achieve even higher gains in upcoming years.

When evaluating the genetic variance across the 30-year simulation, it can be observed that each method presents similar profile for all levels of G×E interaction (Figure 3 and Supplementary material – Figure S4). Over time, and as expected, there is a long-term reduction in genetic variance for both levels of trait heritability. The early strategies (CONVe and GSe) presented higher levels of genetic variance regardless the G×E interaction. The GSe returned a quicker reduction in genetic variance when compared to CONVe under the three levels of G×E interaction. The GSe also presented the lowest values at the end of 30 years (*e.g*. G×E 0: 1.92 (GSe), 4.90 (GS), 27.03 (CONVe), 5.40 (CONV); G×E 120: 1.78 (GSe), 5.47 (GS), 14.90 (CONVe), 5.45 (CONV)). It can also be observed that in the absence of G×E interaction, the CONVe scenario led to the smallest reduction in the genetic variance among all scenarios tested. The scenarios CONV and GS presented a similar performance in genetic variance, regardless of the level of G×E interaction. Similar results were also found for the trait with 0.7 of heritability, whereas the range of genetic variance was smaller compared with the trait with 0.3 of heritability.

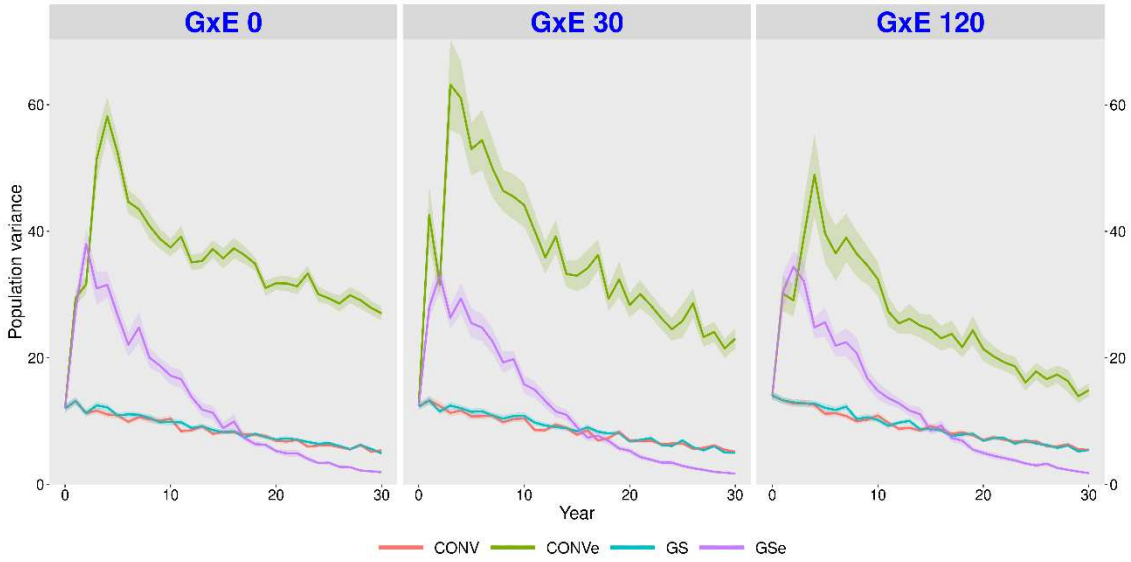


Figure 3 - Genetic variance overtime for the four simulated scenarios for a trait with a broad sense heritability of 0.3. In this graph, the variance was measured on the parents. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 1 (G×E 30), and 4 (G×E 120) times the additive effect. The gain is plotted as a mean of the parents for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the genetic mean for the 50 replicates and the shading represents the standard error. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Deployment of genomic selection versus high throughput phenotyping

After the evaluation of selection strategies, we implemented scenarios with HTP in a way to compare them. In general, implementation of HTP into a breeding program increases the hybrid performance, compared to scenarios without HTP, *i.e*. GSe_HTP vs. GSe and CONVe_HTP vs. CONVe (Figure 4, Table 3, and Supplementary material – Figure S5 and Tables S4 and S5). With a heritability of 0.3, the GSe_HTP increases the final hybrid gain compared with the GSe by 7% and 4%, for G×E 0 and G×E 120, respectively. The final

performance of CONVe_HTP increased by 11% when compared to CONVe, for both G×E 0 and G×E 120.

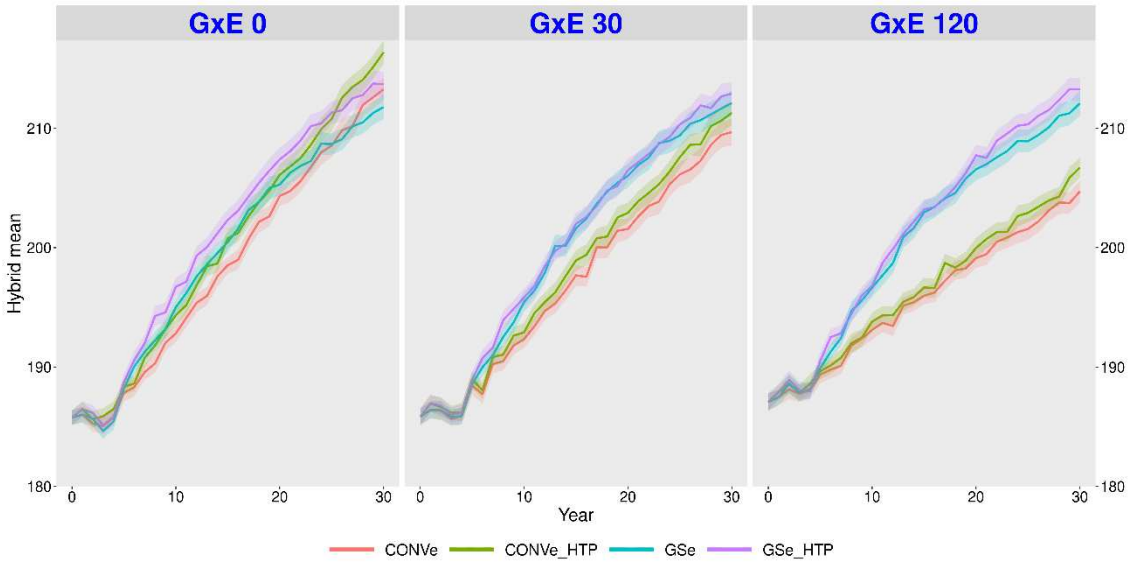


Figure 4 - Hybrid gain overtime for the four simulated scenarios for a trait with a broad sense heritability of 0.3. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 1 (G×E 30), and 4 (G×E 120) times the additive effect. The hybrid gain is plotted as a mean of the hybrids for each cycle. Each line represents the hybrid gain for the 50 replicates and the shading represents the standard error. CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: conventional breeding program with genomic selection. GSe_HTP: conventional breeding program with genomic selection and high throughput phenotyping.

Table 3 - Hybrid gains in unit, costs to implement each scenario in unit, and costs divided by hybrid gains. All reports are made under three different level of G×E interaction (G×E 0, G×E 30, and G×E 120) for the four scenarios with and without the implementation of highthroughput phenotyping. In this strategy we doubled the number of repetitions and increase the heritability in trait assessment. The results of hybrid gains represented the final value of each scenario (year 30), considering the means of 50 repetitions of each pipeline, for two trait heritabilities (0.3 and 0.7).

Scenario/G×E	Hybrid gain			Cost	Cost by gain		
	0	30	120		0	30	120
Heritability equals to 0.3							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.11	1.07	1.11	1.22	1.10	1.14	1.10
GSe	0.95	1.10	1.42	1.43	1.52	1.30	1.01
GSe_HTP	1.02	1.14	1.48	1.67	1.65	1.47	1.13
Heritability equals to 0.7							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.06	1.02	1.08	1.22	1.15	1.20	1.13
GSe	0.90	1.07	1.45	1.43	1.60	1.35	0.99
GSe_HTP	0.92	1.13	1.51	1.67	1.81	1.47	1.11

CONVe: conventional breeding program. CONVe_HTP: conventional breeding and high throughput phenotyping. GSe: genomic selection breeding program. GSe_HTP: genomic selection breeding program with high throughput phenotyping.

The results on hybrid performance demonstrated that the scenarios display a similar performance across years. For instance, in the absence of G×E interaction, CONVe_HTP presented the best final performance. Alternatively, when G×E interaction is present, scenarios with genomic selection were better. For the trait with heritability of 0.7, in the absence of G×E interaction, scenarios using only phenotypic selection (CONVe and CONVe_HTP) were superior (Table 2) and in the presence of G×E interaction scenarios with genomic selection were better.

The level of G×E interaction and whether HTP was used in the program had an impact on the costs of gain. When examining HTP implantation within a program, we determined there was an increase in cost when compared to those programs without HTP and that cost did not necessarily mean greater increases in gain when compared to the inclusion of GS methods. The cost per hybrid gain decreased with GS (GSe and GSe_HTP) when we increased the G×E interaction effect. This pattern was even more pronounced in a trait with a heritability of 0.7. In general, as G×E interaction increases, the cost per unit of gain is lowered to implement genomic selection (GSe scenario) and HTP (CONVe_HTP and GSe_HTP) in a breeding program.

When we consider a bigger set of simulations (Supplementary material – Table S6), similar patterns were found. However, programs with a G×E interaction of 120, scenarios that included genomic selection (GSe and GSe_HTP) presented a lower cost per gain, even lower than the conventional scenarios (CONVe and CONVe_HTP), for both trait heritabilities. Interestingly, the final gains in the larger breeding programs increased compared to the final gains in the small breeding program (*i.e.* GSe_HTP: 1.18, 1.40, 1.86 (large) vs. 0.95, 1.10, 1.42 (small) for G×E 0, 30, and 120, respectively, with a heritability of 0.3 (Supplementary material – Table S6). In the absence of a G×E interaction, higher prediction accuracies were found in scenarios with only phenotypic selection (CONVe and CONVe_HTP), for both heritabilities (Figure 5 and Supplementary material – Figure S6). The prediction accuracy of the scenarios with 0.7 presented higher values compare to scenarios with 0.3 of trait heritability. The trend in selective accuracy also demonstrates that the G×E interaction plays an important role in the performance of non-GS scenarios, decreasing the model reliability, whereas it did not imply in significant changes in accuracy performance of GS scenarios, in across different levels of G×E.

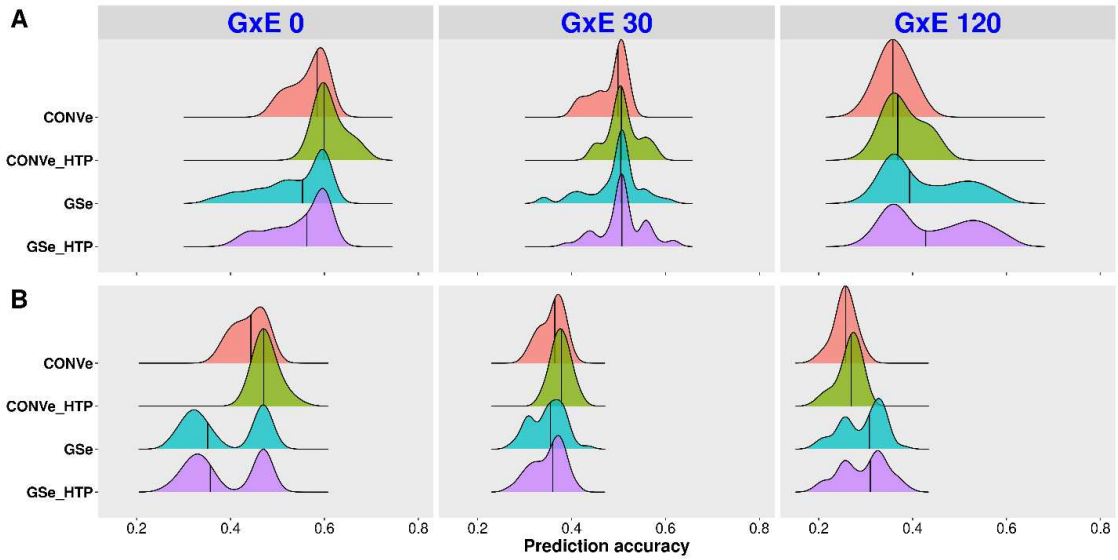


Figure 5 - Prediction accuracy of each scenario at F3 and F5, for the 30 years of breeding program, four scenarios, and a trait heritability of 0.3. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The distribution within each panel represents the four simulated scenarios for 50 replicates. CONVe: conventional breeding program. CONVe_HTP: conventional breeding and high throughput phenotyping. GSe: genomic selection breeding program. GSe_HTP: genomic selection breeding program with high throughput phenotyping.

The two scenarios with HTP displayed a higher selective accuracy compared to scenarios without the HTP (*i.e.* GSe_HTP versus GSe and CONVe_HTP versus CONVe) (Table 3). We can observe that the predictive accuracy of these scenarios tends to decrease over the years. In addition, for all scenarios, the accuracy for the selection at F5 was lower than the accuracy for F3 population for scenarios for both heritability levels. The large set of simulations provided similar results with those reported before (Supplementary material – Figure S7 and S8, Table S6). The results with the trait heritability of 0.3 emphasized even more the prediction accuracy of the scenarios with genomic selection compared to scenarios without in higher G×E interaction. This was also seen in the accuracy prediction from scenarios with a trait heritability of 0.7, but the presence of G×E interaction had less impact

(reduce less the value) of the prediction accuracy compared with the trait with 0.3 of heritability.

DISCUSSION

Modern breeding programs require efficient breeding approaches that utilize the advancements in breeding to achieve higher performance in terms of genetic gains. The evaluation of breeding schemes through simulation has opened a new window of possibilities to examine these advancements and inferences to be made about the right strategy to implement (Bančič et al., 2021; Silva et al., 2021; Lara et al., 2022). For instance, our results demonstrate that scenarios that incorporate genomic selection are superior for dealing with situations where the G×E interaction is a factor that introduces noise in the selection process. While this result in itself is not surprising, the magnitude of the difference and the impact will be specific to the training population size, magnitude of G×E and breeding strategy. Breeding simulations can then provide a more complete evaluation of such impact and the specific details where they are maximized. For example, in our evaluated scenarios, the implementation of genomic selection was only effective when paired with early parental selection. In the long-term, early parental selection was found to be more valuable than scenarios that considered parental selection based on general combining ability (CONV and GS). Historically, breeders have avoided early recycling of parents due to the confounded impact that dominance can have in the phenotypic performance of early inbred lines. Here, we show promising results for a trait that is controlled by partial dominance, even when we apply phenotypic selection.

Another factor that contributes with genetic gain for each scenario was the prediction accuracy in the selection for each population. It varied among the scenarios once we had different tools being applied in the selection (i.e. phenotypic selection, genomic selection, etc.). In general, genomic selection may boost breeding programs by increasing both, the

number of cycles per unit of time and the prediction accuracy (Gaynor et al., 2017). As in our case, the breeding program are comparable in matter of time, genomic selection impacted by improving the accuracy of the individual's selection in each cycle. In the absence of G×E interaction, CONVe scenario results in higher prediction accuracy values compared to all other scenarios. This higher accuracy would result in larger gains over the long term. However, when increasing the noise by adding G×E interaction, the reliability of scenarios with genomic selection remains high, enabling these methods to circumvent this issue and maintain the long-term gain, regardless of trait heritability.

In breeding programs, it is not only important to achieve faster genetic gain while maintaining those increased levels, but to track the loss of diversity as it can be indicative of the potential a breeding program possesses to keep improving genetic gain, especially when GS drives the increasing of inbreeding rates (Obšteter et al., 2019; Meuwissen et al., 2020; Pocrnic et al., 2023). In the long-term, GS scenarios returned higher gains compared with phenotypic program, as largely registered in different crops that tracked the long-term GS impacts (Allier et al., 2020; Powell et al., 2020; Lubanga et al., 2022). However, the steady decrease in genetic diversity in scenarios with GS highlight the needed for the incorporation of tools that constrain inbreeding levels in lower values, once it limits long-term genetic gains (Cobb et al., 2019). Strategies that balances genetic gain and genetic loss are indicated, such as optimal contribution selection (Meuwissen and Sonesson, 1998), genomic optimum contribution selection (Sonesson et al., 2012), optimal cross selection (Gorjanc et al., 2018), and usefulness criterion for parental contribution (Allier et al., 2019b).

In breeding programs, it is not only important to achieve faster genetic gain while maintaining those increased levels, but to track the loss of diversity as it can be indicative of the potential a breeding program possesses to keep improving genetic gain. Bottom line, our

results clearly advocate that scenario with genomic selection and early parental selection can help to circumvent problems related to G×E interactions, especially in breeding programs that use non-target environments for selection. On the other hand, the GS scenario without early selection also led to the lowest performance in the pipeline. It is possible that GS by itself will be better than conventional programs in cases with even higher levels of G×E interaction, or cases where the genomic models are more accurate. Many different approaches to improve prediction accuracy have been previously shown, such as including non-additive effects in the model (Dias et al., 2018; Oliveira et al., 2020; Zystro et al., 2021b), the relationship in-between the individuals from the training population and the target population (Li et al., 2021), among others.

In this study, while we referred as one of the tested scenarios as the use of high throughput phenotyping, what we evaluated was the improvement in the estimate of trait heritability, and the increase of the number of replicates. From that point of view, it makes sense that HTP presented better results, since more and more accurate data is being simulated. As mentioned by Wang et al. (2018) sensing technologies (e.g. ultrasonic sensor, LiDAR, digital cameras) return the most accurate traits estimations, when compared with manual measurements. Similarly, a new phenotyping technology is not technically needed in order for the breeder to increase replication size or the number of genotypes examined. However, in the conventional pipeline, the breeder is typically constrained by resources (time, space, funds) and we then assumed here that new HTP technologies were needed in a breeding program to reduce labor costs, as also discussed by several authors (Watanabe et al., 2017; Crossa et al., 2021; Persa et al., 2021). From a broader point of view, the study highlights the importance of maximizing the precision of the phenotyping while also maximizing the number of genotypes. While it wasn't the topic of the study, previous work has shown similar results when comparing

advanced experimental designs such as augmented p-rep designs or controlling for spatial variability in the field (Bernardeli et al., 2021; Coelho et al., 2021).

The impacts of HTP seemed to be higher in scenarios with only phenotypic selection (CONVe/CONVe-HTP), in terms of hybrid gain. Therefore, scenarios that accounted for genomic selection also aided from the implementation of HTP, especially in cases where G×E interaction is higher. As an outcome, those results states that it is important to consider HTP for breeding programs such our pipeline, with target and off-season environments. Considering the trait with higher heritability and scenarios with high G×E interaction, the impact of better phenotyping was notable.

Nowadays, phenotyping is considered as one of the factors limiting the accuracy of the genomic selection model. If one can improve the information used to calibrate the model, one can then improve of the prediction accuracy of the model (Montesinos-López et al., 2017; Atkinson et al., 2018). To improve the phenotypic accuracy, one needs to consider cost and the amount of time needed to infer the phenotype (Crossa et al., 2021), both being easily manageable by the implementation of a HTP platform. These results clearly exemplify the importance and the positive impact of combining both tools, GS and HTP, inside a breeding program. They also indicate that inclusion of improved phenotyping methods can aid in the improvement of model accuracy. Based on the hybrid gains, we highly recommend the implementation of both, genomic selection and HTP into a breeding program to increase the long-term genetic gain, with both small and larger heritabilities.

Based on the cost, the implementation of genomic selection scenario (GSe) should be considered, once the cost per gain of this scenario is lower for both levels of heritability in higher levels of G×E interaction. The cost-effectiveness of any tool applied to a breeding program is considered as critical and relevant for the maintenance of the program (Riedelsheimer and Melchinger, 2013; Crossa et al., 2017; Muleta et al., 2019). However, to

deal with high levels of G×E, the possibility of implementation of good phenotyping tools (GSe_HTP) also has its values and can return higher gains even though the cost per gain is relatively high compared with GSe.

As we mentioned before, the implementation of HTP in a breeding program has some implications. If we considered the implementation of HTP without changing the breeding program size, the most advantageous strategy is to implement both, genomic selection and HTP. On the other hand, only increasing the number of field plots generates similar results to those of the main strategy applied here (increase the number of repetitions and the trait heritability). However, for traits with small heritabilities, the scenario GSe_HTP returned a superior result that should be considered for this situation. This outcome was also observed in the large set of simulations. It seems that when we consider a robust breeding program, the implementations of the scenarios GSe_HTP represented the best strategy returning the highest gains at the lowest cost per gain compared with the conventional scenarios (CONVe and CONVe_HTP).

In practice, breeding programs consider several traits at a time for selection purposes and it may use the correlation among those traits in the analyses (Montesinos-López et al., 2016, 2019), which could improve the model's reliability (Covarrubias-Pazaran et al., 2018). Thus, the costs of phenotyping and genotyping per gain may decrease, if we consider more than one trait at a time and, therefore, increase the prediction accuracy. In addition, if we keep reducing the price of genotyping and phenotyping technologies, the implementation of both, which already return improved results, will be unbeatable. Consequently, it will cause an impact in the way that we handle breeding programs, especially focused on traits largely controlled by the environment, which represents the reality in most breeding programs.

CONCLUSION

The results presented above addresses questions regarding the best way to include tools to optimize breeding programs aimed at releasing hybrid cultivars. We provide insights based on a sweet corn pipeline with two growing seasons in a target and non-target environment. For complex traits in breeding programs that account for higher levels of G×E interaction to invest in genomic selection represents the best strategy for create and maintain long-term hybrid gain. In addition, with an unconstrained budget the investment in both, genomic selection and HTP, seems to be the optimal choice.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Food and Agriculture SCRI 2018-51181-28419 and AFRI 2019-05410 to M.F.R.R.). We also thank the financial support from the Brazilian Government through the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) through the CAPES-PrInt scholarship. This study was financed in part by the CAPES - Finance Code 001.

SUPPLEMENTAL MATERIAL

Table S1. Summary of number of plots and annual total costs at each scenario for the small breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased trait heritability.

Table S2. Summary of number of plots and annual total costs at each scenario for the small breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased the number of repetitions for trait assessment.

Table S3. Summary of number of plots and annual total costs at each scenario for the large breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased the number of repetitions and the heritability of the target trait.

Table S4. Hybrid gain in unit per levels of G×E interaction (G×E 0, G×E 30, and G×E 120) of four scenarios considering the implementation of high throughput phenotyping. In this set of simulations, we increase the trait heritability for the scenarios with HTP.

Table S5. Hybrid gains in unit, costs to implement each scenario in unit, and costs divided by hybrid gains. All reports are made under three different level of G×E interaction (G×E 0, G×E 30, and G×E 120) for the four scenarios with and without the implementation of highthroughput phenotyping. In this strategy we doubled the number of repetitions and increase the heritability in trait assessment.

Table S6. Hybrid gains in unit, costs to implement each scenario in unit, and costs divided by hybrid gains. All reports are made under three different level of G×E interaction (G×E 0, G×E 30, and G×E 120) for the four scenarios with and without the implementation of highthroughput phenotyping. In this strategy we doubled the number of repetitions and increase the heritability in trait assessment. The results presented were a report of the breeding program simulation 10 times larger than the sweet corn breeding program.

Figure S1. Genetic gain overtime for a trait with 0.7 of broad sense heritability for the scenarios: CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Figure S2. Prediction accuracy of each scenario at F3 and F5, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.3, for the scenarios: CONV:

conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Figure S3. Prediction accuracy of each scenario at F3 and F5, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.7 for the scenarios: CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Figure S4. Genetic variance overtime for a trait with 0.7 of broad sense heritability for the scenarios: CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

Figure S5. Hybrid gain overtime for a trait with 0.7 of broad sense heritability for the scenarios: CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

Figure S6. Prediction accuracy for the smaller set of simulations, for each scenario at F3 and F5 populations predictions, over the 30 years of breeding program and a trait heritability of 0.7. and the scenarios: CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

Figure S7. Prediction accuracy for the bigger set of simulations, for each scenario at F3 and F5 populations predictions, over the 30 years of breeding program and a trait heritability of 0.3. and the scenarios: CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

Figure S8. Prediction accuracy for the bigger set of simulations, for each scenario at F3 and F5 populations predictions, over the 30 years of breeding program and a trait heritability of 0.7. and the scenarios: CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

REFERENCES

Atkinson, J.A., R.J. Jackson, A.R. Bentley, E. Ober, and D.M. Wells. 2018. Field Phenotyping for the Future. Annual Plant Reviews. Wiley. p. 719–736

Bančič, J., C.R. Werner, R.C. Gaynor, G. Gorjanc, D.A. Odeny, et al. 2021. Modeling Illustrates That Genomic Selection Provides New Opportunities for Intercrop Breeding. Front. Plant Sci. 12. doi: 10.3389/fpls.2021.605172.

Bernardeli, A., J.R.A.S. de C. Rocha, A. Borém, R. Lorenzoni, R. Aguiar, et al. 2021. Modeling spatial trends and enhancing genetic selection: An approach to soybean seed composition breeding. Crop Sci. 61(2): 976–988. doi: 10.1002/csc2.20364.

Bernardo, R. 2021. Upgrading a maize breeding program via two-cycle genomewide selection: Same cost, same or less time, and larger gains. Crop Sci. (March): 2444–2455. doi: 10.1002/csc2.20516.

Chen, G.K., P. Marjoram, and J.D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19(1): 136–142. doi: 10.1101/gr.083634.108.

Coelho, I.F., M.A. Peixoto, T.D.S. Marçal, A. Bernardeli, R.S. Alves, et al. 2021. Accounting for spatial trends in multi-environment diallel analysis in maize breeding (M. Rahimi, editor). PLoS One 16(10): e0258473. doi: 10.1371/journal.pone.0258473.

Covarrubias-Pazaran, G., B. Schlautman, L. Diaz-Garcia, E. Grygleski, J. Polashock, et al. 2018. Multivariate GBLUP improves accuracy of genomic selection for yield and fruit weight in biparental populations of vaccinium macrocarpon Ait. Front. Plant Sci. 9: 1–13. doi: 10.3389/fpls.2018.01310.

Crossa, J., R. Fritsche-Neto, O.A. Montesinos-Lopez, G. Costa-Neto, S. Dreisigacker, et al. 2021. The Modern Plant Breeding Triangle: Optimizing the Use of Genomics, Phenomics, and Enviromics Data. Front. Plant Sci. 12(April): 1–6. doi: 10.3389/fpls.2021.651480.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, et al. 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22(11): 961–975. doi: 10.1016/j.tplants.2017.08.011.

Dias, K.O.G., S.A. Gezan, C.T. Guimarães, A. Nazarian, L.C. Silva, et al. 2018. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity (Edinb). 121(1): 24–37. doi: 10.1038/s41437-018-0053-6.

Feng, L., S. Chen, C. Zhang, Y. Zhang, and Y. He. 2021. A comprehensive review on recent applications of unmanned aerial vehicle remote sensing with various sensors for high-throughput plant phenotyping. Comput. Electron. Agric. 182: 106033.

Ferrão, L.F. V., C.D. Marinho, P.R. Munoz, and M.F.R. Resende. 2020. Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. Crop Sci. 60(2): 666–677. doi: 10.1002/csc2.20096.

Fritsche-Neto, R., G. Galli, K.L.R. Borges, G. Costa-Neto, F.C. Alves, et al. 2021. Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review. Front. Plant Sci. 12(July): 1–16. doi: 10.3389/fpls.2021.658267.

Gaynor, R.C., G. Gorjanc, A.R. Bentley, E.S. Ober, P. Howell, et al. 2017. A two-part strategy for using genomic selection to develop inbred lines. Crop Sci. 57(5): 2372–2386. doi: 10.2135/cropsci2016.09.0742.

Gaynor, R.C., G. Gorjanc, and J.M. Hickey. 2021. AlphaSimR: An R package for breeding program simulations. G3 Genes, Genomes, Genet. 11(2): 1–21. doi: 10.1093/G3JOURNAL/JKAA017.

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, et al. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54(4): 1476–1488. doi: 10.2135/cropsci2013.03.0195.

Hu, Y., V. Colantonio, B.S.F. Müller, K.A. Leach, A. Nanni, et al. 2021. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nat. Commun. 12(1). doi: 10.1038/s41467-021-21380-4.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127(3): 595–607. doi: 10.1007/s00122-013-2243-1.

Jha, S.K., N.K. Singh, and P.K. Agrawal. 2016. Complementation of sweet corn mutants: a

method for grouping sweet corn genotypes. J. Genet. 95(1): 183–187. doi: 10.1007/s12041-015-0608-8.

Lara, L.A.C., I. Pocrnic, T.P. Oliveira, R.C. Gaynor, and G. Gorjanc. 2022. Temporal and genomic analysis of additive genetic variance in breeding programmes. Heredity (Edinb). 128: 21–32. doi: 10.1038/s41437-021-00485-y.

Lertrat, K., and T. Pulam. 2007. Breeding for increased sweetness in sweet corn. Int. J. Plant Breed. 1(1): 27–30.

Li, D., Z. Xu, R. Gu, P. Wang, J. Xu, et al. 2021. Genomic Prediction across Structured Hybrid Populations and Environments in Maize. Plants 10(6): 1174. doi: 10.3390/plants10061174.

Marinho, C.D., I.F. Coelho, M.A. Peixoto, G.A. Carvalho Júnior, and M.F.R. Resende Jr. 2022. Genomic selection as a tool for maize cultivars development. Rev. Bras. Milho e Sorgo 21(e1285).

McPeek, M.S., and T.P. Speed. 1995. Modeling Inference in Genetic Recombination. Genetics 139: 1031–1044. doi: 10.1007/978-1-4614-1347-9_10.

Meuwissen, T.H.. E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829. doi: 11290733.

Mir, R.R., M. Reynolds, F. Pinto, M.A. Khan, and M.A. Bhat. 2019. High-throughput phenotyping for crop improvement in the genomics era. Plant Sci. 282(January): 60–72. doi: 10.1016/j.plantsci.2019.01.007.

Montesinos-López, O.A., A. Montesinos-López, J. Crossa, F.H. Toledo, O. Pérez-Hernández, et al. 2016. A Genomic Bayesian Multi-trait and Multi-environment Model. G3 Genes|Genomes|Genetics 6(9): 2725–2744. doi: 10.1534/g3.116.032359.

Montesinos-López, A., O.A. Montesinos-López, J. Cuevas, W.A. Mata-López, J. Burgueño, et al. 2017. Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. Plant Methods 13(1): 1–29. doi: 10.1186/s13007-017-0212-4.

Montesinos-López, O.A., A. Montesinos-López, R. Tuberosa, M. Maccaferri, G. Sciara, et al. 2019. Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods. Front. Plant Sci. 10(November): 1–12. doi: 10.3389/fpls.2019.01311.

Muleta, K.T., G. Pressoir, and G.P. Morris. 2019. Optimizing genomic selection for a sorghum breeding program in Haiti: A simulation study. G3 Genes, Genomes, Genet. 9(2): 391–401. doi: 10.1534/g3.118.200932.

Neupane, K., and F. Baysal-Gurel. 2021. Automatic identification and monitoring of plant diseases using unmanned aerial vehicles: A review. Remote Sens. 13(19): 3841.

Oliveira, A.A., M.F.R. Resende, L.F.V. Ferrão, R.R. Amadeu, L.J.M. Guimarães, et al. 2020. Genomic prediction applied to multiple traits and environments in second season maize hybrids. Heredity (Edinb). 125(1–2): 60–72. doi: 10.1038/s41437-020-0321-0.

Pérez-Rodríguez, P., J. Crossa, J. Rutkoski, J. Poland, R. Singh, et al. 2017. Single-Step Genomic and Pedigree Genotype × Environment Interaction Models for Predicting Wheat Lines in International Environments. Plant Genome 10(2). doi: 10.3835/plantgenome2016.09.0089.

Persa, R., P.C.O. Ribeiro, and D. Jarquin. 2021. The use of high-throughput phenotyping in

genomic selection context. Crop Breed. Appl. Biotechnol. 21(Special Issue): 1–11. doi: 10.1590/1984-70332021v21Sa19.

Powell, O., R.C. Gaynor, G. Gorjanc, C.R. Werner, and J.M. Hickey. 2020. A two-part strategy using genomic selection in hybrid crop breeding programs. bioRxiv: 1–46. doi: 10.1101/2020.05.24.113258.

Riedelsheimer, C., and A.E. Melchinger. 2013. Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor. Appl. Genet. 126(11): 2835–2848. doi: 10.1007/s00122-013-2175-9.

Silva, É.D.B., A. Xavier, and M.V. Faria. 2021. Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding. Front. Genet. 12(September): 1–12. doi: 10.3389/fgene.2021.637133.

Sun, J., J.E. Rutkoski, J.A. Poland, J. Crossa, J.L. Jannink, et al. 2017. Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. Plant Genome 10(2). doi: 10.3835/plantgenome2016.11.0111.

Volpato, L., F. Pinto, L. González-Pérez, I.G. Thompson, A. Borém, et al. 2021. High Throughput Field Phenotyping for Plant Height Using UAV-Based RGB Imagery in Wheat Breeding Lines: Feasibility and Validation. Front. Plant Sci. 12(February). doi: 10.3389/fpls.2021.591587.

Wang, X., D. Singh, S. Marla, G. Morris, and J. Poland. 2018. Field-based high-throughput phenotyping of plant height in sorghum using different sensing technologies. Plant Methods 14(1): 1–16.

Wang, N., H. Wang, A. Zhang, Y. Liu, D. Yu, et al. 2020. Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. Theor. Appl. Genet. 133(10): 2869–2879.

Watanabe, K., W. Guo, K. Arai, H. Takanashi, H. Kajiya-Kanegae, et al. 2017. High-throughput phenotyping of sorghum plant height using an unmanned aerial vehicle and its application to genomic prediction modeling. Front. Plant Sci. 8: 421.

Yassue, R.M., G. Galli, R.B. Junior, G. Morota, and R. Fritsche-neto. 2021. A low-cost greenhouse-based high-throughput phenotyping platform for genetic studies : a case study in maize under inoculation with plant growth-promoting bacteria.

Zystro, J., T. Peters, K. Miller, and W.F. Tracy. 2021a. Classical and genomic prediction of hybrid sweet corn performance in organic environments. Crop Sci. 61(3): 1698–1708. doi: 10.1002/csc2.20400.

Zystro, J., T. Peters, K. Miller, and W.F. Tracy. 2021b. Classical and genomic prediction of synthetic open-pollinated sweet corn performance in organic environments. Crop Sci. 61(5): 3382–3391. doi: 10.1002/csc2.20531.

SUPPLEMENTARY MATERIAL

Appendix S1

Table S1. Summary of number of plots and annual total costs at each scenario for the small breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased trait heritability.

Year	Scenario			
	CONVe	**CONVe_HTP**	**GSe**	**GSe_HTP**
1	200	200	300	300
2	1000	1000	1000	1000
3	1320	1320	1320	1320
4	960	960	960	960
5	475	475	475	475
Total plots	3955	3955	4055	4055
Total costs ($)	50900	44900	73000	66600

CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: breeding program with genomic selection. GSe_HTP: breeding program with genomic selection and high throughput phenotyping.

Table S2. Summary of number of plots and annual total costs at each scenario for the small breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased the number of repetitions for trait assessment.

Year	Scenario			
	CONVe	**CONVe_HTP**	**GSe**	**GSe_HTP**
1	200	300	300	500
2	1000	1800	1000	1800
3	1320	1620	1320	1620
4	960	1460	960	1460
5	475	775	475	775
Total plots	3955	5955	4055	6155
Total costs ($)	50900	62100	73000	85200

CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: breeding program with genomic selection. GSe_HTP: breeding program with genomic selection and high throughput phenotyping.

Table S3. Summary of number of plots and annual total costs at each scenario for the large breeding program simulated considering the implementation of high throughput phenotyping (HTP) into the sweet corn breeding pipeline. The values represent the strategy where we increased the number of repetitions and the heritability of the target trait.

Year	Scenario			
	CONVe	**CONVe_HTP**	**GSe**	**GSe_HTP**
1	1600	2600	2600	4600
2	10000	18000	10000	18000
3	13200	16200	13200	16200
4	9600	14600	9600	14600
5	4750	7750	4750	7750
Total plots	39150	59150	40150	61150
Total costs ($)	508200	616200	734200	852200

CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: breeding program with genomic selection. GSe_HTP: breeding program with genomic selection and high throughput phenotyping.

Table S4. Hybrid gain in unit per levels of G×E interaction (G×E 0, G×E 30, and G×E 120) of four scenarios considering the implementation of high throughput phenotyping. In this set of simulations, we increase the trait heritability for the scenarios with HTP.

Scenario/GxE	Hybrid gain			Cost	Cost per gain		
	0	30	120		0	30	120
Heritability equals to 0.3							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.11	1.07	1.11	0.88	0.79	0.72	0.79
GSe	0.95	1.10	1.42	1.43	1.52	1.30	1.01
GSe_HTP	1.02	1.14	1.48	1.31	1.29	1.14	0.88
Heritability equals to 0.7							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.04	1.03	1.02	0.88	0.85	0.86	0.86
GSe	0.90	1.07	1.45	1.43	1.60	1.35	0.99
GSe_HTP	0.91	1.06	1.49	1.31	1.44	1.24	0.88

CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: breeding program with genomic selection. GSe_HTP: breeding program with genomic selection and high throughput phenotyping.

Table S5. Hybrid gains in unit, costs to implement each scenario in unit, and costs divided by hybrid gains. All reports are made under three different level of G×E interaction (G×E 0, G×E 30, and G×E 120) for the four scenarios with and without the implementation of highthroughput phenotyping. In this strategy we doubled the number of repetitions and increase the heritability in trait assessment. The results of hybrid gains represented the final value of each scenario (year 30), considering the means of 50 repetitions of each pipeline, for two trait heritabilities (0.3 and 0.7). In this set of simulations, we doubled the number of repetitions in trait assessment for the scenarios with HTP.

Scenario/G×E	Hybrid gain (unit)			Cost	Cost per gain		
	0	30	120		0	30	120
Heritability equals to 0.3							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.03	1.04	1.10	1.22	1.19	1.17	1.11
GSe	0.92	1.09	1.49	1.43	1.56	1.32	0.96
GSe_HTP	1.01	1.19	1.60	1.67	1.65	1.40	1.05
Heritability equals to 0.7							
CONVe	1	1	1	1	1	1	1
CONVe _HTP	1.01	1.00	1.03	1.22	1.21	1.22	1.18
GSe	0.90	1.07	1.48	1.43	1.60	1.34	0.97
GSe_HTP	0.91	1.12	1.52	1.67	1.83	1.50	1.10

CONVe: conventional breeding program. CONVe_HTP: conventional breeding program with high throughput phenotyping. GSe: breeding program with genomic selection. GSe_HTP: breeding program with genomic selection and high throughput phenotyping.

Table S6. Hybrid gains in unit, costs to implement each scenario in unit, and costs divided by hybrid gains. All reports are made under three different level of G×E interaction (G×E 0, G×E 30, and G×E 120) for the four scenarios with and without the implementation of highthroughput phenotyping. In this strategy we doubled the number of repetitions and increase the heritability in trait assessment. The results of hybrid gains represented the final value of each scenario (year 30), considering the means of 50 repetitions of each pipeline, for two trait heritabilities (0.3 and 0.7). In this strategy we doubled the number of repetitions in trait assessment and increase the heritability of the trait. The results presented were a report of the breeding program simulation 10 times larger than the sweet corn breeding program.

Scenario/GxE	Hybrid gain			Cost	Cost by gain		
	0	30	120		0	30	120
Heritability equals to 0.3							
CONVe	1	1	1	1	1	1	1
CONVe_HTP	1.04	1.02	1.07	1.21	1.16	1.18	1.13
GSe	1.09	1.28	1.74	1.44	1.32	1.13	0.83
GSe_HTP	1.18	1.40	1.86	1.68	1.43	1.20	0.90
Heritability equals to 0.7							
CONVe	1	1	1	1	1	1	1
CONVe_HTP	1.01	0.99	1.02	1.21	1.20	1.23	1.19
GSe	1.01	1.29	1.77	1.44	1.43	1.12	0.81
GSe_HTP	1.05	1.34	1.86	1.68	1.60	1.26	0.90

CONVe: conventional breeding program. CONVe_HTP: conventional breeding and high throughput phenotyping. GSe: genomic selection breeding program. GSe_HTP: genomic selection breeding program with high throughput phenotyping.

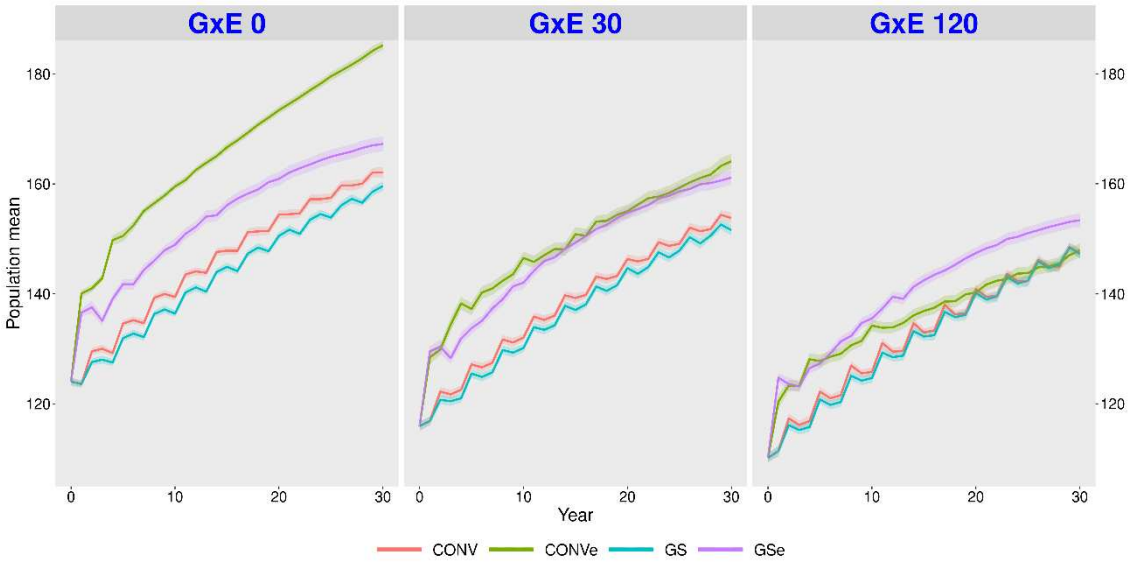


Figure S1. Genetic gain overtime for the four simulated scenarios for the trait with 0.7 of broad sense heritability. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The gain is plotted as a mean of the parents for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the genetic mean for the 50 replicates and the shading represents the standard error. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

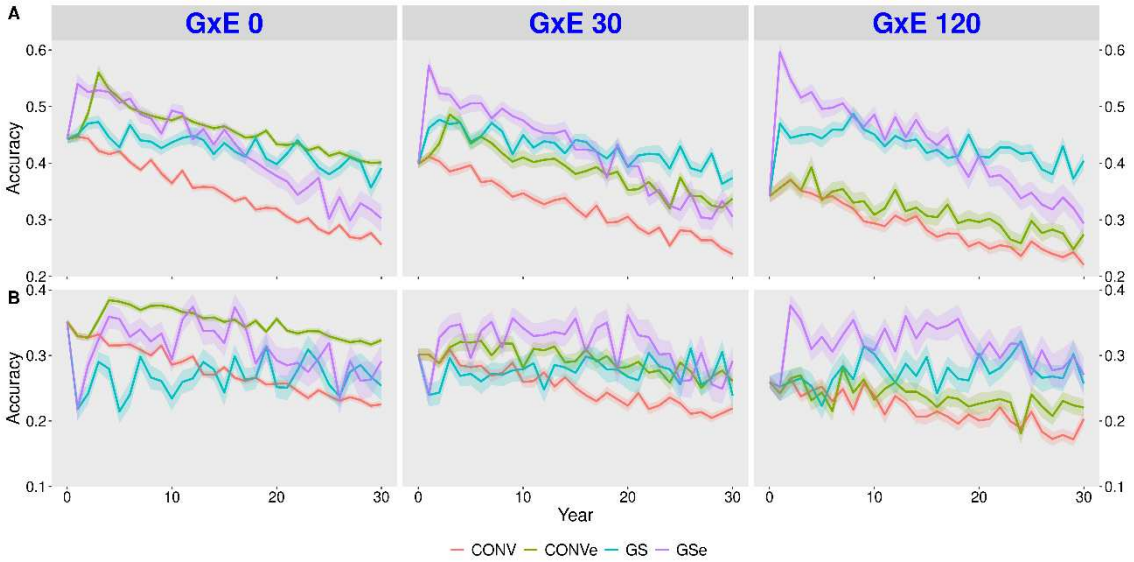


Figure S2. Prediction accuracy of each scenario at F3 and F5, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.3. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The predictive accuracy is plotted as a mean of the model accuracy for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the selective accuracy for the 50 replicates and the shading represents the standard error. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

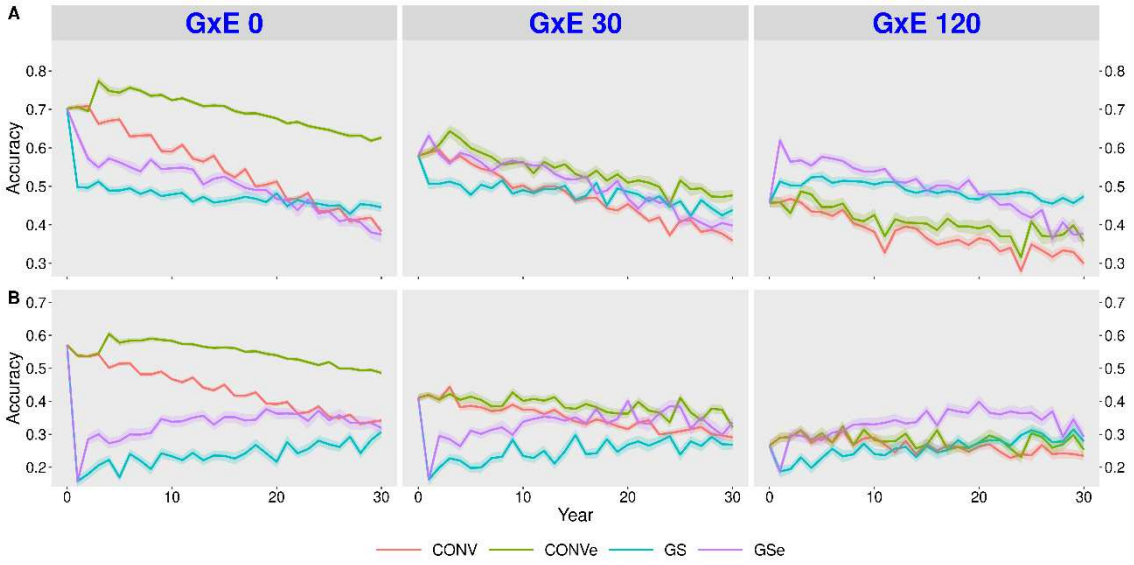


Figure S3. Prediction accuracy of each scenario at F3 and F5, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.7. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The predictive accuracy is plotted as a mean of the model accuracy for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the selective accuracy for the 50 replicates and the shading represents the standard error. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

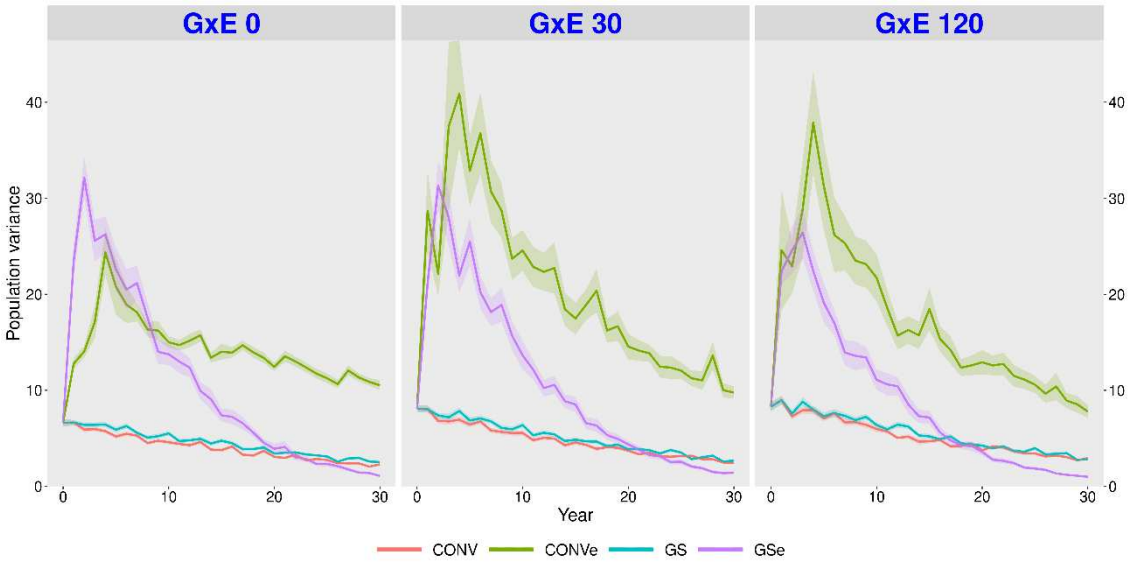


Figure S4. Genetic variance overtime for the four simulated scenarios for the trait with 0.7 of broad sense heritability. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The gain is plotted as a mean of the parents for each cycle. The lines within each of the three panels represents the six simulated scenarios, where each line represents the genetic mean for the 50 replicates and the shading represents the standard error. CONV: conventional breeding program. CONVe: conventional breeding program with early selection. GS: conventional genomic selection breeding program. GSe: genomic selection breeding program with early selection.

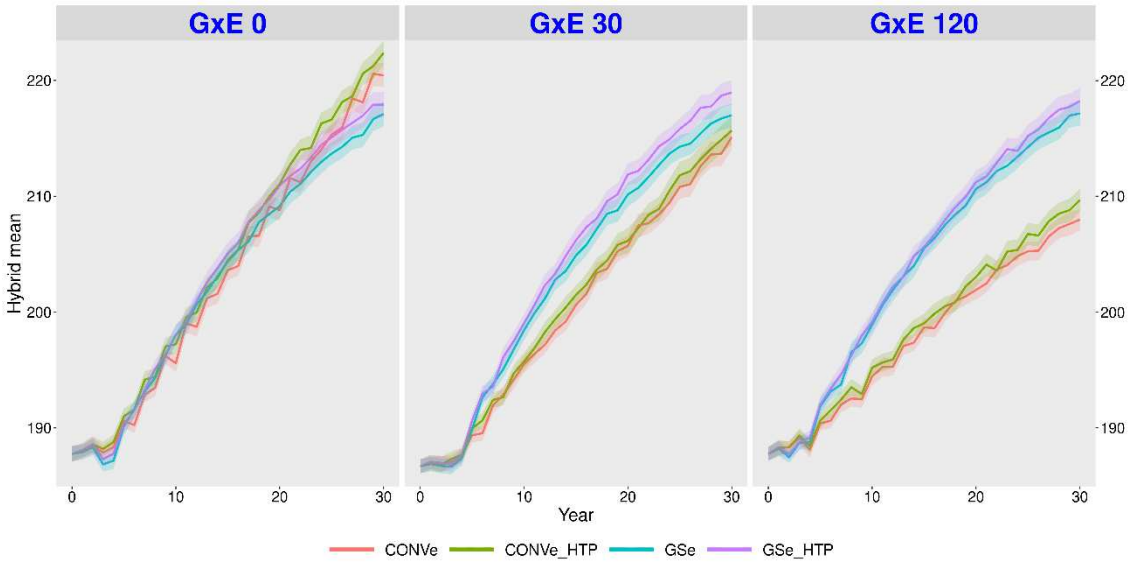


Figure S5. Hybrid gain overtime for the four simulated scenarios for the trait with 0.7 of broad sense heritability. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The hybrid gain is plotted as a mean of the hybrids for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the hybrid gain for the 50 replicates and the shading represents the standard error. CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

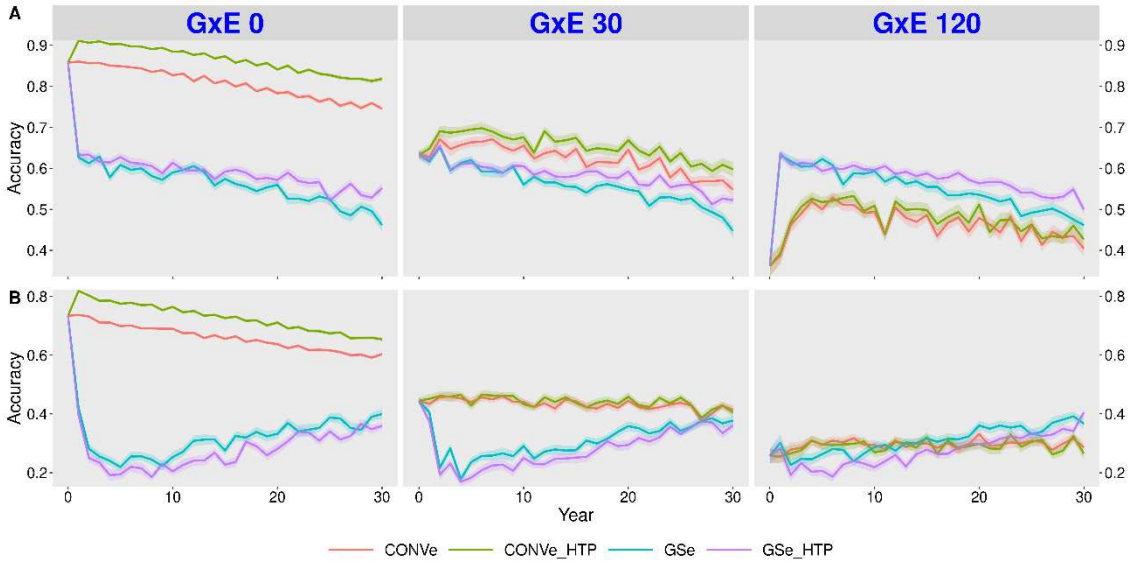


Figure S6. Prediction accuracy of each scenario at F3 and F5, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.7. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The predictive accuracy is plotted as a mean of the model accuracy for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the selective accuracy for the 50 replicates and the shading represents the standard error. CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

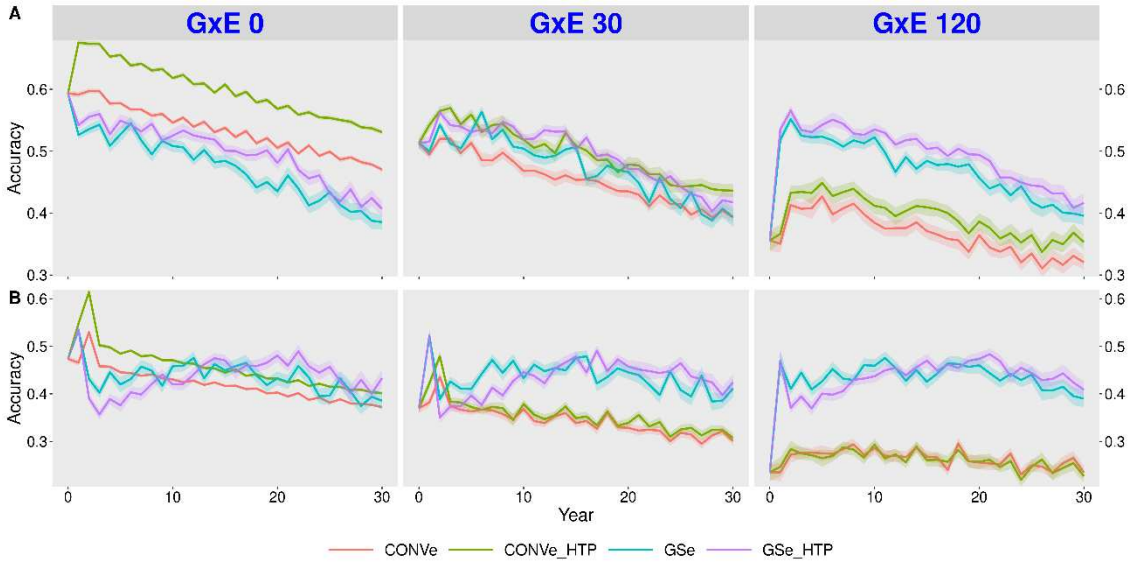


Figure S7. Prediction accuracy for the bigger set of simulation, considering each scenario at F3 and F5 populations predictions, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.3. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The predictive accuracy is plotted as a mean of the model accuracy for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the selective accuracy for the 50 replicates and the shading represents the standard error. CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

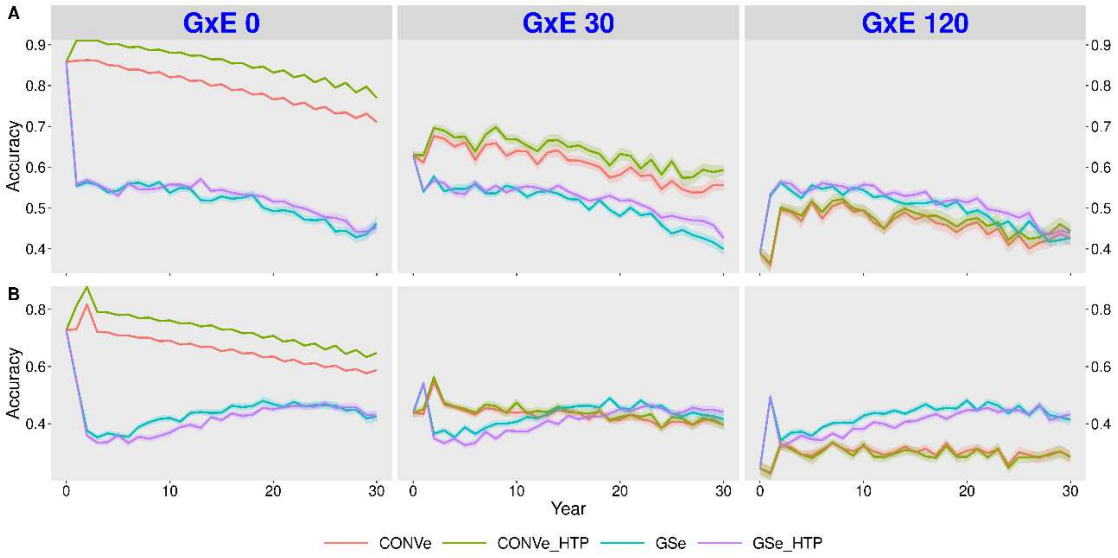


Figure S8. Prediction accuracy for the bigger set of simulation, considering each scenario at F3 and F5 populations predictions, over the 30 years of breeding program for the four scenarios and a trait heritability of 0.7. Results are show under genotype-by-environment (G×E) interaction of 0 (G×E 0), 30 (G×E 30), and 120 (G×E 120). The predictive accuracy is plotted as a mean of the model accuracy for each cycle. The lines within each of the three panels represents the four simulated scenarios, where each line represents the selective accuracy for the 50 replicates and the shading represents the standard error. CONVe: conventional breeding program with early selection. CONVe_HTP: conventional breeding program with early selection and highthroughput phenotyping. GSe: genomic selection breeding program with early selection. GSe_HTP: genomic selection breeding program with early selection and highthroughput phenotyping.

Chapter 2

Genomic prediction of hybrid performance for boosting sweet corn breeding program

Genomic prediction of hybrid performance for boosting sweet corn breeding program

Marco Antônio Peixoto[1,2], Kristen A. Leach[2], Diego Jarquin[3], Patrick Flannery[4], William F. Tracy[4], Jared Zystro[5], Leonardo Lopes Bhering[1], Márcio F. R. Resende[2,*]

[1] Universidade Federal de Viçosa, Laboratório de Biometria, Viçosa, Minas Gerais, Brazil.

[2] Department of Horticultural sciences, University of Florida, Gainesville, Florida, United States

[3] Department of Agronomy, University of Florida, Gainesville, Florida, United States

[4] Department of Agronomy, University of Wisconsin-Madison, Madison, Wisconsin, United States

[5] Organic seed alliance

*** Corresponding author**

Abstract

Sweet corn breeding programs rely on the hybrid as a cultivar. For this reason, in silico prediction can boost hybrid prediction and reduce costs and time in a breeding program. The aim of this study was to explore the potential of implementing genomic selection in a sweet corn breeding program through hybrid prediction in an across-year and across-site framework. We evaluated 506 hybrids in 6 environments (California, Florida, and Wisconsin, in the years of 2020 and 2021). A total of 21 traits from three different groups were measured (vegetative-, ear-, and consumer-related traits) over the six environments. Eight statistical models were considered for prediction, as the combination of two genomic models (GBLUP and RKHS) with to different kernels (additive and additive + dominance), and in a single- and multi-trait framework. Also, three different cross-validations schemes were tested (CV1, CV0, and CV00), regarding observed/unobserved genotypes/environments. At the end, models were compared based on the correlation between the estimated breeding values/total genetic values and phenotypic measurements. Overall, trait heritabilities and correlations vary largely among traits. The RKHS models outperformed the GBLUP models in all cross-validation schemes, ranging from 1.02% to 34% improvement. Models with additive plus dominance kernels presented slight improvement for some of the models examined. In addition, the models for across-site prediction performed better in the CV1 scheme (0.325), follow by CV0 (0.306), and CV00 (0.122). The same pattern was observed for across-year prediction (CV0: 0.323 and CV00: 0.139). Hence, RKHS should be considered as a standard model for sweet corn hybrid prediction. In addition, the implementation of GS in a sweet corn breeding program can boost the results by optimizing the testcross stage and the candidates that reaches field stage.

Keywords: G×E interaction, hybrid prediction, non-additive effects, RKHS model, cross-validation schemes.

1. Introduction

Hybrids represented a final and/or commercial product for multiple crops and, consequently, several breeding programs. They can be generated through a simple cross between individuals that may or may not belong to different genetic groups, such as heterotic groups (Oliveira et al., 2022). Hybrid crop breeding generally advances though an evaluation process where finish lines from a given program are crossed to testers. Hybrids that perform well are then evaluated in an increasing number of representative environments or target population of environments (TPE). Classically, selection of the best hybrids in a breeding program is made based on general combining ability (GCA), which represents how good is a parent, based on the hybrid performance for all crosse that it was part of. Specific combining ability (SCA) is also considered for specific crosses, indicating how superior the hybrid is compared to the mean of its parents.

Hence, to test all possible crosses is not an easy task in a breeding program. For example, with only 100 potential parents (*N*), the possible crosses to make is $C = \frac{N * (N-1)}{2}$, and it reaches ~4950 crosses (without reciprocal crosses and selfies). The number of putative crosses that can be generated among a set of lines and the testers can become quite large rapidly (Zystro et al., 2021a).

A modern breeding program can be divided into two phases: line development followed by product development, where hybrids are created and tested (Cowling et al., 2020; Powell et al., 2020). In such programs, the use of genomic selection (GS) tools can aid in the prediction of a potentially promising hybrids. The deployment of the GS in this stage (product development) uses parental genotypes to predict hybrid performance. Afterwards, the development of hybrid prediction through GS models is expected to increase the probability of

superior hybrids that will reach advanced field-testing stages (Marinho et al., 2022). These methods can also reduce efforts and costs of hybrid assessment.

Sweet corn represents a hybrid crop which aims to produce products that will meet producer and consumer driven needs. This crop has breeding particularities. First, even though there is heterosis in the crossed hybrids, the absence of well-defined heterotic groups makes it impossible to implement reciprocal recurrently selection in the breeding program (Peixoto et al. 2022). In addition, the target outputs of breeding are both marketable product and processing, such as canned and frozen corn. Then, very often some traits are deemed higher in priority by the growers together with high productivity. For example, consumer-related traits, such as those represent ear appearance and sensing traits (flavor and texture) are combined with vegetative traits that are important for growers, such as germination, leaf angle, and days to pollination. In response, a breeder has multiple traits that needed to be accounted for before releasing a line.

Genomic hybrid prediction is broadly applied in maize breeding (Fritsche-Neto et al., 2021). In the same sense, statistical methods have been evolving and developed for leveraging these predictions. Generally, the genomic models used for hybrid prediction for maize breeding are genomic BLUP (GBLUP) and reproducing kernel Hilbert space (RKHS) (Alves et al., 2019; Cuevas et al., 2019; Krause et al., 2020; Marinho et al., 2022). Even though similar in implementation, the differences between GBLUP and RKHS are related to the mathematical equations that is used to build the genomic relationship matrix (or kernels) that is used in the models for prediction (VanRaden, 2008; De Los Campos et al., 2010; Vitezica et al., 2013). In addition, methods that consider both, additive and non-additive effects (such as dominance) have demonstrated significant increases in trait predictability (Ferrão et al., 2020; Oliveira et al., 2020). For instance, the addition of dominance effects has been demonstrated to increases the accuracy for model prediction in sweet corn (Zystro et al., 2021b, 2021a).

Another factor being explore in hybrid prediction is the correlation between traits (Covarrubias-Pazaran et al., 2018). This can help increase the prediction accuracy of traits with lower heritabilities by borrowing information from traits with higher heritability (Mrode, 2014). The use of several traits would greatly benefit trait prediction, especially when there are several targets in a sweet corn breeding program. Lastly, the most challenging factor that impacts genomic prediction is related to the genotype-by-environment interaction (G×E). As demonstrated by Jarquín et al. (2014), the G×E interaction impacts the genotype ranking making it difficult to calibrate the model, and as a result decreases modeling capabilities. The use of multi-year information will help to circumvent this issue when predicting the hybrid performance.

In this study we explore multiple models that can be used to implement genomic selection in a sweet corn breeding, and the possibilities that hybrid prediction could represent in the program.

2. Material and Methods

Plant material

A total of 506 hybrids were created from 62 lines that came from a collaboration between the University of Florida and the University of Wisconsin sweet corn breeding programs. The hybrids were evaluated in three sites: Florida (FLO), California (CAL), and Wisconsin (WIS) across two years, 2020 and 2021. Here, we deemed an environment to be the combination between sites and years (six environments total) (**Supplementary material – Figure S1**). The two trials in CAL site for the years of 2020 (CA20) and 2021 (CA21) had 246 and 39 hybrids measured. At the FLO site for 2020 (FL20) and 2021 (FL21) 418 and 203 hybrids were assessed, respectively. For WIS site, 236 and 39 hybrids were assessed, in 2020 and 2021 (WI20 and WI21, respectively).

Phenotyping

The assessment of hybrids in the field followed a randomized complete block design (RCBD) with one observation per plot at all environments. Initially, a block effect was also used within the repetitions at FL20 (two repetitions and five blocks), FL21 (2 repetitions and three blocks), and CA20 (three repetitions and 5 blocks) environments. However, the hybrid experiments were planted with two repetitions in WI20 and three repetitions in CA21/WI21 environments, all following a RCBD.

In the sweet corn breeding program, several traits related to the ear appearance, vegetative aspects of the crop and flavor are considered important for market specificities. This brings to the equation the needed to breed for several traits at time, and, at the end to measure and build models capable of predicting these traits. A total of 20 traits were measured over the six environments (a complete list of all the traits is presented in the **Supplementary material – Table S1**). The traits could be grouped into three different categories: Vegetative-related traits (STAND: stand count; DTP: days to pollination; DTS: days to silking; PH: plant height; EH: ear height), ear-related traits (EL: ear length. EW: ear width; TPF: tip fill; HP: husk protection; KRN: kernel row number; SOL: solidity; TP: taper; CUR: curvature; HAP: husk appearance; RAP: row appearance), and consumer-related traits (ES: ear shape; CR: color; FL: flavor; TXT: texture; RT: over-all ear rating).

Phenotypic analyses

In the first step of the analysis, the estimation of variance components and prediction of genotypic values for the traits assessed was made via the residual maximum likelihood/best linear unbiased prediction (REML/BLUP) procedure. The statistical model associated with the evaluation of hybrids in a randomized complete block design with one observation per plot was given by the following equations:

$$\mathbf{y} = \mathbf{X}\boldsymbol{r} + \mathbf{Z}\boldsymbol{g} + \boldsymbol{W}\boldsymbol{p} + \boldsymbol{e}, \text{ (i)}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{r} + \mathbf{Z}\boldsymbol{g} + \boldsymbol{e}, \text{ (ii)}$$

where *y* is the vector of phenotypes; *r* is the vector of replication effects (assumed as fixed), added to the overall mean; *g* is the vector genotypic effects [(assumed as random) ($g \sim N\left(0, \sigma_g^2\right)$, where σ_g^2 is the genotypic variance]; *p* is the vector block effects [(assumed as random) $b \sim N(0, \sigma_b^2)$, where σ_b^2 is the block variance]; and *e* is the vector of residuals [(random) $e \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance]. Uppercase letters (***X***, ***Z***, and ***W***) represent the incidence matrices for *r*, *g*, and *p*, respectively. For the random effects, significance was tested by the likelihood ratio test (LRT) using a chi-square statistic with 1 degree of freedom and a 5% probability of error type I (Wilks, 1938). BLUE values (best linear unbiased estimation) were also generated and used for the second step of the analysis.

Genotyping

Whole genome sequence was used to determine the genotypes of the inbred lines (Colantonio et al. in prep). A initial subset of SNP data was generated consisting of 200k markers randomly sampled from the whole genome sequencing using vcftools (Danecek et al., 2011). Single nucleotide polymorphisms with a minor allele frequency of less than 1% and missing data of greater than 30% were removed from the SNP data set. After the process, a final set of 70,562 SNPs was used for the analyses.

Genomic selection models

The models GBLUP and RKHS were evaluated with both, single- and multi-trait framework. The models incorporated additive effect (A) and additive plus dominance effects (AD). For the dominance matrix (RKHS model only), the SNPs matrix was coded as 0 for both homozygous classes (AA and aa) and as 1 for the heterozygous class, whereas the intermediated values (0.5 and 1.5) that came from one homozygous and one heterozygous inbred line were coded as missing data (**Figure 1**). We describe the model's structure below.

Genomic best linear unbiased prediction (GBLUP): The GBLUP model for single-trait structure and additive effect (1) and additive plus dominance effects (2) are represented as follows:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{e}, \text{ [1]}$$

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}, \text{ [2]}$$

where y is the vector of BLUE values of the hybrids predicted in the first stage of the analyses, μ is the overall mean, a represents the vector of additive genetic effects, where $a \sim N(0, \boldsymbol{G}_a \sigma_a^2)$, where $\boldsymbol{G}_a$ is the additive genomic relationship matrix and σ_a^2 is the additive genetic variance, d represents the vector of dominance effects, where $d \sim N(0, \boldsymbol{G}_d \sigma_d^2)$, where $\boldsymbol{G}_d$ is the dominance genomic relationship matrix and σ_d^2 is the dominance variance, and e is the residual error variance, where $e \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represents the incidence matrices to a and d effects, respectively.

The multitrait GBLUP model with additive effect (3) and additive plus dominance effects (4) are given by:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{e}, \text{ [3]}$$

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}, \text{ [4]}$$

where y is the vector with BLUE values of all traits predicted from the hybrids at first stage of the analyses, μ is the overall mean, a represents a matrix of additive genetic effects for all trait, where $a \sim N(0, \boldsymbol{\Sigma} \otimes \boldsymbol{G}_a)$, $\boldsymbol{G}_a$ is the additive genomic relationship matrix and $\boldsymbol{\Sigma}$ is the additive genetic variance-covariance matrix across traits, d represents the vector of dominance effects, where $d \sim N(0, \boldsymbol{\Sigma} \otimes \boldsymbol{G}_d)$, where $\boldsymbol{G}_d$ is the dominance genomic relationship matrix and $\boldsymbol{\Sigma}$ is the dominance variance-covariance matrix across traits and e is the residual error variance, where

$e \sim N(0, I \otimes \mathbf{R})$, where $\boldsymbol{R}$ is the residual variance-covariance matrix for each hybrid and for all traits. $\boldsymbol{Z}a$ and $\boldsymbol{Z}d$ represents the incidence matrices to a and d effects, respectively.

Reporducing Kernel Hilbert Space (RKHS): A model, consisting of the semi-parametric kernel RKHS, was used for single-trait prediction, using additive (5) and additive + dominance effects models (6). The following models were used:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{e}, \text{ [5]}$$

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}, \text{ [6]}$$

where y is the vector of BLUE values from the hybrids predicted in the first stage of the analyses, μ is the overall mean, a represents the vector of additive genetic effects, where $a \sim N(0, \boldsymbol{K}_a \sigma_a^2)$, where $\boldsymbol{K}_a$ is the additive symmetric semipositive definite matrices representing the covariance of the genetic values, d represent a vector of dominance effects, where $a \sim N(0, \boldsymbol{K}_d \sigma_d^2)$, where $\boldsymbol{K}_d$ is the dominance symmetric semipositive definite matrices representing the covariance of the genetic values, and e is the residual error variance, where $e \sim N(0, \sigma_e^2)$. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represents the incidence matrices to a and d effects, respectively.

The multitrait RKHS model were implemented with additive (7) and additive plus dominance effects (8), as follow detailed:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{e}, \text{ [7]}$$

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}, \text{ [8]}$$

where y is the vector of BLUE values from the hybrids predicted in the first stage of the analyses, μ is the overall mean, a represents the vector of additive genetic effects, where $a \sim N(0, \boldsymbol{\Sigma} \otimes \boldsymbol{K}_a)$, where $\boldsymbol{K}_a$ is the additive symmetric semipositive definite matrices representing the covariance of the genetic values and $\boldsymbol{\Sigma}$ is the additive variance-covariance matrix across traits, d represent a vector of dominance effects, where $a \sim N(0, \boldsymbol{\Sigma} \otimes \boldsymbol{K}_d)$, where

$\boldsymbol{K}_d$ is the dominance symmetric semipositive definite matrices representing the covariance of the genetic values and $\boldsymbol{\Sigma}$ is the dominance variance-covariance matrix across traits, and e is the residual error variance, where $e \sim N(0, \boldsymbol{I} \otimes \boldsymbol{R})$, where $\boldsymbol{R}$ is the residual variance-covariance matrix for each hybrid and for all traits. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represents the incidence matrices to a and d effects, respectively.

In the RKHS, $\boldsymbol{K}_a$ and $\boldsymbol{K}_d$ are represented by: $\boldsymbol{K}_a = \exp(-\varphi_{a_r} D_a^2)$, and $\boldsymbol{K}_d = \exp(-\varphi_{d_r} D_d^2)$, where φ_{a_r} and φ_{d_r} represents the bandwidth parameters for additive and dominance models (Pérez and de los Campos, 2014) and D_a^2 and D_d^2 represents the Euclidian distance matrix using the SNP matrix for additive effects, and dominance SNP matrix for dominance effects, respectively. Here, we followed Morota and Gianola (2014) and applied the kernel averaging model, were $\boldsymbol{K}_a$ and $\boldsymbol{K}_d$ were represented for three kernels that came from the three different values of the bandwidth parameters (φ_{a_r} and φ_{d_r}), represented by 5/h, 1/h and 0.2/h, where h is the 5th percentile of the D_a^2 or D_d^2 leading to local, intermediate and global kernels, respectively (Morota and Gianola, 2014).

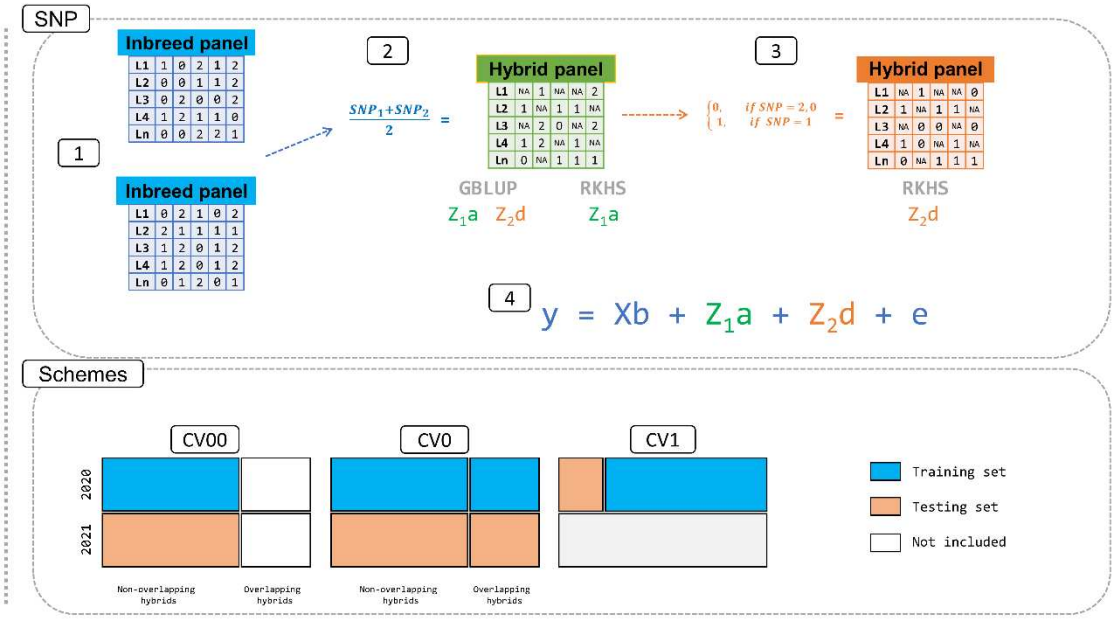


Figure 1 - Pipeline for the prediction in the sweet corn dataset. The three trials are: Florida, Wisconsin, and California. The SNP panel for the lines (1) were combined based on the formula in (2) and coded to the additive (Z_1a) and dominance (Z_1d) kernels. One additional step is needed to implement Z_1d for RKHS model (3). All information generated is then used in step (4) for the prediction of genomic estimated breeding values (GEBV) using GBLUP and RKHS models. The schemes tested in this study are CV00: untested hybrids in untested environments. CV0: tested hybrids in untested environments. CV1: untested hybrids in tested environments. SNP: single nucleotide polymorphism. GBLUP: genomic best linear unbiased prediction. RKHS: reproducing kernel Hilbert space.

Cross validation schemes

Three different cross-validations schemes were implemented in the analyses. (i) within-year cross validation (untested hybrids in tested environments or CV1) with three environments individually (CA20, FL20, and WI20, **Figure 1**). A total of 5 k-folds were used (20% of the hybrids as a testing set and 80% for training), and permutation in-between those groups led to five k-folds of training/testing sets. In addition, we replicate the 5 k-folds process through 20 repetitions. (ii) across-year prediction (tested hybrids in untested environments or CV0), where

hybrid data measured at each site in the first year (2020) was used to training the model and predict the hybrids at each site in 2021, separately (*i.e*. CA20, FL20, and WI20 as training set for the prediction of CA21. FL21, and WI21, respectively). (iii) across-year prediction where we cut-off hybrid information that overlapped between both years (untested hybrids in untested environment, or CV00). Also, we only used the information of the traits EL, EW, and TPF (the only traits that were measured in all environments) to build the CV0 and CV00 schemes. The model accuracy was calculated between the estimated breeding value and the BLUE values estimated in the model in the model.

Software implementation

The first step (phenotypic analyses) was carried out in R package ASREML (Gilmour et al., 2015), with the genomic models being implemented in BGLR package (Pérez and de los Campos, 2014; Pérez-Rodríguez and de los Campos, 2022). The Bayesian models used 30000 iterations, a burn-in of 3000, and a sampling interval (thin) of 10, totaling 2700 iterations. All the codes and the data used on the analyses are available at github (link after acceptance).

3. Results

Here, we implemented genomic hybrid prediction in a sweet corn breeding program. The results demonstrate that implementation is feasible, even though we observed a large amount of variation in the phenotypic datasets. The amount of variation observed for trait heritabilities and correlations between traits led to variation in prediction accuracy for the traits. Overall, the RKHS models outperformed GBLUP models in all cross-validations schemes. In addition, models accounting for additive plus dominance kernels presented a slight improvement for some scenarios.

Trait heritability, significance, and correlations

The LRT indicates that the genotypic effect was significant for all traits, but CUR and SOL (FL21) and FL (WI20) **(Supplementary material – Table S2).** In addition, the block effect was not significant for most traits **(Supplementary material – Table S3)**. Trait heritabilities varied across environments. For traits from the vegetative group, the heritabilities varied from 0.28 (EH in CA21) to 0.82 (DTP in WI20). The ear-related traits presented heritabilities that varied from 0.12 (CUR in FL21) to 0.80 (EL in CA20). Whereas the values for consumer-related traits varied in the range of 0.05-0.58 (FL in WI20 and CR in WI21, respectively). A summary of trait information for all six environments can be found in **Supplementary material – Tables S1-S6**.

The correlation between vectors of pairs of trait BLUEs also demonstrates a wide range of variation (**Supplementary material – Figures S2-S7**). From the total number of pairs of correlations for all traits in the six environments, 151 out of 316 (47.8%) were significant at 10%, representing the presence of a significant linear association with each other. Large values of correlation were found between traits from the same group. For instance, DTP-DTS presented a correlation of 0.93 and 0.94 (WIS20 and WIS21, respectively). This was also observed for PH-EH at 0.72 and 0.86 (WIS20 and CAL21, respectively). The consumer-related traits presented moderated values of correlation, from 0.11 (CR-TXT in WIS20) to 0.77 (RT-TXT in WIS21). In addition, some negative correlation values were reported, such as TPF-EL in CA20 (-0.53) and DTP-STD (-0.6 in CA21).

Within-year hybrid accuracy (CV1)

We compared the prediction of all models using the CV1 approach within each site, for all traits for the year of 2020 (CA20, FL20, and WI20). Means for the repetitions showed that RKHS overperformed GBLUP in all three locations (7.35% higher in California, 1.02% higher in Florida, and 14.25% higher in Wisconsin) (**Supplementary material - Figure S7-S9**). On

the other hand, models accounting for multi-traits reduced trait predictability by 6.0% for CAL, 1.13% for FLO, and 1.4% for WIS. The inclusion of dominance in the model (AD models), were outperformed by models with only additive effects (7.74% at CAL, 0.16% at FLO, 10.53% at WIS).

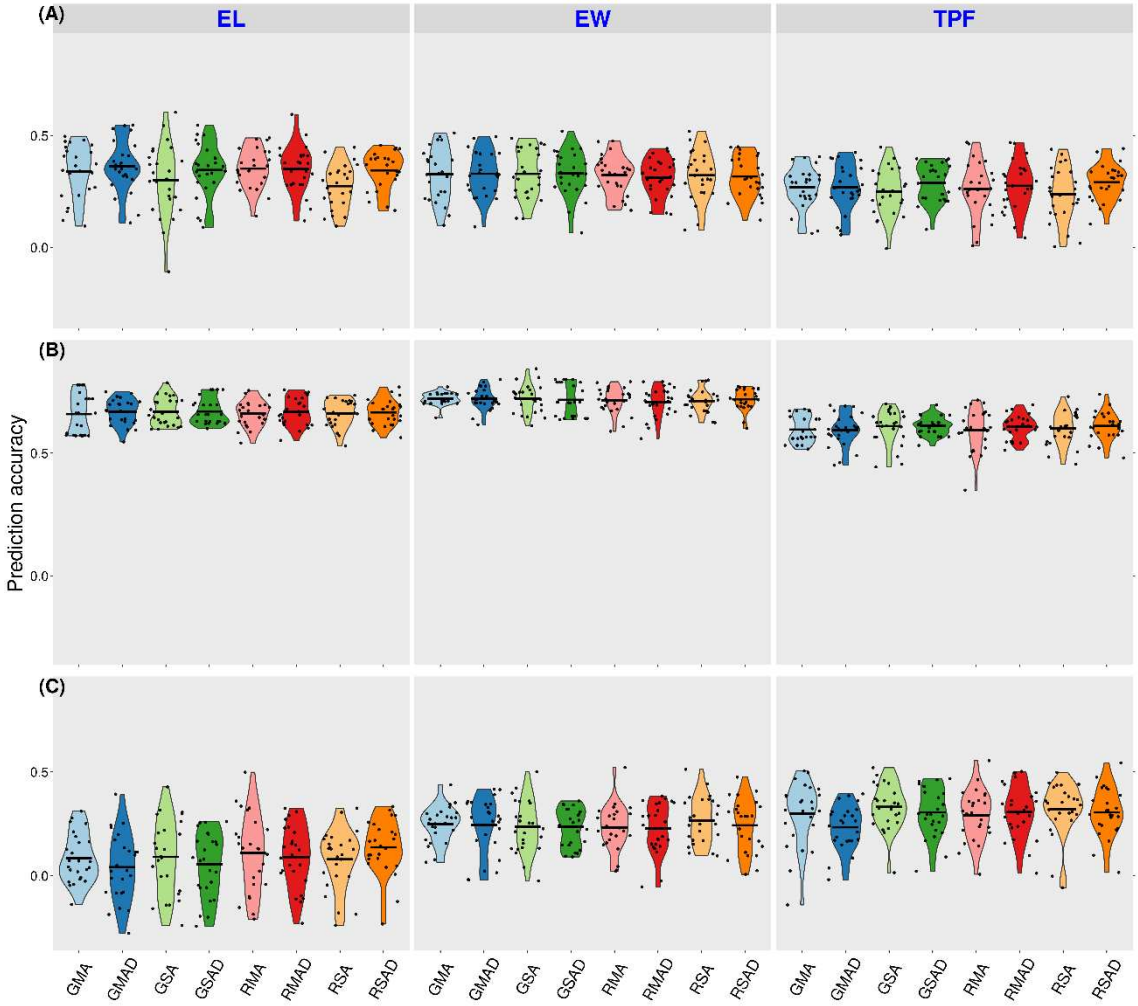


Figure 2 - **Accuracy for prediction of tested hybrids in tested environments (CV1).** GMA: GBLUP multi-trait with additive effect. GMAD: GBLUP multi-trait with additive effect and dominance effect. GSA: GBLUP single-trait with additive effect. GSAD: GBLUP single-trait with additive effect and dominance effect. RMA: RKHS multi-trait with additive effect. RMAD: RKHS multi-trait with additive effect and dominance effect. RSA: RKHS single-trait with additive effect. RSAD: RKHS single-trait with additive effect and dominance effect. (A) California site. (B) Florida site. (C) Wisconsin site. EL = ear length. EW = ear width. TPF = tip-fill.

Across-year hybrid accuracy (CV0)

The results for a model's performance under across-year cross validation scheme are summarized in **Table 1**. The RKHS performed better than GBLUP in the average for all three sites (16%, 34%, 15%, for CA, FL, and WI, respectively). Single trait models were outperformed by the multi-trait model, in average only in CAL (6%) and WIS (4%). In the end,

AD models have higher prediction accuracies than A models alone (13%, 29%, and 14% at CA, FL, and WI, respectively). The average prediction across models was lower at the FLO site (-0.016 for CUR to 0.54 for EL). In addition, higher accuracies were found in WI for the traits KRN (0.68) and EL (0.713) (**Figure 3**).

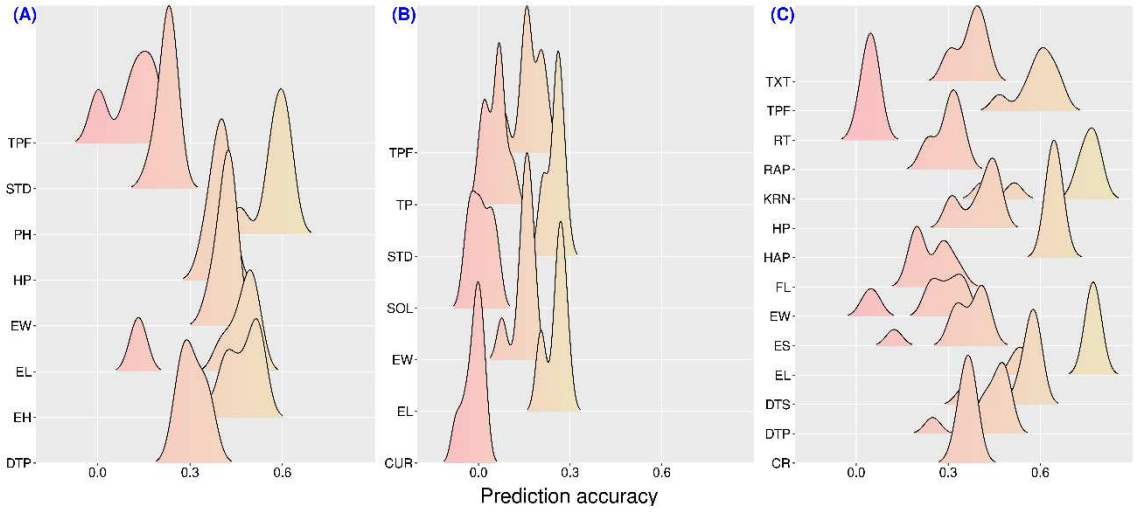


Figure 3 - Accuracy for prediction across methods for tested hybrids in new environments (CV0). (A) California. (B) Florida. (C) Wisconsin. EL: ear length. EW: ear width. TPF: tip fill. STD: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Table 1 - Prediction accuracy of across-year hybrids prediction for California, Florida, and Wisconsin. CV0 = Prediction of tested hybrids in untested environments. EL: ear length. EW: ear width. TPF: tip fill. STD: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Trait	Model							
	GSA	GSAD	GMA	GMAD	RSA	RSAD	RMA	RMAD
California								
DTP	**0.367**	0.353	0.286	0.259	0.337	0.300	0.298	0.268
EH	0.323	0.411	0.489	0.522	0.417	0.455	**0.525**	0.522
EL	0.132	0.400	0.134	0.483	0.446	0.488	0.507	**0.518**
EW	0.393	0.422	0.366	0.423	**0.438**	0.427	0.431	0.431
HP	0.378	0.419	0.344	0.417	0.389	0.382	**0.419**	0.415
PH	0.461	0.569	0.575	0.608	0.568	0.604	0.611	**0.622**
STD	0.225	**0.249**	0.176	0.223	0.242	0.227	0.227	0.240
TPF	-0.002	0.092	0.008	0.158	0.121	0.136	0.184	**0.191**
Florida								
CUR	0.015	-0.005	-0.074	-0.044	-0.005	-0.025	-0.005	**0.018**
EL	0.200	**0.261**	0.209	0.265	0.274	0.293	0.252	0.278
EW	**0.158**	0.170	0.076	0.152	0.185	0.170	0.149	0.137
SOL	-0.043	-0.020	-0.025	0.064	0.004	0.009	0.034	**0.045**
STD	0.207	0.253	0.217	0.256	0.255	**0.272**	0.262	0.284
TP	0.112	0.064	0.019	0.006	0.077	0.068	0.030	**0.064**
TPF	**0.225**	0.142	0.203	0.083	0.201	0.164	0.162	0.157
Wisconsin								
CR	0.373	**0.386**	0.330	0.357	0.380	0.372	0.351	0.355
DTP	0.249	0.422	0.471	0.480	0.399	0.446	0.488	**0.490**
DTS	0.350	0.568	0.565	0.580	0.533	0.586	**0.588**	0.587
EL	0.549	0.763	0.520	**0.785**	0.768	0.771	0.768	0.782
ES	0.123	0.322	0.322	0.408	0.353	0.398	0.409	**0.431**
EW	0.033	0.239	0.063	0.345	0.248	0.287	0.315	**0.356**
FLA	**0.336**	0.271	0.180	0.200	0.300	0.280	0.200	0.207

HAP	0.639	0.638	0.656	0.650	0.651	0.630	**0.669**	0.634
HP	0.310	0.384	0.315	0.441	0.415	0.450	0.446	**0.464**
KRN	0.515	0.773	0.409	0.728	0.773	**0.793**	0.737	0.769
RAP	0.227	0.300	0.242	0.302	0.316	0.342	0.314	**0.343**
RT	0.040	0.014	**0.071**	0.057	0.027	0.041	0.052	0.059
TPF	0.467	0.550	0.595	0.623	0.586	0.612	0.648	**0.667**
TXT	0.312	**0.424**	0.299	0.388	0.397	0.414	0.363	0.384

GMAD: GBLUP multi-trait with additive effect and dominance effect. GSA: GBLUP single-trait with additive effect. GSAD: GBLUP single-trait with additive effect and dominance effect. RMA: RKHS multi-trait with additive effect. RMAD: RKHS multi-trait with additive effect and dominance effect. RSA: RKHS single-trait with additive effect. RSAD: RKHS single-trait with additive effect and dominance effect.

Across-year: new environments and new hybrids predictions (CV00)

In the CV00 prediction (untested hybrids in untested environments), RKHS models were superior to GBLUP in CA and WI (16.12% and 7.6%, respectively), while it was outperformed at FLO (-1.3%) (**Table 2**). However, multi-trait models performed only slightly better in FLO (0.77%), whereas the AD model A overperformed the A model in WI (4.58%). The best mean predictability was for the trait HAP in WI site (0.410), whereas negative values were found for TPF in CA (-0.19) and EW in WI (-0.28) (**Figure 4**).

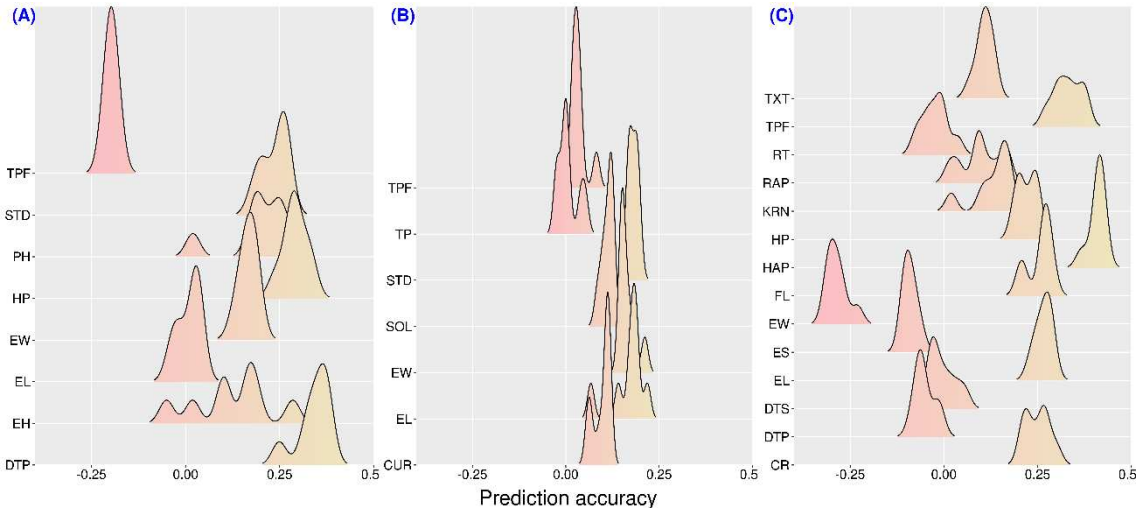


Figure 4 – Accuracy for prediction across-models of new hybrids in new environments (CV00). (A) California, (B) Florida, (C) Wisconsin, EL: ear length. EW: ear width. TPF: tip fill. STD: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Table 2 - Prediction accuracy of across-year hybrids prediction for California, Florida, and Wisconsin. CV00 = prediction of new hybrids in new environments. EL: ear length. EW: ear width. TPF: tip fill. STD: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Trait	Model							
	GSA	GSAD	GMA	GMAD	RSA	RSAD	RMA	RMAD
				California				
DTP	**0.381**	0.375	0.249	0.315	0.366	0.363	0.335	0.342
EH	0.165	0.019	**0.285**	-0.051	0.102	0.101	0.189	0.170
EL	0.029	-0.010	-0.038	-0.026	0.031	**0.036**	0.030	0.021
EW	0.130	0.183	**0.191**	0.155	0.152	0.182	0.162	0.178
HP	0.326	**0.334**	**0.270**	0.238	0.289	0.299	0.280	0.291
PH	0.228	0.174	**0.297**	0.020	0.193	0.195	0.257	0.250
STD	0.245	0.252	0.180	0.211	0.268	**0.271**	0.202	0.263
TPF	-0.190	-0.214	-0.194	-0.204	-0.203	-0.197	**-0.181**	-0.195
				Florida				
CUR	0.036	0.038	**0.082**	0.033	0.023	0.017	0.020	0.027
EL	0.176	0.154	**0.212**	0.166	0.148	0.152	0.150	0.150
EW	0.121	0.125	0.085	0.121	**0.126**	0.104	0.116	0.100
SOL	-0.023	-0.003	0.040	**0.053**	0.001	-0.024	0.006	-0.002
STD	0.173	0.167	0.190	0.172	0.168	0.187	0.190	**0.197**
TP	0.115	0.090	0.060	0.066	**0.116**	0.113	0.109	0.111
TPF	**0.218**	0.141	0.189	0.067	0.183	0.188	0.177	0.171
				Wisconsin				
CR	**0.301**	0.273	0.209	0.217	0.265	0.268	0.222	0.242
DTP	-0.018	-0.047	**-0.009**	-0.066	-0.058	-0.086	-0.064	-0.065
DTS	**0.057**	0.032	-0.005	-0.026	0.011	-0.024	-0.034	-0.040
EL	0.277	0.280	0.231	0.266	0.248	**0.291**	0.255	0.283
ES	-0.109	-0.093	**-0.061**	-0.078	-0.105	-0.100	-0.081	-0.099
EW	-0.308	-0.301	-0.298	-0.287	-0.276	**-0.230**	-0.314	-0.269
FL	0.262	0.205	0.278	0.273	0.211	0.264	0.273	**0.291**
HAP	0.405	0.407	0.368	0.420	0.425	0.415	0.415	**0.426**

HP	0.198	0.205	0.207	0.187	0.240	0.249	0.237	**0.251**
KRN	0.158	0.160	0.020	0.146	**0.180**	0.171	0.101	0.122
RAP	**0.132**	0.093	0.039	0.014	0.164	0.170	0.100	0.088
RT	-0.076	**-0.036**	-0.007	0.037	-0.057	-0.034	-0.006	-0.006
TPF	0.377	0.306	0.302	0.272	0.353	**0.375**	0.326	0.335
TXT	0.110	0.123	0.098	0.127	0.106	**0.135**	0.070	0.100

GMAD: GBLUP multi-trait with additive effect and dominance effect. GSA: GBLUP single-trait with additive effect. GSAD: GBLUP single-trait with additive effect and dominance effect. RMA: RKHS multi-trait with additive effect. RMAD: RKHS multi-trait with additive effect and dominance effect. RSA: RKHS single-trait with additive effect. RSAD: RKHS single-trait with additive effect and dominance effect.

Across-site hybrid accuracy for new environments (CV0)

In this step we compare the information across all locations to train the model in a way to predict each of the locations in 2021 (CA21, FLO21, and WI21). The models MADG (0.62, EL trait), MAG (0.05, EW trait), and MADK (0.63, TPF trait) presented the best results for predicting the CA21 environment (**Table 3 and Supplementary material – Table S7**). In addition, accuracies of 0.28 (EL), 0.30 (EW), and 0.19 (TPF) were estimated by the models ADK, ADK, and MADK, respectively for FLO. Lastly, the models MADG (0.64, EL), MAK (0.10, EW), and ADK (0.61, TPF), perform the best considering the prediction of WI. As can be observed from these results, no one model is superior at trait prediction over another.

Table 3 - Prediction accuracy of across-sites hybrids prediction for California (CA), Florida (FLO), and Wisconsin (WI) for the traits ear length (EL), ear width (EW), and tip fill (TPF). Here, all the information from 2020 (CA20, FLO20, and WI20) was used to training the model. The cross-validation scheme was the CV0 (tested hybrids in untested environments).

Sites	A.G	AD.G	MA.G	MAD.G	A.K	AD.K	MA.K	MAD.K
				EL				
CA	0.237	0.601	0.486	**0.624**	0.592	0.608	0.602	0.581
FLO	0.217	0.223	0.241	0.222	0.276	**0.281**	0.262	0.271
WI	0.223	0.614	0.526	**0.638**	0.604	0.619	0.613	0.589
				EW				
CA	-0.023	0.029	**0.047**	0.034	0.016	0.020	0.036	0.011
FLO	0.175	0.279	0.137	0.269	0.302	**0.307**	0.289	0.293
WI	-0.039	0.092	0.051	0.096	0.081	0.089	**0.097**	0.072
				TPF				
CA	0.141	0.566	0.357	0.569	0.629	0.636	0.611	**0.630**
FLO	0.122	0.188	0.161	0.177	0.186	0.186	0.190	**0.198**
WI	0.140	0.535	0.384	0.531	0.603	**0.609**	0.583	0.601

Across-site newly hybrids for new environments accuracy (CV00)

The prediction of new hybrids in new environments varied for each trait and model used. In the prediction of CA21, accuracies from the different models ranged from 0.016 to 0.156 for EL, from 0.409 to 0.501 for EW, and from 0.190 to 0.292 for TPF **(Table 4 and Supplementary material – Table S8).** When we consider the prediction of FLO21, the accuracies where lower, ranging from -0.073 to -0.001 for EL, -0.065 to 0.048 for EW and -0.058 to 0.152 for TPF. For the prediction of WI21, the values varied from -0.089 to 0.024 (EL), from 0.058 to 0.191 (EW), and from 0.224 to 0.280 (TPF). The prediction of CA presented the best performance for all the three locations, while TPF presented a higher

predictive ability compared to EW and EL (0.209 vs 0.201 and 0.021 for EW and EL, respectively).

Table 4 - Prediction accuracy of across-sites hybrids prediction for the sites of California (CA), Florida (FLO), and Wisconsin (WI) for the traits ear length (EL), ear width (EW), and tip fill (TPF). Here, only the genotypes that were not assessed at the testing site were included in the training set from 2020 sites (CA20, FLO20, and WI20). The cross-validation scheme was the CV00 (untested hybrids in untested environments).

Sites	AG	ADG	MAG	MADG	AK	ADK	MAK	MADK
				EL				
CA	0.109	**0.156**	0.016	0.123	0.148	0.156	0.114	0.139
FLO	-0.016	-0.062	-0.073	-0.058	-0.010	-0.007	-0.035	**-0.001**
WI	-0.022	0.008	-0.089	-0.067	**0.024**	0.015	-0.033	-0.023
				EW				
CA	**0.501**	0.450	0.417	0.409	0.470	0.476	0.453	0.452
FLO	0.043	-0.065	0.034	0.013	0.037	0.038	**0.048**	0.018
WI	0.058	0.154	0.017	0.131	**0.191**	0.169	0.147	0.154
				TPF				
CA	0.190	0.274	0.208	0.275	0.256	0.255	0.281	**0.292**
FLO	0.151	-0.058	0.143	**0.152**	0.145	0.143	**0.152**	0.135
WI	0.232	0.225	0.241	0.224	0.273	0.267	0.275	**0.280**

4. Discussion

As the field of quantitative genomics has evolved over the last years, tools for genomic selection have gained popularity in breeding programs. It has allowed breeders to predict new genotypes and their performance in untested environments. As we know, genomic selection has been proposed for boosting different stages of a breeding program (Allier et al., 2019; Oliveira et al., 2020; Powell et al., 2020). For hybrid crops, the prediction of the potential crosses (hybrids) from genotyped parents represents an important tool that can ensure good candidates reach advanced field stages (Kadam et al., 2021; Oliveira et al., 2022). Routinely, sweet corn breeders assessed hybrids performance in several environments, and they need to keep in the consideration a set of target traits for crop production. There is still a challenge with the implementation of genomic selection for in silico hybrid prediction, which comes from model selection and the implementation of cross-validations schemes to optimize advanced field stages.

Model optimization for hybrid prediction

Historically, sweet corn breeders select for several breeding targets at once, which, ultimately, implies complex decisions in the breeding program. For instance, the traits studied here present complex relationships between them, with correlations varying from high positive to low negative values. In addition, trait heritabilities vary greatly from one trait to another. This complexity can directly affect model accuracy (Montesinos-López et al., 2016). It is common knowledge that, in multi-trait models, traits with lower heritabilities can borrow information from traits with high heritabilities once they are correlated with one another leading to an improve of accuracy (Jia and Jannink, 2012; Mrode, 2014). However, our results indicated that little benefit is gained from MTM over STM models, being STM gives a better performance in some scenarios. These results are not similar to what has been reported for some genomic prediction models, where MTM is slightly or even significantly better (Covarrubias-Pazaran et

al., 2018; Oliveira et al., 2020; Sandhu et al., 2022). The relationship between traits (meaning correlations), did not facilitate the accuracy improvement.

Our results indicate that the addition of non-additive effects in the model has the potential to increase the model accuracy, up to 29% for untested environments (CV0) in the across-year validation scheme. For hybrid prediction in tropical field corn, it is standard that non-additive effects play an important rule and can contribute to increase the model's prediction accuracy, even though it has an increased impact in some traits compared to others (Alves et al., 2019; Ferrão et al., 2020; Rogers et al., 2021). For instance, Coelho et al. (2020) estimates, for field corn in a diallel design, a dominance heritability of 0.32 (grain yield), 0.02 (EL), 0.13 (EH), and 0.05 (PH); whereas Nardino et al. (2020) demonstrate, for yield related traits, that the proportion between the GCA and SCA varies from 0.92 and 7.26 among sites. Together, the complexity and importance of dominance may vary from trait to trait and, ultimately, impact the implementation of a model with non-additive effects in hybrid prediction. On the other hand, the role dominance plays are still emerging in sweet corn genomic selection. However, early evidences suggest that non-additive effects represent an important factor in the model, that can also vary from one trait to the other (Zystro et al., 2021a, 2021b). For instance, Neto et al. (2022) indicated a large impact of the SCA estimates in F1 population performance, with a GCA/SCA ratio smaller than one unit for the traits PH, EH, EL, and KRN.

In Zystro et al. (2021a) the application of five-fold cross-validation (equivalent to our CV1 scheme) returns an increase of 50% in trait prediction for GBLUP model, comparing A/AD models. However, on average, the results of the CV1 scheme with the use of dominance, tends to reduce the prediction accuracy and was outperformed by the additive only effects models by 6%.

The results demonstrated that RKHS models can be more advantageous over conventional GBLUP for hybrid performance in all schemes presented here (CV1, CV0, and CV00). Several

authors demonstrated that RKHS models outperform linear kernel GBLUP (Cuevas et al., 2016, 2017; Bandeira e Sousa et al., 2017; Lopez-Cruz et al., 2021). This study reinforces this observation and adds sweet corn hybrid prediction to the mix. Therefore, GBLUP-based models are largely applied for hybrid prediction (Fritsche-Neto et al., 2021). The GBLUP kernels here explored in the model (A and D, for additive and dominance information) were constructed based on the numerator of the relationship matrix, as proposed by VanRaden, (2008) (for additive kernel) and on the proposal made by Vitezica et al. (2013) (dominance). However, it is possible to build non-parametric matrices to enhance predictive performance, and it is desirable to pick a kernel matrix that captures characteristics of the data. Ultimately, this would leverage prediction accuracy (as it did here). In this way, RKHS, a semi-parametric model, seems to better capture the data complexity and represent the optimum model for hybrid prediction in a sweet corn breeding program.

Implementation of CV schemes in sweet corn hybrid prediction

We compared the implementation of different cross-validations (CV) schemes (Burgueño et al., 2012; Jarquin et al., 2018) in the sweet corn breeding program. Genomic selection can contribute to a breeding program by predicting unobserved genotypes and/or environments, which, ultimately, reduces field trials and ensures that potential candidates advance through the trialing stages. The CV1 scheme precludes the information of untested hybrids in tested environments. It resulted in higher prediction accuracies once all the information is made available. The prediction performance of CV1 is as high as the genetic correlation between the training set and the prediction set (Jarquin et al., 2018). As the set of hybrids came from 61 individual lines, the relatedness in the CV partition boosted the prediction accuracy of the CV1 scheme. In addition, FLO site had the highest performance, outperforming CAL and WIS sites, on average. In this case, the number of hybrids assessed in the 2020 season in FLO was significantly higher than those in CAL and WIS (418, 246, and 236 at FLO, CAL, and WIS

sites, respectively) with the same set of parents, which could have built a stronger correlation between training and prediction sets, thereby, increasing the performance of the model prediction.

In a sweet corn breeding program, hybrid combinations are used in three different and sequential test cross steps (Peixoto et al. 2022). First, new hybrids are generated from the crossing of a large number of advanced lines against a few testers and assessed in a few environments. The parents of the best performing hybrids are selected based on the general and specific combining abilities (GCA and SCA, respectively), and then proceed to next stage where they are crossed to a larger number of testers and assessed across several environments. The selection of the best two parental is made based on GCA and SCA and these two lines are crossed to even more testers and planted in large environment trials to evaluate their performance.

For instance, the CV00 scheme prediction can aid the in-silico prediction of the best hybrids combinations for the first stage of testcrossing. Ultimately, it can guarantee that the best hybrids combinations are included in the first advanced trials. However, CV00 represents the hardest hybrid set to predict by lacking both information, regarding newly hybrids and environments (Burgueño et al., 2012; Jarquin et al., 2018). Furthermore, it represents the lowest accuracy compared with other schemes (Jarquin et al., 2018; Vieira et al., 2022). Then, CV00 prediction for sweet corn hybrids reinforces those hypotheses, once the average accuracy was lower than the other schemes, whereas, for some traits, the predictability achieved good levels of accuracy (*i.e.*, DTP in CAL site: 0.38, EW in FLO site: 0.38, HP in WIS site: 0.42).

The CV0 scheme, can predict how the genetic material will perform in another environment or in a target population of environments based on known performance in a tested location. For the sweet corn breeding program, this can help predict how a hybrid will performance in the second and third filed hybrid stages. Compared to CV00, the CV0 returned the higher values

of prediction accuracy, somehow expected once the genetic material in the CV0 (hybrids) was already assessed in the site in the year before.

The three stages of hybrid prediction deal with the disturbance factors caused by the presence of the G×E interaction. Dealing with the G×E interaction presents a big challenge in hybrid prediction (Schrag et al., 2019; Rogers et al., 2021). This could be observed in our results where the across-site prediction, the results demonstrated the drop in accuracy for the scheme CV0 and CV00 compared with the across-year prediction. In advanced field stages, the presence of G×E interaction can cause differences in field evaluation and, ultimately in the genotype ranking, which impacts the composition of the selection set.

Considering only the across-site prediction, when the information of all sites was used in the training set, the accuracy of the prediction increased. The inclusion of this type of data represents a way to circumvent the G×E interaction in hybrid prediction in sweet corn.

Future directions

This was the first effort to increase the knowledge in hybrid performance through genomic prediction in a multi-trait framework and using different kernels structures in sweet corn. For instance, some traits, such as flavor, color rating, and texture, are related with market preferences. Traits like those demand more systematic approaches to enhance based on not only genetic performance, but also consumer-personal preference. For such, we should use available tools to guide breeding strategy targets. Future directions for enhance consumer-related traits can be done with the application of algorithms to predict consumer preferences based on genomics and also metabolites data (Colantonio et al., 2022), which can bring a new view on consumers inclinations, such as flavor perception. Alternatively, the use of near-infrared technology can add in the inference of the relationship of the lines, allowing for phenomic selection (Krause et al., 2019).

Also, aiming to optimize the testcross stages, sparse testing designs should be considered (Jarquin et al., 2020; Crespo-Herrera et al., 2021). Those authors described that the methodology has the potential to save a substantial number of resources by optimizing the genotypes and environments explored in the trials, by accounting for the G×E interaction. Thus, we can enhance the sweet corn program by considering the sparse designs. On the other hand, several tools could be used to increase the model prediction. First, we can leverage the quality of the phenotypic data collected in the testcross stages and, consequently, increase the efficiency of the genomic models by used high-throughput phenotyping tools (Persa et al., 2021).

Also, hybrid prediction can take advantage of the environmental covariables to build covariance relationship matrix among trials (Bandeira e Sousa et al., 2017; Gevartosky et al., 2023). In addition, those environmental kernels seem to be dependent of germplasm, environment, and traits (Costa-Neto et al., 2021; Rogers and Holland, 2022). Another good alternative of application is the crop growth models (Cooper and Messina, 2021), which deals with the G×E interaction, largely identified in the testing phase of the sweet corn program.

5. Conclusion

Our study presents a comprehensive and detailed application of genomic selection as a tool for hybrid prediction in sweet corn breeding program, exploring several traits of interest for the first time in sweet corn. In summary, the models with multi-trait inference did not present superior results in all scenarios, whereas models with additive-dominance presented at least a slightly improved performance in many the traits. The most prominent inference was for RKHS models, that outperform GBLUP in all scenarios and is recommended as a standard model for sweet corn prediction. In addition, hybrid prediction through genomic selection has great potential to impact sweet corn breeding and different CV schemes can be used in different stages of the breeding program.

Acknowledgements

This work was supported by the National Institute of Food and Agriculture SCRI 2018-51181-28419 and AFRI 2019-05410 to M.F.R.R.). We also thank the financial support from the Brazilian Government through the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) through the CAPES-PrInt scholarship. This study was financed in part by the CAPES - Finance Code 001.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Table S1. Traits measured in each of the six environments here assessed. FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. CA21: California site, 2021. WI20: Wisconsin site, 2020. WI21: Wisconsin site, 2021.

Table S2. Likelihood ratio test for the genotypic effect for the traits assessed in six environments for the individual analyses. FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. CA21: California site, 2021. WI20: Wisconsin site, 2020. WI21: Wisconsin site, 2021. ***** Significative and **^{na}** non-significative by LRT test at 5% of probability and 1 degree of freedom.

References

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., and Lehermeier, C. (2019). Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. *G3 Genes, Genomes, Genet.* 9, 1469–1479. doi: 10.1534/g3.119.400129.

Alves, F. C., Granato, Í. S. C., Galli, G., Lyra, D. H., Fritsche-Neto, R., and de los Campos, G. (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* 15, 14. doi: 10.1186/s13007-019-0388-x.

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

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

Coelho, I. F., Alves, R. S., Rocha, J. R. A. S. C., Peixoto, M. A., Teodoro, L. P. R., Teodoro, P. E., et al. (2020). Multi-trait multi-environment diallel analyses for maize breeding. *Euphytica* 216, 1–17. doi: 10.1007/s10681-020-02677-9.

Colantonio, V., Ferrão, L. F. V., Tieman, D. M., Bliznyuk, N., Sims, C., Klee, H. J., et al. (2022). Metabolomic selection for enhanced fruit flavor. *PNAS* 119, 1–11. doi: 10.1073/pnas.2115865119/-/DCSupplemental.Published.

Cooper, M., and Messina, C. D. (2021). Can We Harness “Enviromics” to Accelerate Crop Improvement by Integrating Breeding and Agronomy? *Front. Plant Sci.* 12, 1932. doi: 10.3389/fpls.2021.735143.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021). Nonlinear kernels, dominance, and

envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity (Edinb).* 126, 92–106. doi: 10.1038/s41437-020-00353-1.

Covarrubias-Pazaran, G., Schlautman, B., Diaz-Garcia, L., Grygleski, E., Polashock, J., Johnson-Cicalese, J., et al. (2018). Multivariate GBLUP improves accuracy of genomic selection for yield and fruit weight in biparental populations of vaccinium macrocarpon Ait. *Front. Plant Sci.* 9, 1–13. doi: 10.3389/fpls.2018.01310.

Cowling, W. A., Gaynor, R. C., Antolín, R., Gorjanc, G., Edwards, S. M., Powell, O., et al. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Sci. Rep.* 10, 1–8. doi: 10.1038/s41598-020-61031-0.

Crespo-Herrera, L., Howard, R., Piepho, H., Pérez-Rodríguez, P., Montesinos-Lopez, O., Burgueño, J., et al. (2021). Genome-enabled prediction for sparse testing in multi-environmental wheat trials. *Plant Genome* 14. doi: 10.1002/tpg2.20151.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian Genomic Prediction with Genotype × Environment Interaction Kernel Models. *G3 Genes|Genomes|Genetics* 7, 41–53. doi: 10.1534/g3.116.035584.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., Campos, G. de los, et al. (2016). Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. *Plant Genome* 9, 1–20. doi: 10.3835/plantgenome2016.03.0024.

Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 Genes, Genomes, Genet.* 9, 2913–2924. doi: 10.1534/g3.119.400493.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011).

The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

De Los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res. (Camb).* 92, 295–308. doi: 10.1017/S0016672310000285.

Ferrão, L. F. V., Marinho, C. D., Munoz, P. R., and Resende, M. F. R. (2020). Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. *Crop Sci.* 60, 666–677. doi: 10.1002/csc2.20096.

Fritsche-Neto, R., Galli, G., Borges, K. L. R., Costa-Neto, G., Alves, F. C., Sabadin, F., et al. (2021). Optimizing Genomic-Enabled Prediction in Small-Scale Maize Hybrid Breeding Programs: A Roadmap Review. *Front. Plant Sci.* 12, 1–16. doi: 10.3389/fpls.2021.658267.

Gevartosky, R., Carvalho, H. F., Costa-Neto, G., Montesinos-López, O. A., Crossa, J., and Fritsche-Neto, R. (2023). Enviromic-based kernels may optimize resource allocation with multi-trait multi-environment genomic prediction for tropical maize. *BMC Plant Biol.* 23, 1–20. doi: 10.1186/s12870-022-03975-1.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S., and Thompson, R. (2015). ASReml user guide release 4.1 structural specification. *Hemel hempstead VSN Int. ltd*.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi: 10.1007/s00122-013-2243-1.

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes, Genomes, Genet.* 10, 2725–2739. doi: 10.1534/g3.120.401349.

Jarquin, D., Howard, R., Xavier, A., and Das Choudhury, S. (2018). Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans. *Agronomy* 8, 51. doi: 10.3390/agronomy8040051.

Jia, Y., and Jannink, J. (2012). Multiple-Trait Genomic Selection Methods Increase genetic valu prediction accuracy. *Genetics* 192, 1513–1522. doi: 10.1534/genetics.112.144246.

Kadam, D. C., Rodriguez, O. R., and Lorenz, A. J. (2021). Optimization of training sets for genomic prediction of early-stage single crosses in maize. *Theor. Appl. Genet.* 134, 687–699. doi: 10.1007/s00122-020-03722-w.

Krause, M. D., Dias, K. O. G., Pedroso Rigal dos Santos, J., Oliveira, A. A., Guimarães, L. J. M., Pastina, M. M., et al. (2020). Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Sci.* 60, 3049–3065. doi: 10.1002/csc2.20253.

Krause, M. R., González-Pérez, L., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O., Singh, R. P., et al. (2019). Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. *G3 Genes, Genomes, Genet.* 9, 1231–1247. doi: 10.1534/g3.118.200856.

Lopez-Cruz, M., Beyene, Y., Gowda, M., Crossa, J., Pérez-Rodríguez, P., and de los Campos, G. (2021). Multi-generation genomic prediction of maize yield using parametric and non-parametric sparse selection indices. *Heredity (Edinb).*, 1–10. doi: 10.1038/s41437-021-00474-1.

Marinho, C. D., Coelho, I. F., Peixoto, M. A., Carvalho Júnior, G. A., and Resende Jr, M. F. R. (2022). Genomic selection as a tool for maize cultivars development. *Rev. Bras. Milho e Sorgo* 21.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., et al. (2016). A Genomic Bayesian Multi-trait and Multi-

environment Model. *G3 Genes|Genomes|Genetics* 6, 2725–2744. doi: 10.1534/g3.116.032359.

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: A review. *Front. Genet.* 5, 1–13. doi: 10.3389/fgene.2014.00363.

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values*. Cabi.

Nardino, M., Barros, W. S., Olivoto, T., Cruz, C. D., Silva, F. F., Pelegrin, A. J., et al. (2020). Multivariate diallel analysis by factor analysis for establish mega-traits. *An. Acad. Bras. Cienc.* 92, 1–19. doi: 10.1590/0001-3765202020180874.

Neto, I. L. S., Figueiredo, A. S. T., Uhdre, R. S., Contreras-Soto, R. I., Scapim, C. A., and Zanotto, M. D. (2022). Combining ability and heterotic pattern in relation to F1 performance of tropical and temperate-adapted sweet corn lines. *Bragantia* 81, 1–12. doi: 10.1590/1678-4499.20220056.

Oliveira, A. A., Resende, M. F. R., Ferrão, L. F. V., Amadeu, R. R., Guimarães, L. J. M., Guimarães, C. T., et al. (2020). Genomic prediction applied to multiple traits and environments in second season maize hybrids. *Heredity (Edinb).* 125, 60–72. doi: 10.1038/s41437-020-0321-0.

Oliveira, I. C. M., Bernardeli, A., Guilhen, J. H. S., and Pastina, M. M. (2022). "Genomic Prediction of Complex Traits in an Allogamous Annual Crop: The Case of Maize Single-Cross Hybrids," in *Complex Trait Prediction* (Springer), 543–567. doi: doi.org/10.1007/978-1-0716-2205-6_20.

Pérez-Rodríguez, P., and de los Campos, G. (2022). Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. *Genetics* 222. doi: 10.1093/genetics/iyac112.

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442.

Persa, R., Ribeiro, P. C. O., and Jarquin, D. (2021). The use of high-throughput phenotyping in genomic selection context. *Crop Breed. Appl. Biotechnol.* 21, 1–11. doi: 10.1590/1984-70332021v21Sa19.

Powell, O., Gaynor, R. C., Gorjanc, G., Werner, C. R., and Hickey, J. M. (2020). A two-part strategy using genomic selection in hybrid crop breeding programs. *bioRxiv*, 1–46. doi: 10.1101/2020.05.24.113258.

Rao, C. R. (1952). *Advanced statistical methods in biometric research*. First. New York, USA: A Division Of Macmillan Publishing Co, Inc New York.

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 Genes, Genomes, Genet.* 11, jkaa050. doi: 10.1093/g3journal/jkaa050.

Rogers, A. R., and Holland, J. B. (2022). Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3 Genes, Genomes, Genet.* 12. doi: 10.1093/g3journal/jkab440.

Sandhu, K. S., Patil, S. S., Aoun, M., and Carter, A. H. (2022). Multi-Trait Multi-Environment Genomic Prediction for End-Use Quality Traits in Winter Wheat. *Front. Genet.* 13, 1–14. doi: 10.3389/fgene.2022.831020.

Schrag, T. A., Schipprack, W., and Melchinger, A. E. (2019). Across-years prediction of hybrid performance in maize using genomics. *Theor. Appl. Genet.* 132, 933–946. doi: 10.1007/s00122-018-3249-5.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980.

Vieira, C. C., Persa, R., Chen, P., and Jarquin, D. (2022). Incorporation of Soil-Derived Covariates in Progeny Testing and Line Selection to Enhance Genomic Prediction

Accuracy in Soybean Breeding. *Front. Genet.* 13, 1–15. doi: 10.3389/fgene.2022.905824.

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi: 10.1534/genetics.113.155176.

Zystro, J., Peters, T., Miller, K., and Tracy, W. F. (2021a). Classical and genomic prediction of hybrid sweet corn performance in organic environments. *Crop Sci.* 61, 1698–1708. doi: 10.1002/csc2.20400.

Zystro, J., Peters, T., Miller, K., and Tracy, W. F. (2021b). Classical and genomic prediction of synthetic open-pollinated sweet corn performance in organic environments. *Crop Sci.* 61, 3382–3391. doi: 10.1002/csc2.20531.

Supplementary material

Genomic prediction of hybrid performance for boosting sweet corn breeding program

Marco Antônio Peixoto[1,2], Kristen A. Leach[2], Diego Jarquin[3], Patrick Flannery[4], William F. Tracy[4], Jared Zystro[5], Leonardo Lopes Bhering[1], Márcio F. R. Resende[2,*]

[1] Universidade Federal de Viçosa, Laboratório de Biometria, Viçosa, Minas Gerais, Brazil

[2] Department of Horticultural sciences, University of Florida, Gainesville, Florida, United States

[3] Department of Agronomy, University of Florida, Gainesville, Florida, United States

[4] Department of Agronomy, University of Wisconsin-Madison, Madison, Wisconsin, United States

[5] Organic seed alliance

*** Corresponding author**

Table S1. Traits measured in each of the six environments here assessed. FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. CA21: California site, 2021. WI20: Wisconsin site, 2020. WI21: Wisconsin site, 2021.

	Environments					
Traits	**FL20**	**FL21**	**CA20**	**CA21**	**WI20**	**WI21**
EL	√	√	√	√	√	√
EW	√	√	√	√	√	√
TPF	√	√	√	√	√	√
STAND	√	√	√	√	-	-
DTP	-	-	√	√	√	√
HP	-	-	√	√	√	√
KRN	√	-	-	-	√	√
PH	-	-	√	√	√	-
EH	-	-	√	√	√	-
SOL	√	√	-	-	-	-
TP	√	√	-	-	-	-
CUR	√	√	-	-	-	-
DTS	-	-	-	-	√	√
HAP	-	-	-	-	√	√
RAP	-	-	-	-	√	√
ES	-	-	-	-	√	√
CR	-	-	-	-	√	√
FL	-	-	-	-	√	√
TXT	-	-	-	-	√	√
RT	-	-	-	-	√	√

√: indicates that the trait was measured at the respective site. EL: ear length. EW: ear width. TPF: tip fill. STAND: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Table S2. Likelihood ratio test for the genotypic effect for the traits assessed in six environments for the individual analyses. FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. CA21: California site, 2021. WI20: Wisconsin site, 2020. WI21: Wisconsin site, 2021. ***** Significative and **[na]** non-significative by LRT test at 5% of probability and 1 degree of freedom.

LRT test for genotypic effect						
Traits	**FL20**	**FL21**	**CA20**	**CA21**	**WI20**	**WI21**
CR	-	-	-	-	0.2073*	0.4036*
CUR	9.4606*	27.5076 [na]	-	-	-	-
DTP	-	-	0.5445*	3.7875*	4.1053*	4.1928*
DTS	-	-	-	-	4.1662*	5.8363*
EH	-	-	145.6411*	21.7719*	104.9425*	-
EL	0.9345*	8.8201*	3.2375*	1.8987*	1.8788*	1.2911*
ES	-	-	-	-	0.3511*	0.2238*
EW	0.0344*	0.0576*	0.0718*	0.0672*	0.0412*	0.0495*
FL	-	-	-	-	0.0188 [na]	0.0784*
HAP	-	-	-	-	0.3368*	0.2411*
HP	-	-	0.3246*	0.3350*	1.0397*	0.6585*
KRN	1.9737*	-	-	-	2.9495*	2.7128*
PH	-	-	320.2704*	96.2332*	223.4988*	-
RAP	-	-	-	-	0.1713*	0.0968*
RT	-	-	-	-	0.2637*	0.1230*
SOL	0.0001*	0.0009 [na]	-	-	-	-
STD	127.3940*	650.5065*	187.3044*	377.6829*	-	-
TP	35.1331*	99.6853*	-	-	-	-
TPF	0.0006*	0.0004*	1.3500*	0.7114*	1.0770*	0.9107*
TXT	-	-	-	-	0.14268*	0.14385*

EL: ear length. EW: ear width. TPF: tip fill. STD: stand count. DTP: days to pollination. HP: husk protection. KRN: kernel row number. PH: plant height. EH: ear height. SOL: solidity. TP: taper. CUR: curvature. DTS: days to silking. HAP: husk appearance. RAP: Row appearance. ES: ear shape. CR: color rate. FL: flavor. TXT: texture. RT: rating.

Table S3. Variance components for and Likelihood ratio test for the block effect for the traits assessed in three environments for the individual analyses. FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. ***** Significative and **na** non-significative by LRT test at 5% of probability and 1 degree of freedom.

LRT test for block effect			
Trait	**FL20**	**FL21**	**CA20**
CUR	0.0000 na	16.7827 na	-
DTP	-	-	0.0000 na
EH	-	-	29.3757*
EL	0.0154 na	0.0388 na	0.0000 na
EW	0.0009*	0.0000 na	0.0211*
HP	-	-	0.0063 na
KRN	0.0000 na	-	-
PH	-	-	63.3606*
SOL	0.0000 na	0.0001 na	-
STD	1.4495 na	11.8974 na	1.4317 na
TP	1.4627*	14.0605 na	-
TPF	0.0000*	0.0000 na	0.0257 na

CUR: curvature. DTP: days to pollination. EH: ear height. EL: ear length. HP: husk protection. KRN: kernel row number. PH: plant height. SOL: solidity. STD: stand count. TP: taper. TPF: tip fill.

Table S4. Phenotypic mean and estimates of traits heritability and for the California site in 2020 and 2021 individual analyses.

California - 2020				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Stand count	STD	78.04	0.77	0.026
Days to pollination	DTP	33.63	0.45	0.117
Husk protection	HP	1.84	0.67	0.036
Ear width	EW	4.655	0.37	0.056
Ear length	EL	20.11	0.80	0.024
Tipfill	TPF	2.97	0.68	0.036
Plant heigh	PH	154.7	0.65	0.054
Ear heigh	EH	70.62	0.63	0.053

California - 2021				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Stand count	STD	48.5	0.78	0.053
Days to pollination	DTP	65.98	0.61	0.082
Husk protection	HP	3.086	0.62	0.083
Ear width	EW	4.06	0.30	0.109
Ear length	EL	17.16	0.60	0.083
Tipfill	TPF	3.554	0.60	0.085
Plant heigh	PH	101.4	0.46	0.098
Ear heigh	EH	39.91	0.28	0.106

Table S5. Phenotypic mean and estimates of traits heritability and for the Florida site in 2020 and 2021 individual analyses.

Florida - 2020				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Stand count	STD	88.91	0.61	0.033
Kernel row number	KRN	15.42	0.65	0.030
Ear length	EL	12.56	0.52	0.038
Ear width	EW	3.60	0.54	0.036
Solidity	SOL	0.95	0.31	0.047
Taper	TP	77.07	0.41	0.043
Curvature	CUR	17.71	0.18	0.050
Tipfill	TPF	0.94	0.42	0.041

Florida - 2021				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Stand count	STD	58.46	0.83	0.031
Tipfill	TPF	0.99	0.35	0.094
Ear length	EL	20.97	0.66	0.062
Ear width	EW	4.80	0.21	0.105
Solidity	SOL	0.82	0.16	0.097
Taper	TP	108.89	0.32	0.089
Curvature	CUR	50.40	0.04	0.097

Table S6. Phenotypic mean and estimates of traits heritability for the Wisconsin site in 2020 and 2021 individual analyses.

Wisconsin - 2020				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Days to pollination	DTP	53.27	0.82	0.021
Days to silking	DTS	55.55	0.81	0.022
Kernel row number	KRN	17.62	0.72	0.031
Husk appearence	HAP	3.839	0.30	0.060
Husk protection	HP	2.828	0.65	0.038
Ear weigth	EW	4.5	0.41	0.054
Ear length	EL	20.34	0.72	0.032
Tipfill	TPF	3.276	0.73	0.030
Row appearence	RAP	3.648	0.38	0.056
Ear shape	ES	3.539	0.49	0.050
Color rate	CR	3.309	0.39	0.060
Flavor	FL	3.551	0.05	0.065
Texture	TXT	2.843	0.31	0.059
Rating	RT	2.807	0.47	0.051
Plant heigth	PH	192.5	0.67	0.036
Ear heigth	EH	77.46	0.43	0.053

Wisconsin - 2021				
Trait	**Abbreviation**	**Phenotypic mean**	**Heritability**	**Standard error**
Days to pollination	DTP	55.17	0.74	0.060
Days to siling	DTS	57.04	0.72	0.064
Kernel row number	KRN	16.9	0.70	0.067
Husk appearence	HAP	3.81	0.24	0.106
Husk protection	HP	3.621	0.68	0.071
Ear weigth	EW	4611	0.52	0.092
Ear length	EL	19.62	0.75	0.058
Tipfill	TPF	3.791	0.69	0.070
Row appearence	RAP	3.748	0.32	0.105
Ear shape	ES	3.73	0.44	0.100
Color rate	CR	3.072	0.58	0.088
Flavor	FL	3.522	0.21	0.106

Texture	TXT	2.896	0.35	0.105
Rating	RT	3.026	0.24	0.106

Table S7. Prediction accuracy of across-site hybrids prediction for the sites of California, Florida, and Wisconsin. EL: ear length. EW: ear width. TPF: tip fill. Here, all the information from 2020 (CA20, FL20, and WI20) was used to training the model. The cross-validation scheme was the CV0 (tested hybrids in untested environments). FLWICA: training set combines CA20, FL20, and WI20 to predict CA21 site. FLCA: training set combines CA20 and FL20 to predict CA21 site. WICA: training set combines CA20 and WI20 to predict CA21 site. CAWIFL: training set combines CA20, FL20, and WI20 to predict FL21 site. CAFL: training set combines CA20 and FL20 and to predict FL21 site. WIFL: training set combines WI20 and CA20 to predict FLO21 site. CAFLWI: training set combines CA20, FL20, and WI20 to predict WI21 site. CAWI: training set combines CA20 and WI20 and to predict WI21 site. FLWI: training set combines FLO20 and WI20 to predict WI21 site.

Sites	GSA	GSAD	GMA	GMAD	RSA	RSAD	RMA	RMAD
				EL				
FLWICA	0.237	0.601	0.486	**0.624**	0.592	0.608	0.602	0.581
FLCA	-0.184	-0.109	-0.162	-0.045	-0.055	**0.006**	-0.081	-0.008
WICA	0.288	0.816	0.809	0.801	**0.833**	0.832	0.811	0.810
CAWIFL	0.217	0.223	0.241	0.222	0.276	**0.281**	0.262	0.271
CAFL	0.256	0.268	0.180	0.232	0.305	**0.318**	0.278	0.296
WIFL	0.165	0.238	0.253	0.279	0.233	0.260	**0.296**	0.280
CAFLWI	0.223	0.614	0.526	**0.638**	0.604	0.619	0.613	0.589
CAWI	0.320	**0.842**	0.848	0.839	0.650	0.653	0.667	0.666
FLWI	0.165	0.318	0.137	**0.332**	0.241	0.254	0.186	0.165
				EW				
FLWICA	-0.023	0.029	**0.047**	0.034	0.016	0.020	0.036	0.011
FLCA	0.131	0.158	0.241	0.214	0.310	**0.376**	0.276	0.272
WICA	-0.048	0.025	0.064	0.160	0.059	0.067	**0.171**	0.164
CAWIFL	0.175	0.279	0.137	0.269	0.302	**0.307**	0.289	0.293
CAFL	0.183	0.256	0.174	0.251	0.272	**0.287**	0.271	0.277
WIFL	0.199	0.207	0.190	0.211	**0.230**	0.227	0.211	0.230
CAFLWI	-0.039	0.092	0.051	0.096	0.081	0.089	**0.097**	0.072
CAWI	0.009	0.188	0.116	0.286	0.609	0.601	**0.657**	0.648
FLWI	0.085	**0.421**	0.143	0.415	0.400	0.401	0.370	0.366

				TPF				
FLWICA	0.141	0.566	0.357	0.569	0.629	0.636	0.611	**0.630**
FLCA	0.209	0.267	0.198	**0.275**	0.252	0.262	0.241	0.248
WICA	0.389	0.663	0.577	0.652	0.664	0.665	0.670	**0.677**
CAWIFL	0.122	0.188	0.161	0.177	0.186	0.186	0.190	**0.198**
CAFL	0.093	0.165	0.102	0.175	0.154	0.176	0.183	**0.196**
WIFL	0.088	0.096	0.117	**0.134**	0.113	0.110	0.117	0.125
CAFLWI	0.140	0.535	0.384	0.531	0.603	**0.609**	0.583	0.601
CAWI	0.401	0.653	0.530	0.619	0.670	0.666	**0.715**	0.715
FLWI	0.169	0.201	0.177	**0.229**	0.180	0.183	0.198	0.227

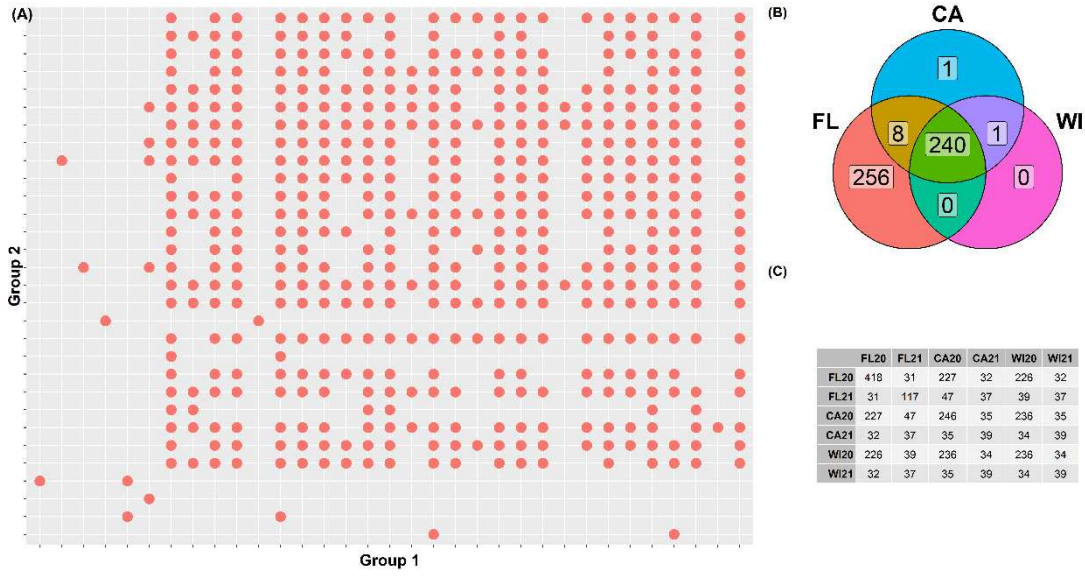
GMAD: GBLUP multi-trait with additive effect and dominance effect. GSA: GBLUP single-trait with additive effect. GSAD: GBLUP single-trait with additive effect and dominance effect. RMA: RKHS multi-trait with additive effect. RMAD: RKHS multi-trait with additive effect and dominance effect. RSA: RKHS single-trait with additive effect. RSAD: RKHS single-trait with additive effect and dominance effect. Sites:

Table S8. Prediction accuracy of across-site hybrids prediction for the sites of California, Florida, and Wisconsin. EL: ear length. EW: ear width. TPF: tip fill. Here, only the genotypes that were not assessed at the testing site were included in the training set from 2020 sites (CA20, FL20, and WI20). The cross-validation scheme was the CV00 (untested hybrids in untested environments). FLWICA: training set combines CA20, FL20, and WI20 to predict CA21 site. FLCA: training set combines CA20 and FL20 to predict CA21 site. WICA: training set combines CA20 and WI20 to predict CA21 site. CAWIFL: training set combines CA20, FL20, and WI20 to predict FL21 site. CAFL: training set combines CA20 and FLO20 and to predict FL21 site. WIFL: training set combines WI20 and CA20 to predict FL21 site. CAFLWI: training set combines CA20, FL20, and WI20 to predict WI21 site. CAWI: training set combines CA20 and WI20 and to predict WI21 site. FLWI: training set combines FL20 and WI20 to predict WI21 site.

Env	GSA	GSAD	GMA	GMAD	RSA	RSAD	RMA	RMAD
				EL				
FLWICA	0.109	**0.156**	0.016	0.123	0.148	0.156	0.114	0.139
FLCA	0.107	0.161	-0.016	0.097	**0.162**	0.148	0.095	0.123
WICA	0.039	-0.065	0.007	-0.001	0.029	0.048	**0.068**	0.039
CAWIFL	-0.016	-0.062	-0.073	-0.058	-0.010	-0.007	-0.035	**-0.001**
CAFL	**0.040**	-0.019	0.025	0.015	0.013	0.008	0.013	0.001
WIFL	0.092	0.057	**0.167**	0.112	0.056	0.078	0.114	0.142
CAFLWI	-0.022	0.008	-0.089	-0.067	**0.024**	0.015	-0.033	-0.023
CAWI	0.115	-0.058	0.082	0.005	**0.870**	0.866	0.863	0.864
FLWI	-0.050	-0.042	-0.010	**0.022**	-0.043	-0.030	-0.012	-0.011
				EW				
FLWICA	**0.501**	0.450	0.417	0.409	0.470	0.476	0.453	0.452
FLCA	**0.414**	0.413	0.392	0.388	0.397	0.399	0.385	0.388
WICA	0.190	0.182	0.115	0.173	0.196	**0.231**	0.197	0.192
CAWIFL	0.043	-0.065	0.034	0.013	0.037	0.038	**0.048**	0.018
CAFL	**0.046**	0.020	0.034	0.017	0.045	0.020	0.019	0.012
WIFL	**0.114**	0.047	0.096	0.035	0.042	0.031	0.040	0.035
CAFLWI	0.058	0.154	0.017	0.131	**0.191**	0.169	0.147	0.154
CAWI	-0.140	-0.073	-0.202	-0.167	0.373	0.377	0.395	**0.405**
FLWI	0.260	0.270	0.231	0.285	0.289	0.267	0.284	**0.297**

	TPF							
FLWICA	0.190	0.274	0.208	0.275	0.256	0.255	0.281	**0.292**
FLCA	0.132	0.211	0.184	**0.322**	0.193	0.197	0.198	0.235
WICA	-0.088	**-0.080**	-0.109	-0.091	-0.095	-0.114	-0.103	-0.118
CAWIFL	0.151	-0.058	0.143	**0.152**	0.145	0.143	**0.152**	0.135
CAFL	0.121	0.131	**0.135**	0.126	0.124	0.120	0.113	0.119
WIFL	0.109	0.090	0.133	**0.146**	0.072	0.090	0.118	0.114
CAFLWI	0.232	0.225	0.241	0.224	0.273	0.267	0.275	**0.280**
CAWI	0.018	0.129	0.050	0.095	0.488	0.509	0.564	**0.565**
FLWI	0.105	0.170	0.102	**0.193**	0.136	0.117	0.109	0.130

GMAD: GBLUP multi-trait with additive effect and dominance effect. GSA: GBLUP single-trait with additive effect. GSAD: GBLUP single-trait with additive effect and dominance effect. RMA: RKHS multi-trait with additive effect. RMAD: RKHS multi-trait with additive effect and dominance effect. RSA: RKHS single-trait with additive effect. RSAD: RKHS single-trait with additive effect and dominance effect.



	FL20	FL21	CA20	CA21	WI20	WI21
FL20	418	31	227	32	226	32
FL21	31	117	47	37	39	37
CA20	227	47	246	35	236	35
CA21	32	37	35	39	34	39
WI20	226	39	236	34	236	34
WI21	32	37	35	39	34	39

Figure S1. Summary of the hybrids and lines used in the study. (A) Schematic representation of the tested hybrid (506) parents by crossing the parentals (Group 1 and Group 2), where each red dot represents one cross between two parents. (B) A Venn diagram representing the number of hybrids planted that was shared among the three environments. FLO: Florida site, CAL: California site, WIS: Wisconsin site. (C) Number of hybrids assessed at each environment (diagonal) and number of hybrids shared among the six environments (off-diagonal). FL20: Florida site, 2020. FL21: Florida site, 2021. CA20: California site, 2020. CA21: California site, 2021. WI20: Wisconsin site, 2020. WI21: Wisconsin site, 2021.

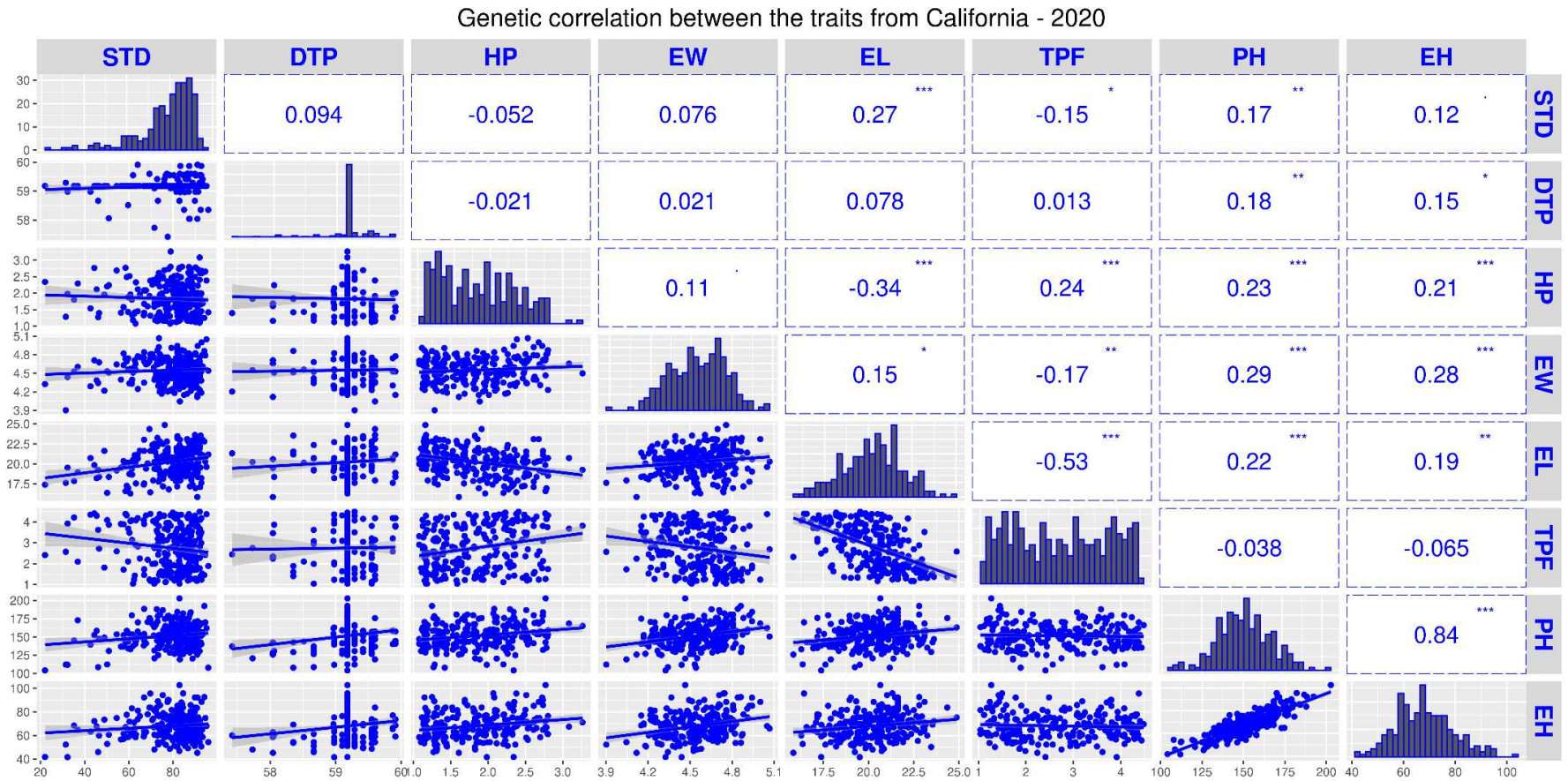


Figure S2. Pearson correlations between vector of BLUEs for California 2020 environment. DTP: days to pollination. EH: ear heigth. EL: ear length. EW: ear width. HP: husk protection. PH: plant heigth. STD: stand count. TPF: tip fill. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

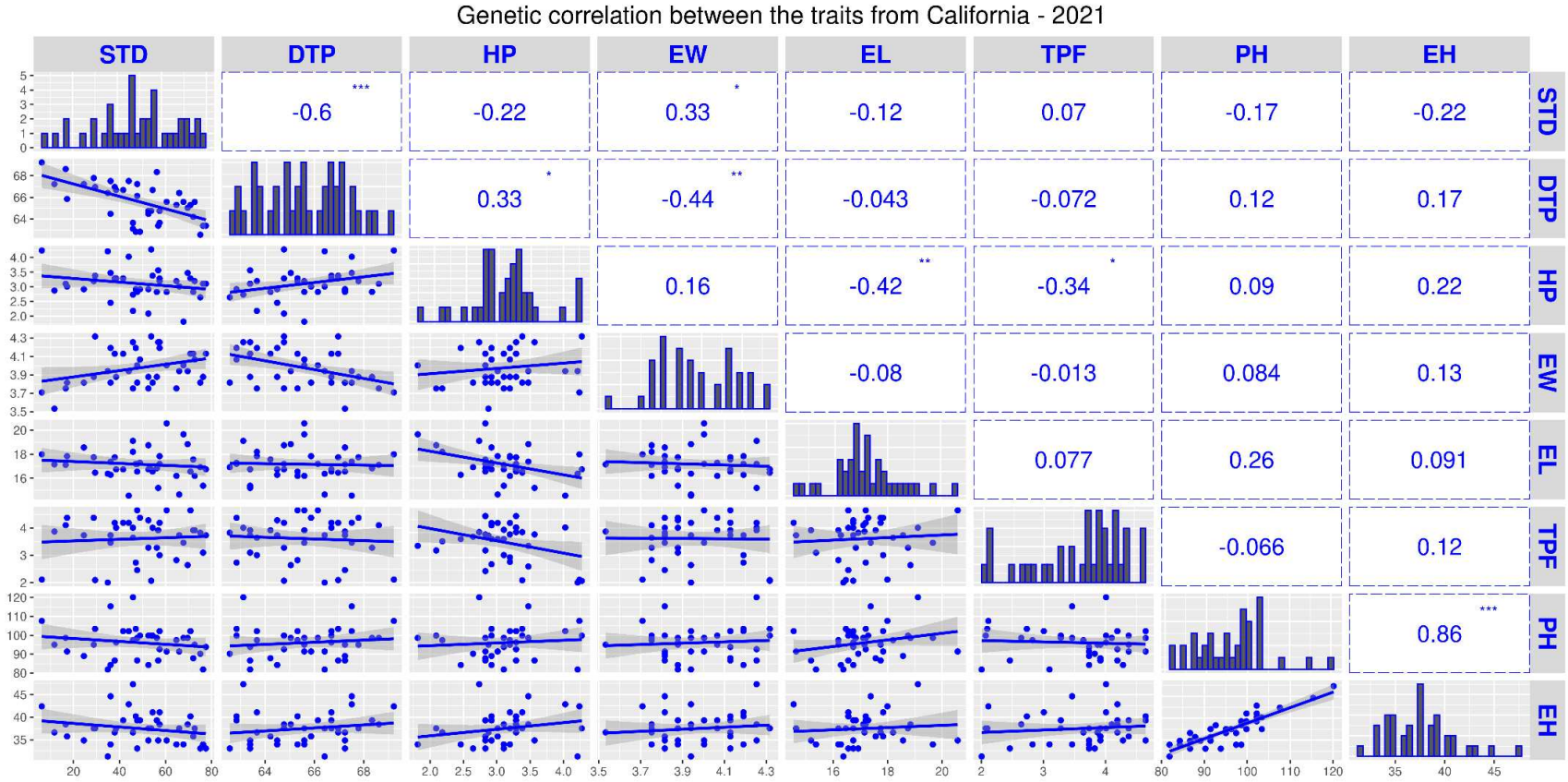


Figure S3. Pearson correlations between vector of BLUEs for California 2021 environment. DTP: days to pollination. EH: ear heigth. EL: ear length. EW: ear width. HP: husk protection. PH: plant heigth. STD: stand count. TPF: tip fill. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

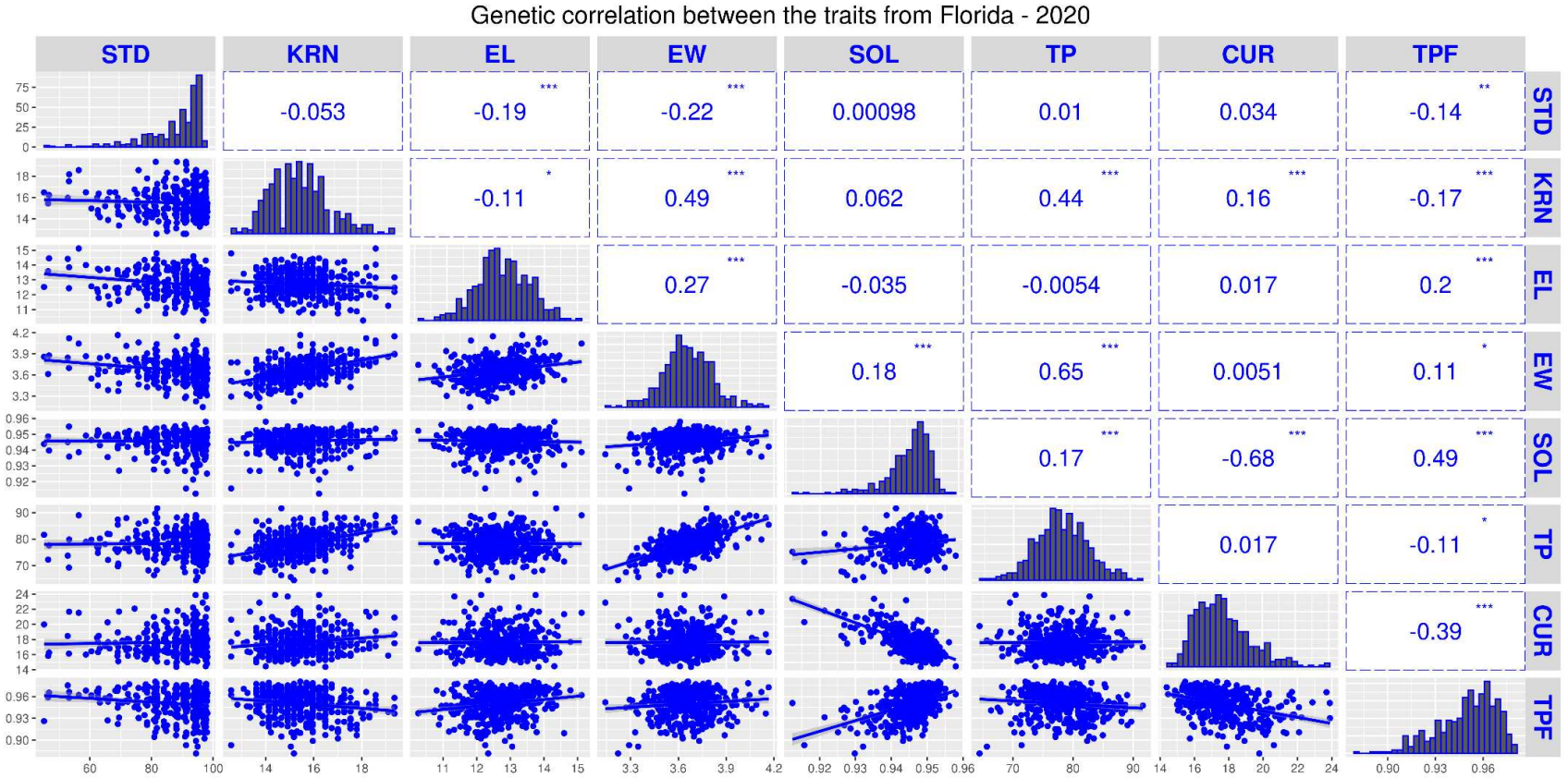


Figure S4. Pearson correlations between vector of BLUEs for Florida 2020 environment. CUR: curvature. EL: ear length. EW: ear width. KRN: kernel row number. SOL: solidity, STD: stand count. TP: taper. TPF: tip fill. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

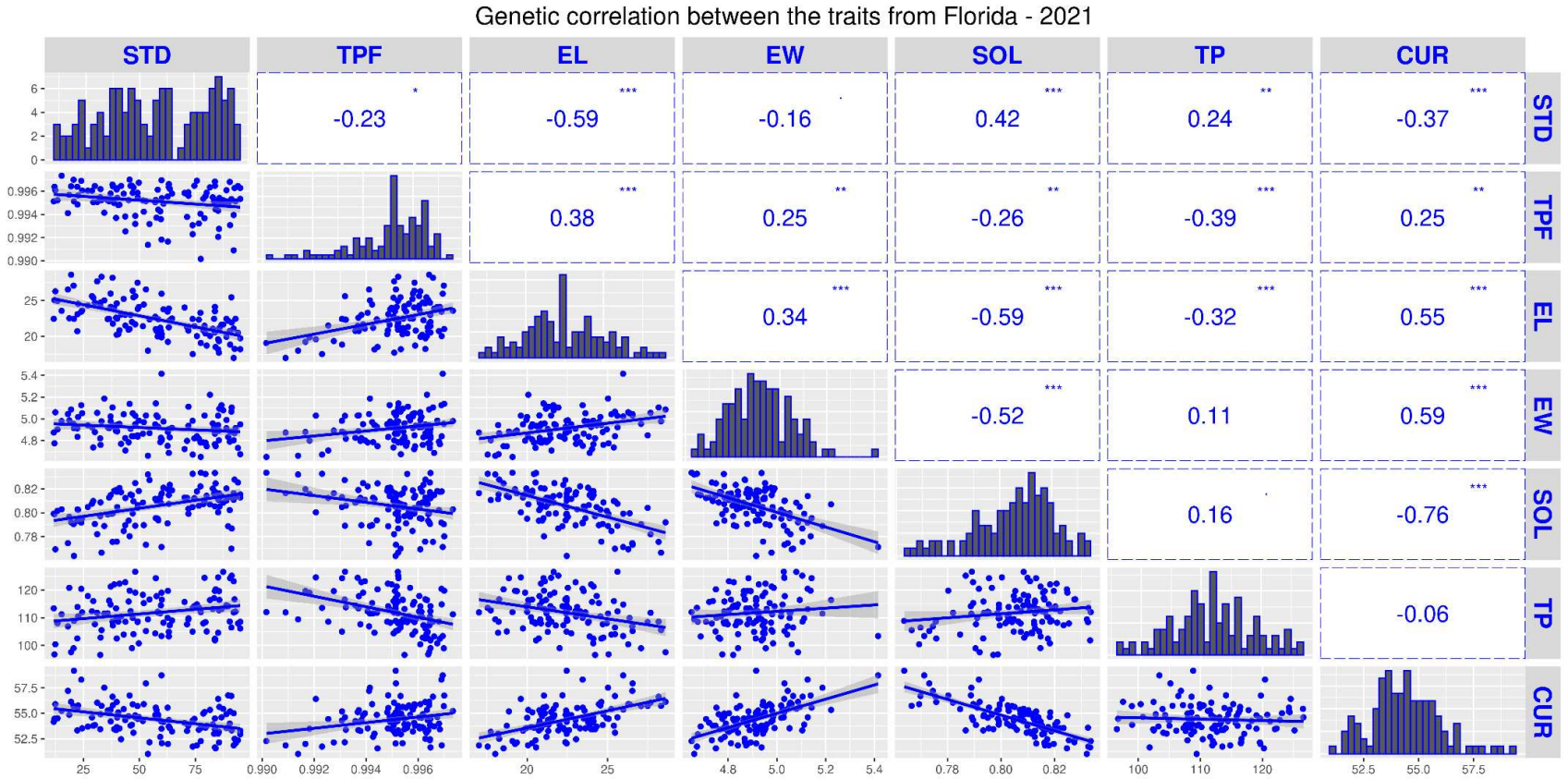


Figure S5. Pearson correlations between vector of BLUEs for Florida 2021 environment. CUR: curvature. EL: ear length. EW: ear width. KRN: kernel row number. SOL: solidity, STD: stand count. TP: taper. TPF: tip fill. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

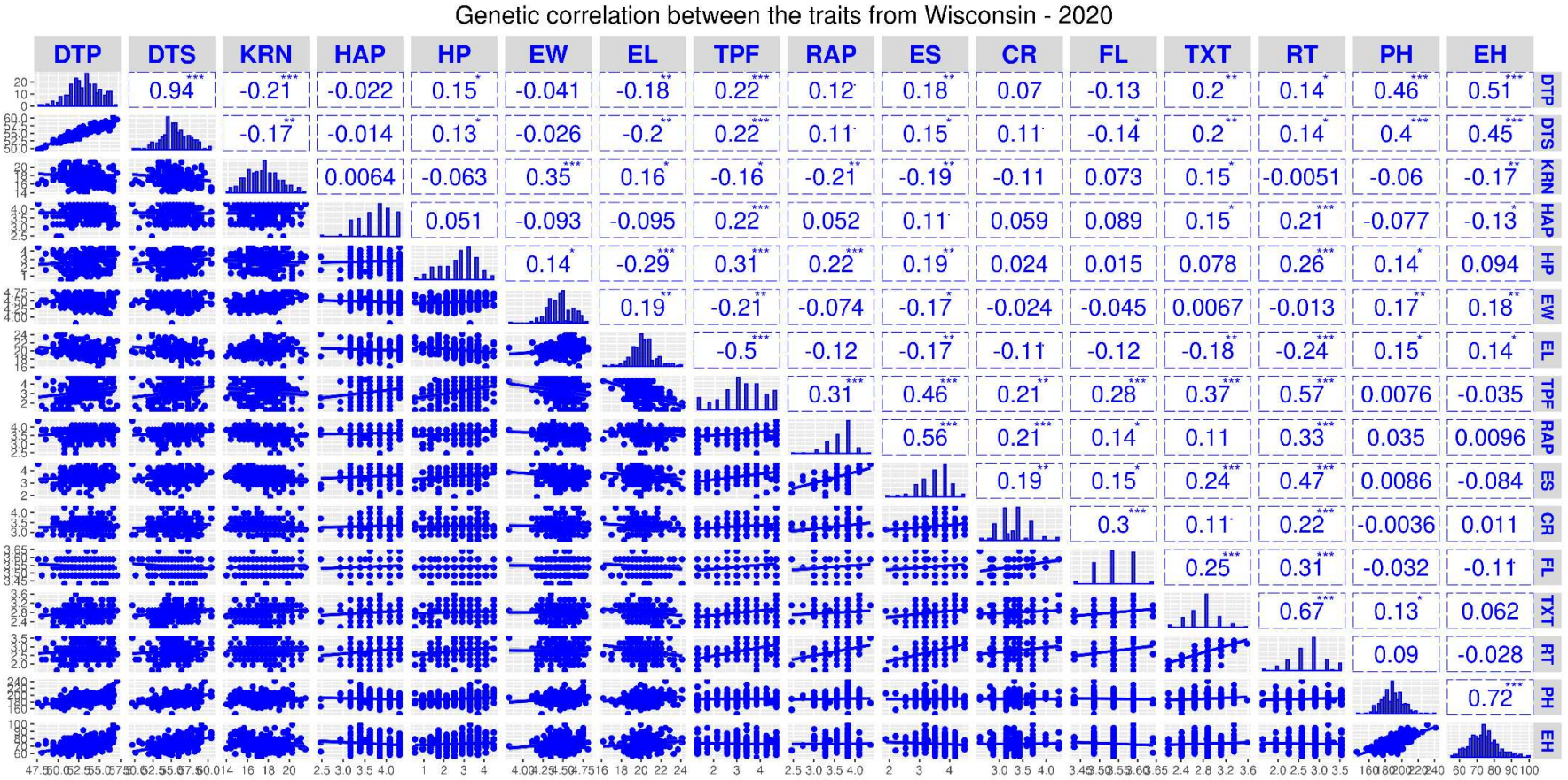


Figure S6. Pearson correlations between vector of BLUEs for Wisconsin 2020 environment. CR: color rate. DTP: days to pollination. DTS: days to silking. EH: ear height. EL: ear length. ES: ear shape. EW: ear width. FL: flavor. HAP: husk appearance. HP: husk protection. KRN: kernel row number. PH: plant height. RAP: Row appearance. RT: rating. TPF: tip fill. TXT: texture. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

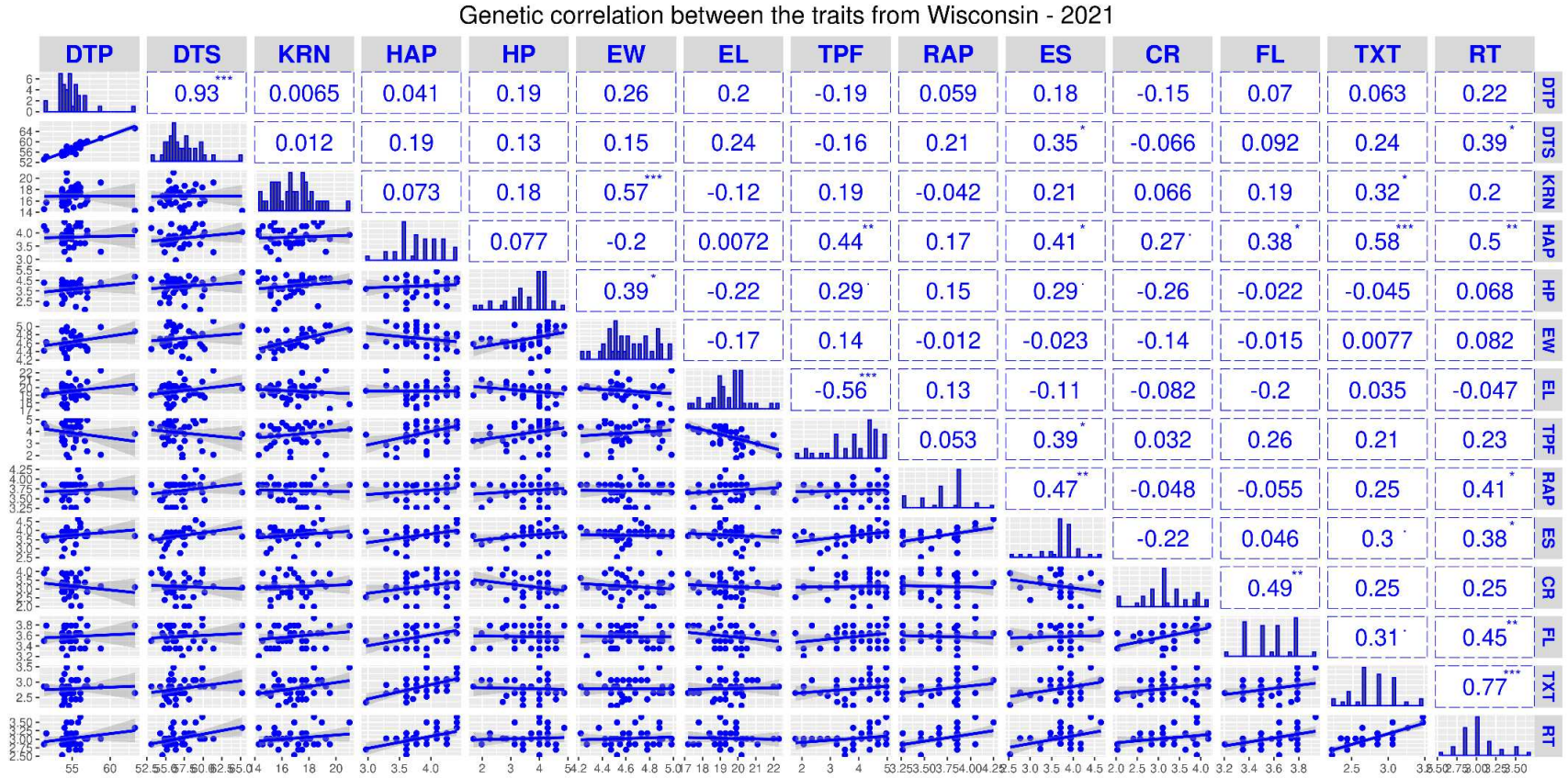


Figure S7. Pearson correlations between vector of BLUEs for Wisconsin 2021 environment. CR: color rate. DTP: days to pollination. DTS: days to silking. EL: ear length. ES: ear shape. EW: ear width. FL: flavor. HAP: husk appearance. HP: husk protection. KRN: kernel row number. RAP: Row appearance. RT: rating. TPF: tip fill. TXT: texture. *,**,*** represents the significance of the correlations at 0.10, 0.05, and 0.01 levels.

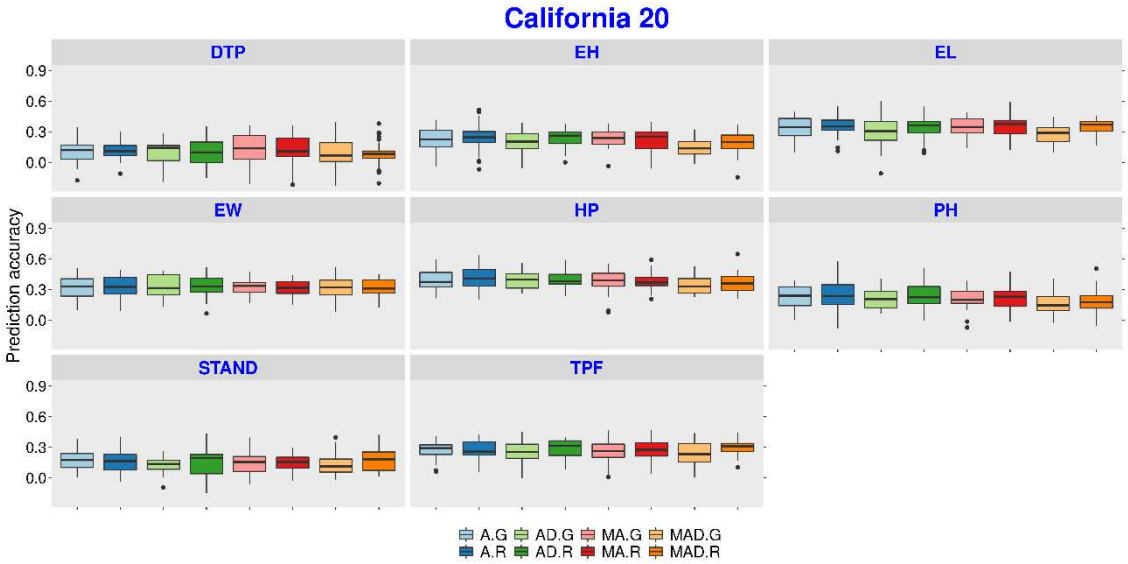


Figure S8. Prediction accuracy for eigth traits using eigth four different models under GBLUP and RKHS predictions for California 2020 site. DTP: days to pollination. EH: ear heigth. EL: ear length. EW: ear width. HP: husk protection. PH: plant heigth. STD: stand count. TPF: tip fill.

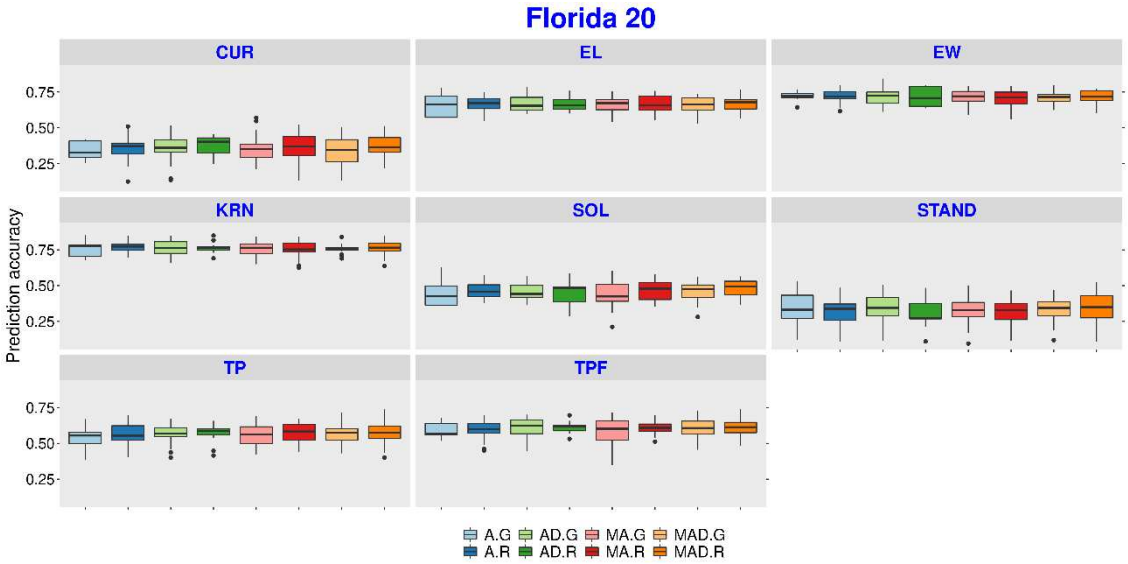


Figure S9. Prediction accuracy for eigth traits using eigth four different models under GBLUP and RKHS predictions for Florida 2020 site. CUR: curvature. EL: ear length. EW: ear width. KRN: kernel row number. SOL: solidity, STD: stand count. TP: taper. TPF: tip fill.

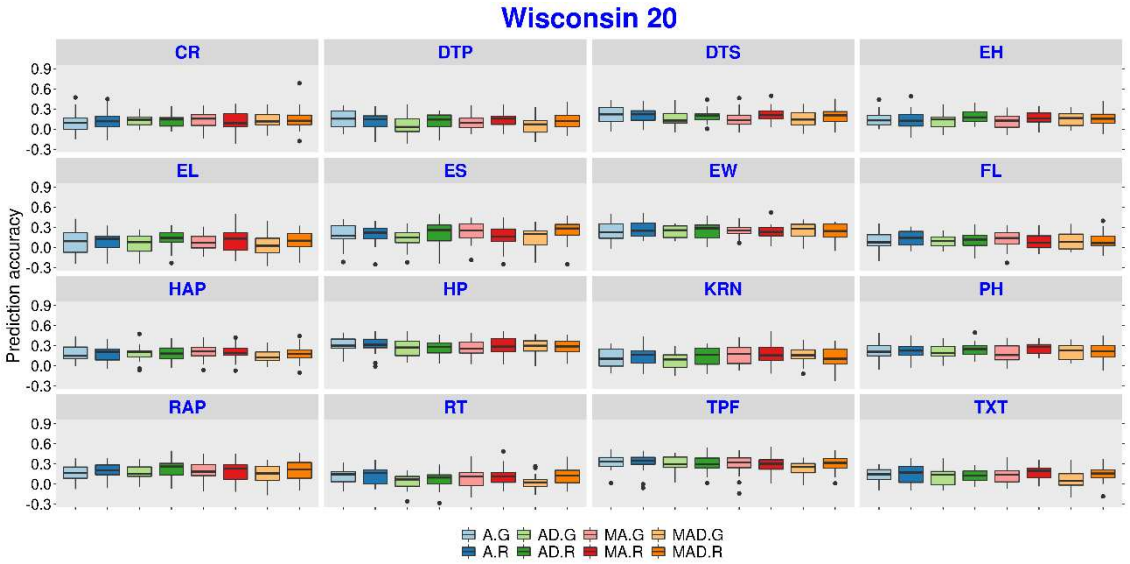


Figure S10. Prediction accuracy for eigth traits using eigth four different models under GBLUP and RKHS predictions for Wisconsin 2020 site. CR: color rate. DTP: days to pollination. DTS: days to silking. EH: ear height. EL: ear length. ES: ear shape. EW: ear width. FL: flavor. HAP: husk appearance. HP: husk protection. KRN: kernel row number. PH: plant height. RAP: Row appearance. RT: rating. TPF: tip fill. TXT: texture.

Chapter 3

SMate: R-package for cross prediction and optimization using genomic selection

SMate: R-package for cross prediction and optimization using genomic selection

Marco Antônio Peixoto[1,2], Rodrigo Rampazo Amadeu[3], Leonardo Lopes Bhering[1], Patrício R. Munoz[4], Felipe Ferrão[4], Márcio F. R. Resende[2,*]

[1] Laboratório de Biometria, Universidade Federal de Viçosa, Viçosa, Minas Gerais State, Brazil.

[2] Sweet Corn Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

[3] Bayer Crop Science, Chesterfield, Missouri State, United States

[3] Blueberry Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

*** Corresponding author**

Abstract

Plant breeding programs rely on balancing long-term genetic diversity and genetic gains, which are conflicting goals. A method to deal with this corollary is through cross selection, where we can jointly optimize the selection of crosses and the maintenance of genetic diversity. Given the problem detailed, here we described **SMate**, a flexible R package to cross prediction and optimization. The package represents a tool for breeding programs to balance genetic gains and inbreeding rate levels. A valid mate plan is built based on two core aspects: (i) prediction of usefulness for potential cross and (ii) optimization of the set of crosses. Usefulness accounted for the mean and variance of each cross recovered from marker data, markers effects, and linkage disequilibrium matrix. The mean and variances can be predicted /estimated for traits with additive control or additive and dominance control. In addition, multi-trait scenarios are allowed by building a cross-selection index based on weights for each trait (for mean and variance of a cross). The optimization algorithm maximizes the usefulness and minimizes next generation inbreeding, which, at the end, circumvent the reduction of genetic diversity over breeding cycles by crossing less related individuals. The user may provide the ideal number of crosses, the maximum/minimum number of crosses per parent, and a maximum value of relationship between parents that can be used to create a cross. Then, it builds a mating plan from the target parental population. An example of implementing **SMate** in a maize breeding program with two heterotic groups is given. In conclusion, **SMate** package enables to optimize cross selection in breeding programs targeting long term genetic gains.

Keywords: breeding program optimization, inbreeding, genetic gain, usefulness, non-additive effects.

1. Introduction

Genomic selection (GS) is a well-stablished tool in plant breeding programs, which leverages short-term genetic gain (Marinho *et al.* 2022; Sandhu *et al.* 2022; Vieira *et al.* 2022). The foundation of GS resembles in the association of high-density markers to quantitative trait loci (QTL) that controls the phenotypic expression of a quantitative trait under selection. In short-term, plant breeding programs can improve performance by selecting individuals with the most outstanding genetic values for composing the next generation. Unfortunately, GS also accelerates the decrease of genetic diversity (Bančič *et al.* 2021; Sabadin *et al.* 2022), once it works together with the selection process quickly fixing the loci with large effects changing allele frequency. In addition, genetic variation may decrease over cycles by random drift, as well as by the accumulation of negative linkage disequilibrium (LD), which is known as Bulmer effect (Bulmer 1971; Walsh and Lynch 2018).

According to conventional wisdom, genetic variation is necessary for long-term genetic gains, and while truncated selection maximizes gain in the next cycle, it may not necessarily maximize long-term genetic gain (Sonesson *et al.* 2012). In GS, related individuals are ranked higher through selection, which leads to more related individuals being mated in the next breeding cycle, resulting in an increase in a measure of relatedness called inbreeding (Falconer and Mackay 1996). However, with an increase in population inbreeding comes a decrease in genetic variation and potentially decreases long-term genetic gain.

Based on the problem stated, several authors have put together efforts to maintain population' genetic variation in higher levels while constrains inbreeding rates in the next generation, especially after the inclusion of GS into breeding pipelines (Jannink 2010; Daetwyler *et al.* 2015; Akdemir and Sánchez 2016; Gorjanc and Hickey 2018; Allier *et al.* 2019b). The first proposal came from Meuwissen and Sonesson (1998), were genetic value is

maximized while inbreeding is constrained to give the optimal contributions of parents to the next generation. The optimum contribution selection has been seeming like a solid base for all following contributions, especially after the idealization and adoption of GS by breeding programs. For example, Jannink (2010) proposed to weight allelic effects by their frequency and magnitude, aiming to preserve rare favorable alleles with large effects on the genetic value. Another idea was the optimal selection value, where outbred individuals are selected for their predicted value to the best doubled haploids lines (Daetwyler *et al.* 2015). In addition, the optimal cross selection puts together the optimization of selection intensity, inbreeding rate, coancestry rate, and cross allocation (Gorjanc *et al.* 2018).

A concept that is becoming more popular toward the same goal before stated is the usefulness (UC). It unifies genetic mean and genetic gain of a cross given a quantitative trait (Schnell and Utz 1976): $\boldsymbol{UC = \mu + ih\sigma}$, where: $\boldsymbol{\mu}$ represents the cross mean, $\boldsymbol{i}$ is the intensity of selection, $\boldsymbol{h}$ is the heritability square root, and $\boldsymbol{\sigma}$ represents the standard error from the genetic variance of a cross. That equation relates two important components for a plant breeding program: the genetic mean ($\boldsymbol{\mu}$) expected for a cross and its expected gain ($\boldsymbol{ih\sigma}$). One of the most important concepts from UC is the genetic variance (or its standard deviation). Over the last years several efforts have being put forward in cross variance estimation. In one hand, Quaas, (1988) proposed and Legarra et al., (2009) expanded to molecular markers dataset a useful way to calculate crossbreed variance, and, ultimately, cross variance (Akdemir and Sánchez 2016). However, in contrast to this idea, Lehermeier et al., (2017) proved that the use of LD and its contribution to cross variance represents the best alternative, being largely applied since then (Allier *et al.* 2019b, 2019a; Wolfe *et al.* 2021).

After the advance of molecular markers, diversity in a population is generally measured by the level of heterozygosity (as the proportion of heterozygotes). Analogously, the inbreeding

coefficient can be derived from the same idea (F_t): 1-f_h, where f_h is the frequency of heterozygotes in the population (Falconer and Mackay 1996; Toro *et al.* 2014). In addition, the covariances from a marker-based relationship matrix divided by two represents the value of the coancestry between a pair of genotypes (Toro *et al.* 2014; Meuwissen *et al.* 2020), and it represents the inbreeding in the next generation.

The optimum cross selection here addressed delivers a crossing plan that maximizes a criterion (cross mean, cross usefulness, or even and trait-based index) under constrains for expected inbreeding, number of parents, and minimum/maximum number of crosses per parent. Specifically, the goal is: to maximize $\boldsymbol{g_t}$, where $\boldsymbol{g}$ is a vector of contributions of selection candidates to the next generation (***t+1***), while constraining the selected group coancestry $\boldsymbol{g'_t G_t g_t}$, given a minimum or lower boundary ($\boldsymbol{l}$) and maximum or higher boundary ($\boldsymbol{u}$) number of crosses. The optimization problem, is given by:

$$\max \boldsymbol{g_{t+1}} = \boldsymbol{g_t}$$

$$\text{subject to } \bar{\boldsymbol{G}}_{\boldsymbol{t+1}} = \frac{\boldsymbol{g'_t G_t g_t}}{\boldsymbol{2}},$$

$$\boldsymbol{l} \leq \boldsymbol{g_t} \leq \boldsymbol{u}$$

where $\bar{G}$ represents the average relationship matrix all candidates in next generation (***t+1***). The optimization mentioned above works to improve long-term genetic gains by reducing inbreeding. There are two key aspects of the optimization that achieve this goal. Firstly, by keeping parental coancestry at low levels, long-term inbreeding is also reduced. Secondly, the avoidance of mating relatives helps to maintain short-term inbreeding at lower levels (Kinghorn 2011). By reducing both short and long-term inbreeding, genetic variation is preserved, which is essential for achieving long-term genetic gains.

122 To date, just a few algorithms, packages, or programs has been proposed for cross
123 prediction and optimization targeting breeding programs (Goda and Isik; Kinghorn 2011;
124 Akdemir and Sánchez 2016; Gorjanc and Hickey 2018). However, the implementation of
125 usefulness, for quantitative traits with different genetic control (additive and additive plus
126 dominance) and the easiness to implement it in breeding programs with several traits at a time
127 together with the optimization, has never being a subject of a package.

128 The aim of this study was to develop a free package on R environment that can integrates
129 cross prediction and optimization aiming long-term genetic gains. **SMate** does: (i) prediction
130 of cross usefulness (genetic gain and genetic mean), (ii) optimal cross selection with an
131 optimization algorithm to increase crosses usefulness and decrease inbreeding rates, and (iii)
132 incorporates the potentialities to several traits at a time and complex traits with additive and
133 dominance effects. The package is freely available on GitHub (available after acceptance). In
134 the following, we describe the theory and technical implementation of the package, demonstrate
135 the how to use it, show an implementation in a maize breeding program, and discuss the
136 potentialities and futures directions.

2. Use

Here, we describe the package functionality while connecting the functions with its genetic/statistical equations. We describe: (i) the initial data input needed; (i) a function to thin out individuals by relatedness; (ii) a function to create a cross plan to be optimized; (iii) how to calculates the usefulness; (iv) to organize the data input for optimization; (v) the implementation of the optimization. Furthermore, the basic idea is to predict a criterion for each potential cross together with the rate of future inbreeding of the cross. Then, leverages criteria's level while decrease the rates of future inbreeding by means of the optimization algorithm.

Data input

The initial step is the data preparation. Four imputes are needed, but not all required, by the **SMate**.

1. Allele dosage matrix: matrix that contains the single nucleotide polymorphism (SNP) markers coded as 0,1,2 for aa, Aa, AA respectively. It should include all potential candidates to be parents. Here, the matrix is used to predict the contribution of each duo of parents to the future cross.

2. Relationship matrix: a matrix that could be a genomic-based relationship matrix (**G**), built up from an allele dosage matrix, or a pedigree-based relationship matrix (**A**). The package uses that information of the covariances between individuals as a measure of coancestry for constraining inbreeding in the next generation.

3. Linkage disequilibrium matrix (LD): a matrix calculated for the potential candidates to be parents. Linkage disequilibrium represents a population property, being easily calculated from a unphased SNP dataset. It describes the association of alleles at different loci and could

describe both evolutionary and biological aspects of the population (Ragsdale and Gravel 2019).

4. Markers effects: vector (for one trait) or data frame (for more than one trait) containing the effect of each marker presented in the allele dosage matrix. For a trait controlled by additive and dominance effects, we can supply both, additive and dominance markers effects. For several traits, the user may supply markers effects for all traits from the same set of SNPs.

5. Estimated breeding value (*ebv*) or total genetic value (*tgv*): vector (for one trait) or data frame (for more than one trait) containing the values for all individuals and traits. It is optional (see description below).

Install and load it

After organizing the input dataset, the next step is to install the package **SMate** and load it. The user may load it from GitHub, using the 'devtools' package, as follows:

```
> devtools::install_github("...")
> require("...")
```

Thinning by relatedness

Subsequently, to thin out some of the individuals that are candidates to parents, the function `relateThinning()` can be used, even though it is an optional step. Hence, when we decrease the number of selected candidates by cutting individuals by relatedness (*i.e.*, those closely related) we decrease the time for prediction of usefulness and for cross optimization, increasing the efficiency of the process. A compelling example explaining better the function is given in **Supplementary material S1**. First, it selects the top individuals within a family by the relatedness threshold. For example, if *max.per.cluster* = 2 and *threshold* = 0.5 the algorithm will identify all groups of individuals that shared more than 0.5 in the relationship matrix and

select the top 2 individuals with highest criterion to be used as possible parents in the mating plan. Besides those arguments, we need to give to the function the relationship matrix (*K* argument) between the parents ($n \times n$, being n the number of candidates to parents). It could be the relationship matrix based on markers information (*G*) or based on pedigree information (*A*). In addition, for the *criterion* argument we should add an information of the individual's performance for the target trait ($n \times 1$). Here, the user can set the estimated breeding value of the individual, total breeding value, or the BLUP of the candidates.

```
> keep = relateThinning(K = G,
+                       criterion = gen_value,
+                       threshold=0.8,
+                       max.per.cluster=4)
96 genotypes kept out of 100
```

For more than one trait, the option is to set weights for each trait and construct an index at trait-level. **SMate** offers the function `calc.index()` for such end. Four arguments must be set: *Gen_value*: a data frame (*n* x *t*) with all genetic values of each individual (ebv, tgv, or BLUP); *Gen_ID*: a vector with all individuals identification (*n* x 1); *Weights*: a vector with the weights for each trait; and a flag to scale or not the data frame with the genetic values (*Scale*). The output can be used into the function `relateThinning()` for thinning the dataset.

```
> # Trait-level index
> outp = calc.index(Gen_value = criteria,
+                   Gen_ID = indiv_ID,
+                   Weights = c(0.5,0.6),
+                   Scale = TRUE)
```

Planning the crosses

From the list of individuals that we should keep for the cross prediction (*keep*), we can use the function `cross_plan()` to generate a data frame with all crosses to be optimized. In the function, the argument *K* represents the relationship matrix of potential parents (*G* or *A* matrix).

The argument *Indiv* asks for a list of all individuals kept after the thin out process (or you can give all the individuals, once the thin out process is not mandatory). The argument *Type* represents the type of crosses that should be generated, and '*half*' indicates that no reciprocals nor parents are included in the crosses.

```
> plan = cross_plan(K = G,
+                   Indiv = keep,
+                   Type = 'half')
Number of crosses generated: 4950

> head(plan,10)
   Par_1 Par_2 idcross
1     G1    G2   G1_G2
2     G1    G3   G1_G3
3     G1    G4   G1_G4
4     G1    G5   G1_G5
5     G1    G6   G1_G6
6     G1    G7   G1_G7
7     G1    G8   G1_G8
8     G1    G9   G1_G9
9     G1   G10  G1_G10
10    G1   G11  G1_G11
```

Usefulness criterion

After tracking the potential crosses for the population, the next step is to run the prediction of usefulness for them. A thoroughly description of the calculation and implementation of the formulas for the usefulness estimates are given in **Supplementary material S2**. Four functions inside *SMate* can do so, where each one can be implemented in one different scenario: for one trait mainly controlled by additive effects (*usefA*), one trait controlled by additive and dominance effects (*usefAD*), more than one trait controlled by additive effect (*usefA*_mt), and more than one trait controlled by additive and dominance effects (*usefAD*_mt). Here, we describe the implementation of one trait controlled by only additive effects (we used *usefA*). The inputs are the allele dosage matrix (*Markers*), coded as 2,1,0 for AA, Aa and aa (3000 x 3000), the additive effects for each one of the markers (*Markers.effA*,

3000 in the example), the matrix of linkage disequilibrium (*LDMat*) for all candidate's parents (100 parents), the intensity of selection (*Intsel*), and cross's plan (*PlanV*).

```
> cross_usef = usefA(Markers = Markers,
+                    Markers.effA = addEff,
+                    LDMat = Mat.ld,
+                    Intsel = 0.05,
+                    PlanV = plan)

> head(cross_usef, 10)
     idcross Par_1 Par_2       Var     MuE     UCt
4410 G73_G93   G73   G93 12.878668 1.04165 8.44408
4203 G69_G93   G69   G93 12.765963 1.02911 8.39908
4411 G73_G94   G73   G94 12.195343 1.08674 8.29011
4204 G69_G94   G69   G94 12.097538 1.07420 8.24863
4355 G71_G93   G71   G93 12.328579 0.96791 8.21052
4356 G71_G94   G71   G94 11.661523 1.01300 8.05695
4415 G73_G98   G73   G98  9.966154 1.25323 7.76505
4208 G69_G98   G69   G98  9.874297 1.24068 7.72242
4360 G71_G98   G71   G98  9.500105 1.17949 7.53723
4414 G73_G97   G73   G97  9.057009 1.32835 7.53606
```

The function output is a data frame with the cross id (idcross), parents (Par_1 and Par_2), variance (Var), mean (MuE), and usefulness (UCt) for each cross.

Organizing the input for the optimization

The optimization algorithm works into a four-column data frame, as given: P1 and P2 represents the parents of the target cross, Y represents the criteria used for maximization and K is a column with the relationship values that came from the *G* or from an *A* matrix. While we provide a way to calculates usefulness from a SNP matrix, the user can provide their own criteria, such as BLUPs for the individuals together with the values of K that can come from a pedigree matrix. For breaking down the *G* (or *A*) we used the function `meltK()`.

```
> # Melting the matrix G for pair-pair covariance (using the auxiliary function me
ltK)
> Gmat_melted = meltK(G)
> # Transform the info into a data frame
```

```
> par_info = data.frame(idcross = paste0(Gmat_melted$P1, "_", Gmat_melted$P2),
+                  K = Gmat_melted$K)
> # Combining the information
> df = merge(cross_usef, par_info, by = "idcross")
> # Build up the input for the optimization
> cand.crosses = data.frame(P1 = df$Par_1,
+                           P2 = df$Par_2,
+                           Y = df$UC,
+                           K = df$K)
> head(cand.crosses,10)


   P1   P2       Y           K
1  G1  G10 1.30135 -0.05286433
2  G1 G100 1.72526 -0.11340694
3  G1  G11 2.76706 -0.10777084
4  G1  G12 3.69125 -0.12404280
5  G1  G13 2.54194 -0.16381366
6  G1  G14 2.76373 -0.10754358
7  G1  G15 1.99167 -0.03136534
8  G1  G16 1.97006  0.04294934
9  G1  G17 2.67456 -0.01236624
10 G1  G18 4.83667 -0.12695175
```

Running the optimization

After building the input file, we may use the function for optimization (**`selectCrosses()`**). The user must give the set of candidate crosses together with the criteria (*data*), and the *G/A* matrix (*K*), inform the number of crosses that are needed (*n.cross*), the maximum (*max.cross*) and minimum contribution (*min.cross*) of each parent, the maximum number of crosses (*i.e.*, lines in the dataset) to search (*max.cross.to.search*) and the culling pairwise (*culling.pairwise.k*) based on the coancestry (Figure 4).

```
 # Optimization
> maxGainPlan = selectCrosses(data=cand.crosses,
+                             K=G,
+                             n.cross=50,
+                             max.cross = 5,
+                             min.cross = 1,
+                             max.cross.to.search=100000,
+                             culling.pairwise.k = 0.275)
```

```
> maxGainPlan[[1]]
  culling.pairwise.k target.Y    target.K
1              0.275 6.744127 -0.02018028



 # Mating plan
> head(maxGainPlan[[2]],10)
      P1  P2       Y          K
4410 G73 G93 8.44408 -0.1017711
4203 G69 G93 8.39908 -0.2582183
4411 G73 G94 8.29011 -0.1947213
4204 G69 G94 8.24863 -0.2420827
4355 G71 G93 8.21052 -0.2509005
4356 G71 G94 8.05695 -0.2620363
4415 G73 G98 7.76505 -0.1888579
4208 G69 G98 7.72242 -0.2362193
4360 G71 G98 7.53723 -0.1061800
4414 G73 G97 7.53606 -0.1974030
```

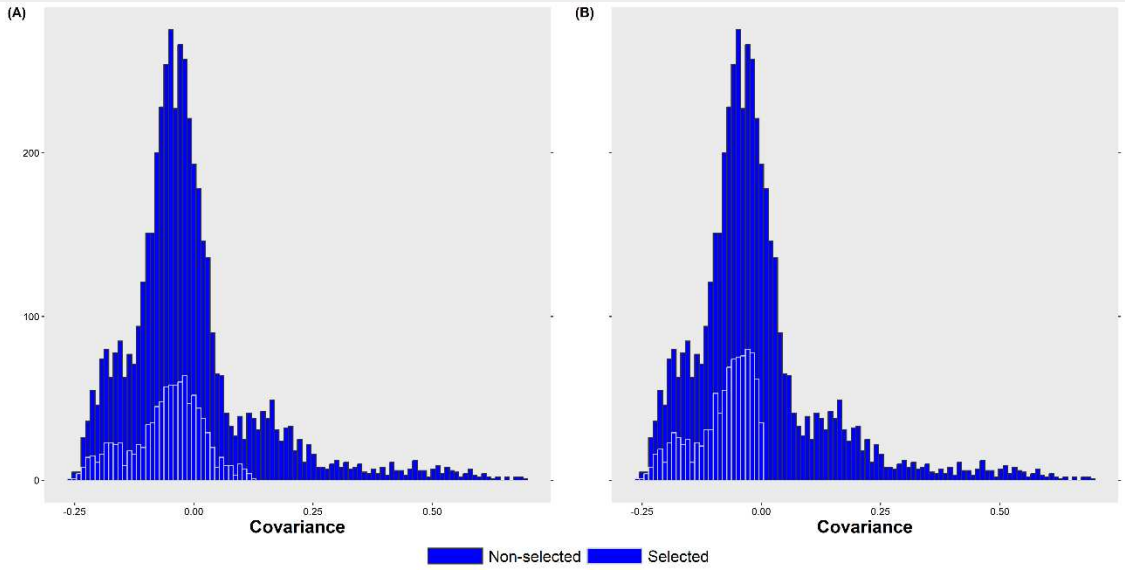


Figure 1 – Histogram of the covariance estimates (as function of coancestry) of non-selected and selected crosses after optimization. A) *culling.parwise.k* set to 0.125. B) *culling.pairwise.k* set to 0.

The results were based on a specific coancestry rate, number of crosses, maximum and minimum parental crosses, candidate list and in the G matrix (Figure 5). The output is a two twofold object, where we can assess the optimization parameters used (culling.pariwise.k, target.Y, and target.k) and in the second, the mating plan accordingly.

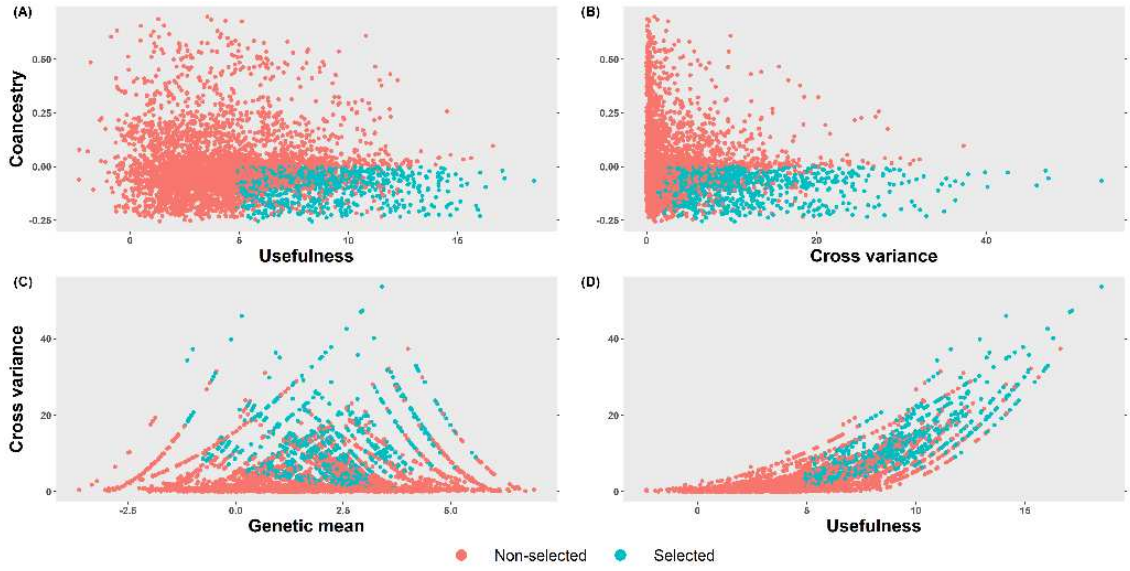
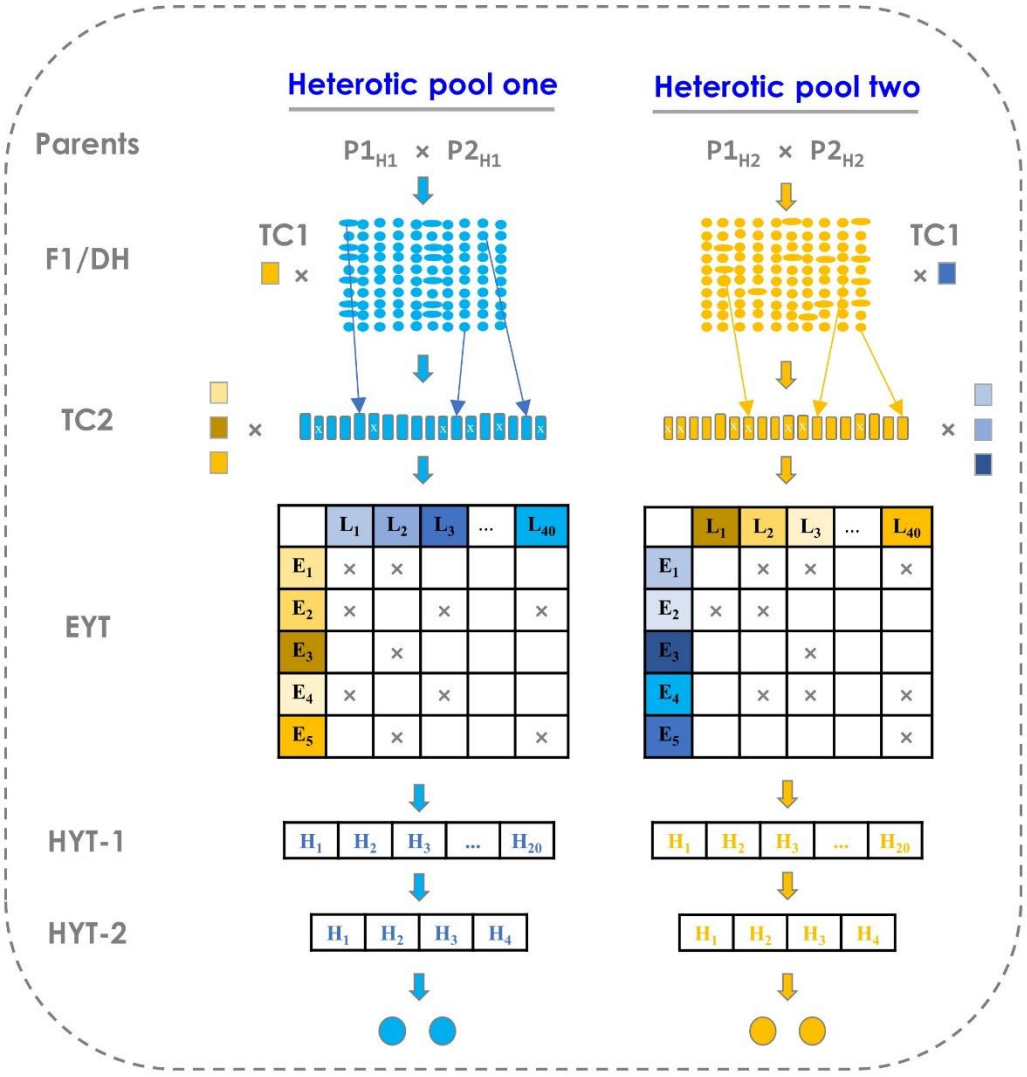


Figure 2 - Selected and non-selected crosses as optimization output. The coancestry level used was 0. A) Coancestry by Usefulness, B) Coancestry by Cross variance, C) Cross variance by Genetic mean, and D) Cross variance by Usefulness. A total of 4950 crosses were plotted. Red dots represent the non-selected crosses by the algorithm and green dots represent the selected crosses.

3.

Demonstration

We demonstrate the use of **SMate** by means of a set of stochastic simulations. We mimic a reciprocal recurrently selection maize breeding program, using doubled haploid technology to generate homogenous lines from F_1 populations (**Figure 1**). All tested scenarios were generated over a 25-year breeding program. A 15-year burn-in phase was implemented so each scenario would start at the same point. A burn-in phase represents the past years of the breeding program while the 25 years represent the upcoming years, where the strategies will be evaluated. Each scenario was replicated 50 times.



Base genome simulation

Figure 3 - Schematical representation of a simulated maize breeding program. Two heterotic pools where simulated. In the first stage, doubled haploid were generated from F_1 lines. Two set of test-crosses were implemented (TC1 and TC2). The hybrids where created crossing lines from one heterotic group (colors orange and Blue) with elite lines that came from the other heterotic group. At the end, the selection was made based on the hybrid performance in three different years and two best hybrids from each heterotic group were released.

We implemented a base genome containing 10 chromosome pairs via the Markovian Coalescent Simulator (Chen *et al.* 2009), available in AlphaSimR (Gaynor *et al.* 2020; R

Development Core Team 2022). As standard for the 'Maize' argument, each chromosome pair has a genetic length of 2.0 Morgans and a physical length of 2×108 base pairs. The recombination and mutation rate options were also included, being 1.25×10-8 and 2.5×10-8, respectively (Hickey et al. 2014; Powell et al. 2020). Then, we set evenly distributed 300 quantitative trait nucleotides (QTNs) per pair of chromosomes (3000 QTNs total). A split of the base genome into two different core genomes was simulated 100 generations ago. That split mimics the two heterotic groups structured in maize (*i.e.*, flint and dent groups), here used as two different groups for reciprocal recurrently selection program.

The biological model for this set of simulations was an additive, dominance, and genotype-by-environment interaction model. Each one of the 3000 QTN was assigned an additive genetic value, where the values were sampled from a standard normal distribution, its interaction with the environment, referred to as GE effect of each QTN, and a dominance effect due to the interaction between alleles in the same locus. In addition, for the implementation of genomic selection in the breeding program, we randomly allocated 100 single nucleotide polymorphisms (SNPs) randomly across each chromosome pair, representing a total of 1000 SNPs for each genotype.

Breeding program simulation – an example with maize breeding

We simulated one trait, with genetic mean of 70, genetic variance of 20, genotype-by-environment of 40, and residual variance of 270. The dominance effect was calculated according to the dominance degree of the trait, with mean of 0.93 and variance of 0.3. These values assumed for the mean and variance of the dominance degree were chosen to follow the historical levels that roughly represents the heterosis in field corn (Troyer and Wellin 2009; Powell *et al.* 2020). The average dominance degree represents the presence of a dominance deviation (δ) from the genetic mean. Due to an interaction between alleles, the dominance effect

shifts the mean, demonstrating the presence of heterosis and agreeing with historical data of commercial maize (Troyer and Wellin 2009).

Two scenarios were implemented for comparison. Before the implementation, a burn-in period was set to mimic the development the maize breeding in the past years, and it uses the reciprocal recurrently selection as the breeding strategy. This strategy starts by creating a new round of 80 F_1's from 80 elite parents in the program followed by successive cycles of reciprocal recurrently selection. All selections in the burn-in phase were based on observed phenotypic values. After the burn-in phase the two different scenarios were run. The first one was a reciprocal recurrently selection with doubled haploids production and genomic selection (hereafter called *DHGS* program) (Figure 3). The 80 crosses were random assigned from the group of 80 parents. The second scenario was outlined in the optimization of usefulness as the criteria together with the inbreeding coefficient to create a mate plan with 80 parents and generating 80 crosses per cycle (hereafter called *OCS*). We set maximum/minimum number of crosses of each parent to 8 and 1, respectively, whilst culling.parwise.k was to set to 0. A second set of simulations were done without G×E interaction.

The performance of each scenario was compared through the genetic gain measured at each cycle. In addition, the inbreeding rate (ΔF) was measured as the heterozygosity level, as follows (Falconer and Mackey, 1996):

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}}$$

where F_t measures inbreeding based on the expected heterozygosity of the SNP markers:

$$F_t = 1 - \frac{1}{m}\sum_i \sum_j p_{ij}^2$$

where *m* is the number of markers and p_{ij} is the frequency of the j^{th} allele of the i^{th} marker.

4. Results

Here, we compare **SMate** with a maize recurrently selection program by means of 25years simulation with 50 simulation replicates. In each cycle, 80 parents were selected and randomly mated (80 crosses) in the benchmark scenario, while in the scenario with **SMate** we assigned the crosses using the usefulness as criteria together with the inbreeding rates (Figure 2).

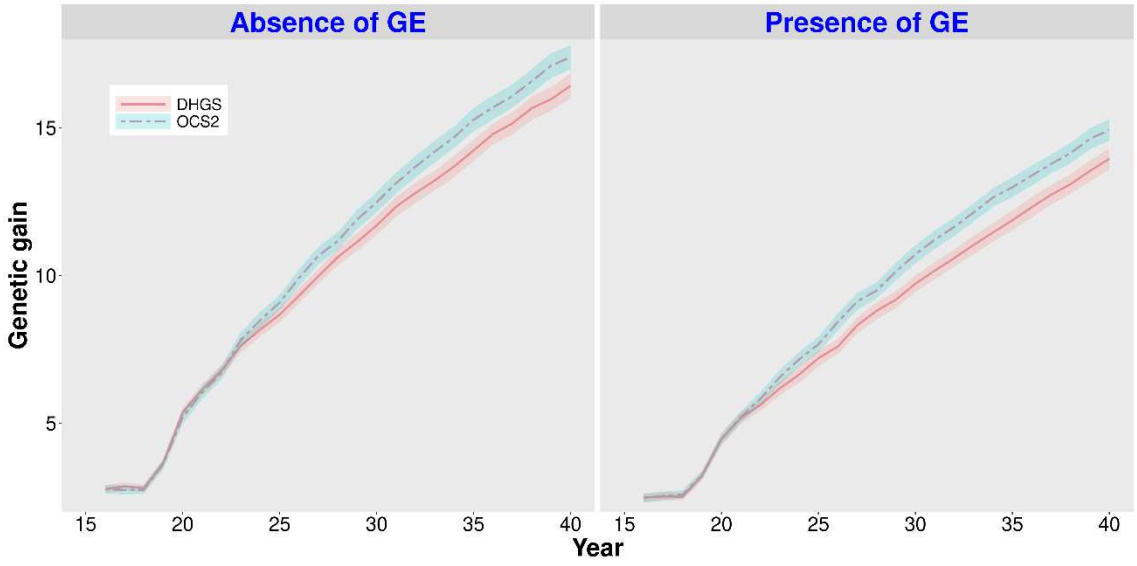


Figure 4 - Genetic gain overtime for two scenarios of a simulated maize breeding. A) absence of genotype-by-environment interaction and B) presence of genotype-by-environment interaction. The scenarios are plotted as the mean of the parents for each cycle, with 50 replicates. The shading around the line represents the standard error of the mean. DHGS: doubled haploid scenario with genomic selection and random crosses; OCS: doubled haploid scenario with genomic selection and optimal cross selection using **SMate** package.

The population mean performance over a 25-year simulation period showed that the OCS scenario resulted in a higher long-term genetic gain compared to the benchmark scenario

(5.90% and 6.95% superior for scenarios with absence and presence of G×E interaction, respectively, Table 1). Based on the average observed heterozygosity, the inbreeding rates of DHGS scenarios (absence and presence of G×E) were higher than the OCS scenarios (6.54 vs. 4.20 6.30 vs 1.84, for absence and presence of G×E, respectively).

These results suggest that the use of **SMate** controls the inbreeding in a long-term impacting the long-term genetic gain in breeding programs. However, further research is needed to consider other scenarios such as trait heritabilities, the influence of QTL number controlling the traits, marker panel size, and larger number of traits.

Table 1 - Genetic mean for both scenarios implemented in the simulations. The values referred to the year 40. DHGS: doubled haploid scenario with genomic selection and random crosses; OCS: doubled haploid scenario with genomic selection and optimal cross selection using **SMate** package. ΔF is the rate of inbreeding based on the observed heterozygosity for the year 40. Gain (%) represents the gain in percentage comparing DHGS and OCS scenarios.

Scenario	DHGS		OCS		Gain (%)
	Gain	ΔF*100	Gain	ΔF*100	
Absence of GE	16.42	6.54	17.39	4.20	5.90%
Presence of GE	13.95	6.30	14.92	1.84	6.95%

A pipeline for cross prediction and optimization in detailed in Supplementary material S3. **SMate** has a toy dataset that comes together with the pipeline before mentioned. This dataset comprises a A trait with heritability of 0.5, simulated in AlphaSimR package. A population of 100 doubled haploid lines were used as potential parents. An additive trait was simulated, with genetic mean equals to 10, genetic variance equals to 5. Five objects are available: (i) a total of 3000 SNP markers, coded as 2,1,0, (ii) the additive effects for each SNP, (iii) a G matrix (VanRaden 2008) together with (iv) an LD matrix, and with (v) BLUPs for all individuals.

5. Discussion

The **SMate** is a friendly R package for optimum cross selection, suitable for use by any breeding program focusing on long-term genetic gain. The main idea is to balance genetic mean and genetic gain (by means of usefulness) and inbreeding coefficients rates (by means of coancestry). The results indicated that the **SMate** can control the rates of inbreeding and return long-term genetic gains, even applying genomic selection in breeding program. This benefit of cross optimization were reported for animal and for plant breeding programs (Allier et al., 2019, Gorjanc et al. 2018, Pocrnic et al. 2023). Altogether, this reiterates the benefits of optimal cross selection and the importance of developing tools that can help breeding programs to implement such tool.

SMate was written in R language, a popular language within biological, statistical, and genetical data analyses' field. The package could be a particular interest of plant breeding programs, once most of them already uses R packages for routine data analyses, such as Asreml-R (Butler *et al.* 2009), BGLR (Pérez and de Los Campos 2011; Pérez-Rodríguez and de los Campos 2022), AGHMatrix (Amadeu *et al.* 2016), Sommer (Covarrubias-Pazaran 2016), and rrBLUP (Endelman 2011). Then, the user will easily connect **SMate** with the currently breeding pipeline.

SMate can be used for a wide of breeding program and situations. For example, breeding programs that breeds for several traits at time, once **SMate** permits the estimation of usefulness criterion for a group of traits, allowing the optimization under the criteria that is represented by the usefulness index. In addition, for several crops, dominance effects are impactful in the selection of potential crosses. The **SMate** enables the prediction of usefulness for quantitative traits controlled by additive and additive plus dominance effects.

Furthermore, even in breeding programs that do not use frequently genomic selection and there are no molecular markers available for the candidate parents, the cross optimization is still

possible. Generally, they keep track of pedigree information for the individuals in the breeding program. Then, the use of the BLUP solely as criteria for optimization and the **A** matrix for inbreeding constrain is encouraged. Such scenario could contribute to small breeding programs, generating information to data-driven decisions.

Future implementations that are target by the **SMate** development are regarding additional features and improvements in the already implemented functions. For example, the implementation of cross variance prediction in another computer language such as C++ via Rcpp package (Eddelbuettel and François 2011) together with the core optimization, should increase the package speed. In addition, a validation of the cross-prediction outcome using real data is needed, even though the results from simulation demonstrated to be superior in the long term. It will increase the reliability of the prediction so far only tested via simulations.

The optimization algorithm here used seems to work well for a few thousand of crosses. However, for many crosses it can be time expensive. One alternative should be to test other algorithms such as pareto frontier (Akdemir and Sánchez 2016; Gorjanc and Hickey 2018; Allier *et al.* 2019a), genetic algorithms (Gebregiwergis *et al.* 2020), and simulated annealing, that have being tested for cross optimization and mate allocation in plant and animal breeding. Altogether, those improvements can leverage the cross prediction and optimization and increase the power of the **SMate** for large datasets and broad community of users.

6. Conclusion

Here, we described **SMate** package that predicts crosses and can optimize usefulness and inbreeding rates in breeding programs. The package enables both, animal and plant breeding programs to achieve long-term genetic gains while using genomic selection and facilitates new research opportunities.

Acknowledgements

This work was supported by the National Institute of Food and Agriculture SCRI 2018-51181-28419 and AFRI 2019-05410 to M.F.R.R.). We also thank the financial support from the Brazilian Government through the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) through the CAPES-PrInt scholarship. This study was financed in part by the CAPES - Finance Code 001.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

The supplementary material can be found at the online version of this manuscript.

References

Akdemir, D., and J. I. Sánchez, 2016 Efficient breeding by genomic mating. Front. Genet. 7: 1–12.

Allier, A., C. Lehermeier, A. Charcosset, L. Moreau, and S. Teyssèdre, 2019a Improving short-and long-term genetic gain by accounting for within-family variance in optimal cross-selection. Front. Genet. 10: 1–15.

Allier, A., L. Moreau, A. Charcosset, S. Teyssèdre, and C. Lehermeier, 2019b Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. G3 Genes, Genomes, Genet. 9: 1469–1479.

Amadeu, R. R., C. Cellon, J. W. Olmstead, A. A. F. Garcia, M. F. R. Resende *et al.*, 2016 AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. Plant Genome 9:.

Bančič, J., C. R. Werner, R. C. Gaynor, G. Gorjanc, D. A. Odeny *et al.*, 2021 Modeling Illustrates That Genomic Selection Provides New Opportunities for Intercrop Breeding. Front. Plant Sci. 12:.

Bulmer, M., 1971 The effect of selection on genetic variability. Am. Nat. 105: 201–211.

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel, 2009 ASReml-R reference manual (version 3). State Queensland, Dep. Prim. Ind. Fish. Brisbane, Qld.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

Covarrubias-Pazaran, G., 2016 Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS One 11: e0156744.

Daetwyler, H. D., M. J. Hayden, G. C. Spangenberg, and B. J. Hayes, 2015 Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity. Genetics 200: 1341–1348.

Eddelbuettel, D., and R. François, 2011 Rcpp: Seamless R and C++ integration. J. Stat. Softw. 40: 1–18.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250–255.

Falconer, D. S., and T. F. C. Mackay, 1996 Introduction to quantitative genetics. Harlow, Essex, UK Longmans Green 3: 280.

Gaynor, R. C., G. Gorjanc, and J. Hickey, 2020 AlphaSimR: An R-package for Breeding

Program Simulations. G3 Genes|Genomes|Genetics 1–21.

Gebregiwergis, G. T., A. C. Sørensen, M. Henryon, and T. Meuwissen, 2020 Controlling Coancestry and Thereby Future Inbreeding by Optimum-Contribution Selection Using Alternative Genomic-Relationship Matrices. Front. Genet. 11: 1–8.

Goda, K., and F. Isik AgMate : An Optimal Mating Software for Monoecious Species versus other mate pair designing methods on long-term breeding of Pinus taeda. 0–3.

Gorjanc, G., R. C. Gaynor, and J. M. Hickey, 2018 Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor. Appl. Genet. 131: 1953–1966.

Gorjanc, G., and J. M. Hickey, 2018 AlphaMate : a program for optimizing selection , maintenance of diversity and mate allocation in breeding programs. Bioinformatics 34: 3408–3411.

Jannink, J., 2010 Dynamics of long-term genomic selection. Genet. Sel. Evol. 42: 1–11.

Kinghorn, B. P., 2011 An algorithm for efficient constrained mate selection. Genet. Sel. Evol. 43: 1–9.

Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656–4663.

Lehermeier, C., S. Teyssèdre, and C. C. Schön, 2017 Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. Genetics 207: 1651–1661.

Marinho, C. D., I. F. Coelho, M. A. Peixoto, G. A. Carvalho Júnior, and M. F. R. Resende Jr, 2022 Genomic selection as a tool for maize cultivars development. Rev. Bras. Milho e Sorgo 21:.

Meuwissen, T. H. E., and A. K. Sonesson, 1998 Maximizing the response of selection with a predefined rate of inbreeding: overlapping generations. J. Anim. Sci. 76: 2575–2583.

Meuwissen, T. H. E., A. K. Sonesson, G. Gebregiwergis, and J. A. Woolliams, 2020 Management of Genetic Diversity in the Era of Genomics. Front. Genet. 11: 1–16.

Pérez-Rodríguez, P., and G. de los Campos, 2022 Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package (E. Chesler, Ed.). Genetics 222:.

Pérez, P., and G. de Los Campos, 2011 BGLR: A Statistical Package for Whole Genome Regression and Prediction. 21.

Powell, O., R. C. Gaynor, G. Gorjanc, C. R. Werner, and J. M. Hickey, 2020 A two-part strategy using genomic selection in hybrid crop breeding programs. bioRxiv 1–46.

Quaas, R. L., 1988 Additive Genetic Model with Groups and Relationships. J. Dairy Sci. 71: 1338–1345.

R Development Core Team, 2022 R: A language and environment for statistical computing.

Ragsdale, A. P., and S. Gravel, 2019 Unbiased estimation of linkage disequilibrium from unphased data. Mol. Biol. Evol. 37: 923–932.

Sabadin, F., J. C. DoVale, J. D. Platten, and R. Fritsche-Neto, 2022 Optimizing self-pollinated crop breeding employing genomic selection: From schemes to updating training sets. Front. Plant Sci. 13:.

Sandhu, K. S., S. S. Patil, M. Aoun, and A. H. Carter, 2022 Multi-Trait Multi-Environment Genomic Prediction for End-Use Quality Traits in Winter Wheat. Front. Genet. 13: 1–14.

Schnell, F. W., and H. F. Utz, 1976 F1 Leistung und Elternwahl in der Zuchtung von Selbstbefruchtern. Ber Arbeitstag Arbeitsgem Saatzuchtleiter.

Sonesson, A. K., J. A. Woolliams, and T. H. E. Meuwissen, 2012 Genomic selection requires

genomic control of inbreeding. Genet. Sel. Evol. 44: 1–10.

Toro, M. A., B. Villanueva, and J. Fernández, 2014 Genomics applied to management strategies in conservation programmes $. Livest. Sci. 166: 48–53.

Troyer, A. F., and E. J. Wellin, 2009 Heterosis Decreasing in Hybrids: Yield Test Inbreds. Crop Sci. 49: 1969–1976.

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

Vieira, C. C., R. Persa, P. Chen, and D. Jarquin, 2022 Incorporation of Soil-Derived Covariates in Progeny Testing and Line Selection to Enhance Genomic Prediction Accuracy in Soybean Breeding. Front. Genet. 13: 1–15.

Walsh, B., and M. Lynch, 2018 *Evolution and selection of quantitative traits*. Oxford University Press.

Wolfe, M. D., A. W. Chan, P. Kulakow, I. Rabbi, and J. L. Jannink, 2021 Genomic mating in outbred species: Predicting cross usefulness with additive and total genetic covariance matrices. Genetics 220:.

Supplementary material

Supplementary file S1

SMate: An R-package for cross prediction and optimization using genomic selection

Marco Antônio Peixoto[1,2], Rodrigo Amadeu[3], Leonardo Lopes Bhering[1], Patrício R. Munoz[4], Felipe Ferrão[4], Márcio F. R. Resende[2,*]

[1] Laboratório de Biometria, Universidade Federal de Viçosa, Viçosa, Minas Gerais State, Brazil.

[2] Sweet Corn Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

[3] Bayer Crop Science, Chesterfield, Missouri State, United States

[3] Blueberry Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

*** Corresponding author**

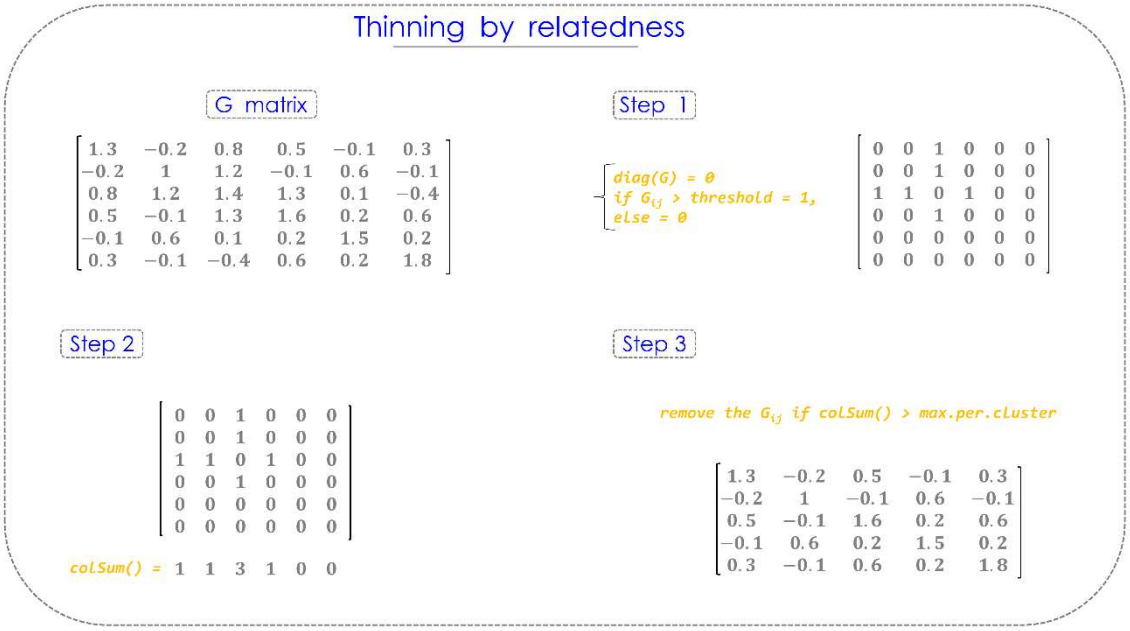


Figure S1 - Simplified representation of the function '`relateThinning`()' from the SimpleMating. The G matrix is coded in a matrix with 0 and 1 based on the level of relatedness of a target genotype with the others, based on a given *threshold*. Every value above the threshold receives the number 1, otherwise 0 (Step 1). After, all columns are summed (Step 2) and after one iteration per row number, we remove the genotype with sum higher than *max.per.cluster* limit (Step 3).

Supplementary file S2

SMate: R-package for cross prediction and optimization using genomic selection

Marco Antônio Peixoto[1,2], Rodrigo Amadeu[3], Leonardo Lopes Bhering[1], Patrício R. Munoz[4], Felipe Ferrão[4], Márcio F. R. Resende[2,*]

[1] Laboratório de Biometria, Universidade Federal de Viçosa, Viçosa, Minas Gerais State, Brazil.

[2] Sweet Corn Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

[3] Bayer Crop Science, Chesterfield, Missouri State, United States

[3] Blueberry Breeding and Genomics Lab, University of Florida, Gainesville, Florida State, United States

*** Corresponding author**

Formulas

Prediction of cross mean

For the prediction of the performance of every specific cross, we considered the prediction for both, only additive effects (estimated breeding value – *ebv*) and both additive and dominance effects (total genetic value – *tgv*). Then, the performance was predicted as follows (Falconer & Mackey):

$$\mu_{ebv} = \frac{GEBV_{P1} + GEBV_{P2}}{2}$$

$$\mu_{tgv} = \sum_{k=1}^{p} a_k(p_{ik} + \mathrm{q}_{ik} + \mathrm{y}_k) + \mathrm{d}_k[2\mathrm{p}_{ik}\mathrm{q}_{ik} + \mathrm{y}_k(\mathrm{p}_{ik} - \mathrm{q}_{ik})]$$

where *GEBV* represents the genomic *ebv* for each parent, *a* and *d* represents the additive and dominance effects of the markers, respectively, *p* and *q* represent the allele frequency for a marker, and y represents the different in the allele frequency between individuals. The same formulas were used to prediction the performance of several traits. Here, we used the information of them to build an index, as follows:

$$\mu_{ebv} = v' g_{ebv}$$

$$\mu_{tgv} = v' g_{tgv}$$

where *v* is a vector of 1 × *t* (*t* represents the number of traits) and g_{ebv} and g_{tgv} represents a matrix with the dimensions *n* × *t* (*n* represents the number of crosses), containing the *ebv* and *tgv*, respectively.

Prediction of cross variance

In a similar way than the prediction of the cross' mean, the cross variance was also implemented to the prediction of additive variance ($\hat{\sigma}_a^2$) and dominance variance ($\hat{\sigma}_d^2$). So, the

variances of a specific cross were calculated following the implementation of Lynch and Walsh (1998), as follows:

$$\hat{\sigma}_a^2 = \alpha' D \alpha$$

$$\hat{\sigma}_d^2 = \beta' D^2 \beta$$

where α represents the vector of additive effects of the markers, β represents a vector of the dominance effects of the markers, and D is a matrix of variance and covariance estimated for each cross (being it $n \times n$, where n is the number of parent candidates). The squared D matrix is a simple multiplication of the elements of D by itself (D ʘ D).

The variance and covariance matrix (D) was proposed by Lehermeier et al. (2017). It was based on the linkage disequilibrium (LD) between loci i and j (M_{ij}), derived from two parental lines (Figure 1) and in the expected frequency of recombinants in the generation 1 (C_{ij}), as follows:

$$D = 4M_{ij}(1 - 2C_{ij})$$

So far, the propositions for the calculus of the frequency of recombinant (C_{ij}) accounted for a genetic map (Lehermeier et al., 2017; Allier et al., 2019; Wolfe et al., 2021). However, genetic maps are not available for all crops. Thinking about that, we propose to build the LD at population level and subtract one unit from the LD matrix, that should represent the frequency of recombination in the next generation.

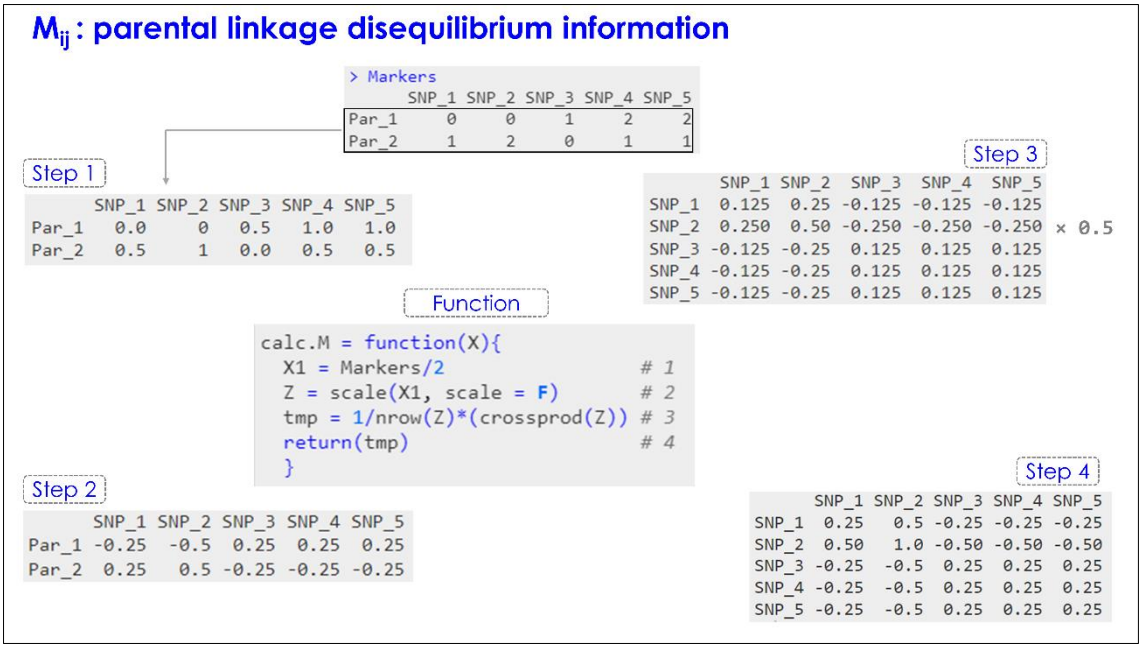


Figure 1 – parental linkage disequilibrium estimates. The function **`calc.M()`** returns a matrix in the same dimension of the number of markers (markers x markers). Step 1: divide the markers by 2. Step 2: center the markers. Step 3: multiply the markers centralized matrix by its transposed matrix and multiply the outcome by 0.5. Step 4: Return the matrix.

When we investigate a multi-trait framework, the variance estimated for several traits follows the construction of covariance matrix with the dimension of the number of traits analyzed ($t \times t$). For such, we build a matrix of covariances (T) by calculating in the first step every trait variance and the covariances between each pair of traits, as follows:

$$\hat{\sigma}^2_{T1} = \alpha_{T1}{}' D \alpha_{T1}$$

$$\hat{\sigma}^2_{T2} = \alpha_{T2}{}' D \alpha_{T2}$$

$$\hat{\sigma}^2_{T1T2} = \alpha_{T1}{}' D \alpha_{T2}$$

$$T = \begin{bmatrix} \hat{\sigma}^2_{T1} & \hat{\sigma}^2_{T1T2} \\ \hat{\sigma}^2_{T1T2} & \hat{\sigma}^2_{T2} \end{bmatrix}$$

Then, we use this T matrix in the following equation for the estimation of a cross variance for a multi-trait framework ($\hat{\sigma}^2_{SI}$):

$$\hat{\sigma}^2_{SI} = v'Tv$$

In this case, we were able to implement the same calculation for both, additive effects (where D is equal to D) and dominance effects (where D is equal to D*D).

Usefulness criterion

To identify the crosses that generate the individuals with higher performance can boost the mean of the next generation. The usefulness criterion (UC) was proposed by Schnell and Utz (1975) and represents an alternative to maximize the performance of the next generation once it combines the genetic mean and the genetic gain of it. Its implementation has proved to returns higher gains than the selection of combinations based on only GEBV (Lehermeier et al., 2017; Allier et al., 2019). So, UC follows the equation:

$$UC = \mu + ih\hat{\sigma}_g$$

where μ is the mean of a cross (genetic mean), *i* represents the intensity of selection, *h* is the squared root of the heritability, and $\hat{\sigma}_g$ represents the progeny genetic standard deviation within family for a target cross ($ih\hat{\sigma}_g$ is the genetic gain). Selection intensity (*i*) represents the mean of deviations from the population mean, and can be implemented as (Falconer & Mackay, 1996):

$$i = \Phi(x)/p,$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

where $\Phi(x)$ is the probability density function of the standard normal distribution for *x*, *p* is the selection proportion, and *x* is the mean deviation of the selected group:

$$x = \sum(y_i - \hat{y})/n$$

where y_i is the i^{th} selected observation, $\hat{y}$ is the population mean, and *n* refers to the number of selected observations. For the implementation in the analyses, we use the R formulae, as follows (Mackey, 2020):

$$i = qnorm(dnorm(1 - p))/p,$$

where *qnorm* is the quantile normal distribution and *dnorm* is the density of the normal distribution. It returns to us the selection intensity for any level of selection proportion (*p*).

References

Allier, A., Moreau, L., Charcosset, A., Teyssèdre, S., & Lehermeier, C. (2019). Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. *G3: Genes, Genomes, Genetics*, *9*, 1469–1479. https://doi.org/10.1534/g3.119.400129

Falconer, D.S., & Mackay, T.F.C. (1996). Introduction to quantitative genetics. *Harlow, Essex, UK: Longmans Green*, *3*, 280

Lehermeier, C., Teyssèdre, S., & Schön, C.C. (2017). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics*, *207*, 1651–1661. https://doi.org/10.1534/genetics.117.300403

Mackey, I. (2020). Selection Intensity

Wolfe, M.D., Chan, A.W., Kulakow, P., Rabbi, I., & Jannink, J.L. (2021). Genomic mating in outbred species: Predicting cross usefulness with additive and total genetic covariance matrices. *Genetics*, *220*. https://doi.org/10.1093/genetics/iyab122

Appendix S3 - Toy example

1. Installing (from github) and load the package

```
# install.packages("devtools")
# devtools::install_github("...")


## Load
source('aux_fun.R')
```

2. Example

The toy data here used came from a simulation using AlphaSimR package (Gaynor et al. 2020). A population of 100 doubled haploid lines were used as potential parents. An additive trait was simulated, with genetic mean equal to 10, genetic variance equal to 5, and heritability of 0.5. A total of 3000 markers were simulated.

```
## Data example
load("Example.RData")
```

1. Allele dosage matrix

```
# Markers
Markers[1:5,1:5]
```

```
   SNP1 SNP2 SNP3 SNP4 SNP5
G1    2    2    2    0    2
G2    2    2    0    0    2
G3    2    2    0    0    2
```

```
G4    2    2    0    0    2
G5    2    0    0    0    2
```

2. Relationship matrix

```
# Relationship matrix
G[1:5,1:5]
```

```
            G1           G2           G3           G4           G5
G1  1.95440487  1.074400818  0.233031310  0.127990797 -0.011229926
G2  1.07440082  2.007947808  0.071175288  0.020677664 -0.004912042
G3  0.23303131  0.071175288  2.025128818  1.179214069 -0.009957259
G4  0.12799080  0.020677664  1.179214069  2.046855068 -0.001366754
G5 -0.01122993 -0.004912042 -0.009957259 -0.001366754  2.009311380
```

3. Linkage disequilibrium matrix (LD)

```
# Linkage disequilibrium matrix
Mat.ld[1:5,1:5]
```

```
           SNP1   SNP10  SNP100 SNP1000 SNP1001
SNP1    1.00000 0.11574 0.00082 0.00327 0.01316
SNP10   0.11574 1.00000 0.01194 0.00218 0.00219
SNP100  0.00082 0.01194 1.00000 0.00843 0.06810
SNP1000 0.00327 0.00218 0.00843 1.00000 0.05573
SNP1001 0.01316 0.00219 0.06810 0.05573 1.00000
```

4. Markers effects

```
# Additive effects for the markers
head(addEff,10)
```

```
              [,1]
SNP1  -0.008823571
SNP2  -0.002925676
SNP3  -0.001385556
SNP4   0.009077221
SNP5  -0.006523342
SNP6   0.002473636
```

```
SNP7  -0.007839470
SNP8  -0.005418003
SNP9   0.004485988
SNP10 -0.003954733
```

5.Genetic value

```
# Genetic value for each individual
head(gen_value,10)
```

```
         [,1]
G1  12.155992
G2  12.210842
G3   8.536672
G4  10.568421
G5  10.536782
G6   9.884344
G7   9.602494
G8   7.526259
G9  10.402562
G10 10.912716
```

3. Thinning by relatedness

The **relateThinning** is the first function to be used and it is an optional step. It will increase the efficiency of the prediction by cutting individuals by relatedness and, consequently, decreases time for the prediction and cross optimization. It selects the top individuals within a family by the relatedness *threshold*. For example, if *max.per.cluster*=2 and *threshold*=0.5 the algorithm will identify all groups of individuals that shared more than 0.5 in the relationship matrix and select the top 2 individuals with highest criterion to be used as possible parents in the mating plan. Besides those arguments, we need to give to the function the relationship matrix (K) between the parents. It could be the relationship matrix based on markers information (G) or based on pedigree information (A). In addition, a *criteria* should be added, such as the estimated breeding value, total breeding value, or the BLUP of the candidates.

```
keep = relateThinning(K = G,
                      criterion = gen_value,
                      threshold=0.8,
                      max.per.cluster=1)
```

```
100 genotypes kept out of 100
```

4. Planning the crosses

After the thinning step, we are able to create all the potential crosses to be optimized.In the function **cross_plan** the argument K represents the relationship matrix of potential parents (G or A matrix). The argument $Indiv$ gives to the program the individuals to keep after the thinning process. The argument $Type$ represents the type of crosses that should be generated, and $half$ indicates that no reciprocals nor parents are included in the crosses.

```
plan = cross_plan(K = G,
                  Indiv = keep,
                  Type = 'half')
```

```
Number of crosses generated: 4950
```

```
head(plan)
```

```
  Par_1 Par_2 idcross
1    G1    G2   G1_G2
2    G1    G3   G1_G3
3    G1    G4   G1_G4
4    G1    G5   G1_G5
5    G1    G6   G1_G6
6    G1    G7   G1_G7
```

5. Usefulness criterion

For the calculus of the usefulness of each potential cross (see appendix S1 for more details), we use the function **usefA**. It predicts the usefulness for one trait controlled only by additive effects, following the equation:

$$UC = u + i\sigma_g$$

The inputs are the markers panel ($Markers$), coded as 0,1,2 for AA, Aa and aa, the additive effects for each one of the 3000 markers ($addEff$), the matrix of linkage disequilibrium ($Mat.ld$) for all candidates parents (100 parents), the intensity of selection (0.05), and the a plan with all potential crosses $plan$.

```r
cross_usef = usefA(Markers = Markers,
                   Markers.effA = addEff,
                   LDMat = Mat.ld,
                   Intsel = 0.05,
                   PlanV = plan)

head(cross_usef, 10)
```

```
     idcross Par_1 Par_2       Var     MuE     UCt
4410 G73_G93   G73   G93 12.878668 1.04165 8.44408
4203 G69_G93   G69   G93 12.765963 1.02911 8.39908
4411 G73_G94   G73   G94 12.195343 1.08674 8.29011
4204 G69_G94   G69   G94 12.097538 1.07420 8.24863
4355 G71_G93   G71   G93 12.328579 0.96791 8.21052
4356 G71_G94   G71   G94 11.661523 1.01300 8.05695
4415 G73_G98   G73   G98  9.966154 1.25323 7.76505
4208 G69_G98   G69   G98  9.874297 1.24068 7.72242
4360 G71_G98   G71   G98  9.500105 1.17949 7.53723
4414 G73_G97   G73   G97  9.057009 1.32835 7.53606
```

6. Optimization

For the optimization step, we need to organize the data into a data frame containing 4 columns: $P1$ represents the first parent; $P2$ Representing the second parent; Y represents the target vector to be optimized (usefulness of the cross or another criteria); K represents the covariance between pair of parents (these information can be accessed in the relationship matrix, G or A).

```r
# Melting the matrix G for pair-pair covariance (using the auxiliary function **meltK**)
Gmat_melted = meltK(G)

# Transform the info into a data frame
par_info = data.frame(idcross = paste0(Gmat_melted$P1, "_", Gmat_melted$P2),
                      K = Gmat_melted$K)

# Combining the information
df = merge(cross_usef, par_info, by = "idcross")

# Build up the input for the optimization
cand.crosses = data.frame(P1 = df$Par_1,
```

```
                          P2 = df$Par_2,
                          Y = df$UC,
                          K = df$K)

head(cand.crosses,10)
```

```
   P1   P2       Y           K
1  G1  G10 1.30135 -0.05286433
2  G1 G100 1.72526 -0.11340694
3  G1  G11 2.76706 -0.10777084
4  G1  G12 3.69125 -0.12404280
5  G1  G13 2.54194 -0.16381366
6  G1  G14 2.76373 -0.10754358
7  G1  G15 1.99167 -0.03136534
8  G1  G16 1.97006  0.04294934
9  G1  G17 2.67456 -0.01236624
10 G1  G18 4.83667 -0.12695175
```

The core optimization algorithm proposed is **selectCrosses**. Seven arguments need to be set prior to the optimization.

(i) *data*: represents the candidate crosses, together with the criteria to maximize and the covariance for each parent to minimize.
(ii) *K*: relationship matrix. The used can set both, a G matrix (that is constructed from markers) or a A matrix (that is constructed from pedigree information).
(iii) *n.cross*: number of crosses in the mating plan.
(iv) *max.cross*: maximum number of crosses per parent (expected).
(v) *min.cross*: minimum number of crosses per parent (this is the target, some parents might be used just one time).
(vi) *max.cross.to.search*: maximum number of rows in the input to search for the solution
(vii) *culling.pairwise.k*: don't allow crosses with more than this threshold of relatedness.

```
## Optimization
maxGainPlan = selectCrosses(data=cand.crosses,
                            K=G,
                            n.cross=50,
                            max.cross = 5,
                            min.cross = 1,
                            max.cross.to.search=100000,
                            culling.pairwise.k = 0.275)
```

```r
# Statistics of the plan
maxGainPlan[[1]]
```

```
  culling.pairwise.k target.Y    target.K
1              0.275 6.744127 -0.02018028
```

```r
# Mating plan
head(maxGainPlan[[2]],10) #mating plan
```

```
      P1  P2       Y          K
4410 G73 G93 8.44408 -0.1017711
4203 G69 G93 8.39908 -0.2582183
4411 G73 G94 8.29011 -0.1947213
4204 G69 G94 8.24863 -0.2420827
4355 G71 G93 8.21052 -0.2509005
4356 G71 G94 8.05695 -0.2620363
4415 G73 G98 7.76505 -0.1888579
4208 G69 G98 7.72242 -0.2362193
4360 G71 G98 7.53723 -0.1061800
4414 G73 G97 7.53606 -0.1974030
```