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

New strategies for implementing genomic selection in breeding programs of clonally propagated crops

Lorena Guimarães Batista

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

Piracicaba
2019

Lorena Guimarães Batista
Agronomist

New strategies for implementing genomic selection in breeding programs of clonally propagated crops

Advisor:
Prof. Dr. **GABRIEL RODRIGUES ALVES MARGARIDO**

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

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação

DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Batista, Lorena Guimarães

New strategies for implementing genomic selection in breeding programs of clonally propagated crops / Lorena Guimarães Batista. - - Piracicaba, 2019.

57 p.

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

1. Seleção Genômica 2. Cana-de-açúcar 3. Dosagem alélica 4. Poliploides 5. Índice de seleção I. Título

ACKNOWLEDGEMENTS

To my parents and siblings, for the guaranteed source of love that kept me going through all these years.

To my colleagues at the Laboratory of Bioinformatics Applied to Bioenergy, for sharing with me the countless moments of back pain, for praying with me during days of strong wind and possible power blackouts, for sharing with me the taste for nerdy jokes and despicable puns, for keeping the supply of coffee, tea and biscuits running, for bearing with my constant random rants in the lab, and for being such a nice crew to work with.

To my beloved friends from my home state, is really nice to see that our friendship survived the distance and that we keep surviving the impending doom of adulthood knocking our doors. We must gather and drink to that asap.

To the beloved friends I made here in Piracicaba. You really made my life way funnier during these years. I appreciate every single meal that we had in our flat, every card game and every beer that helped soothe some stressful times.

To my colleagues at the Genetics and Plant Breeding graduate program. Special thanks to the people at the Statistical Genetics Lab and the Allogamous Plant Breeding Lab.

To my supervisor Gabriel for all the support, patience and always being very helpful. I learned a lot from you, bad jokes included.

To my foreign supervisor John and colleagues at the University of Edinburgh. Also to Anete and Augusto, I owe you all big time.

And, finally, to CAPES and CNPq, for providing the financial support for my research.

SUMÁRIO

RESUMO

Novas estratégias para a implementação de seleção genômica em programas de melhoramento de espécies de propagação vegetativa

A seleção genômica consiste no uso de efeitos preditos de marcadores genéticos para predizer os valores genéticos e/ou genotípicos de indivíduos genotipados. Desta forma, a seleção de genótipos superiores pode ser feita baseada apenas em valores genéticos preditos, reduzindo a necessidade de avaliações fenotípicas subsequentes. Isto representa um grande avanço em termos de custos e eficiência da seleção em programas de melhoramento de todos os tipos de culturas. No primeiro capítulo deste trabalho, nós exploramos uma das maneiras com que a seleção genômica pode ser utilizada para aumentar a eficiência no melhoramento simultâneo para múltiplos caráteres em espécies de propagação vegetativa. Utilizando simulações estocásticas, nós mostramos que um índice de seleção econômico deve ser utilizado no lugar da eliminação independente (*independent culling*). Os resultados mostram que o uso da seleção genômica pode tornar o custo-benefício da eliminação independente obsoleto se indivíduos em gerações iniciais forem genotipados e predições acuradas para todos os caráteres estiverem disponíveis desde o início. Apesar dos potenciais benefícios de realizar a seleção com base em valores genéticos preditos, para algumas espécies de propagação vegetativa a complexidade de seus genomas é um fator limitante para a efetiva implementação da seleção genômica em programas de melhoramento. Considerando que incluir a informação de dosagem alélica melhorou a performance de modelos de seleção genômica em espécies autotetraploides, nosso objetivo no segundo capítulo deste trabalho foi avaliar a acurácia da predição genômica com informação de dosagem alélica em cana-de-açúcar, que é uma complexa espécie poliploide. Neste capítulo, nós expandimos modelos GBLUP de seleção genômica desenvolvidos para autotetraploides para incluir níveis mais altos de ploidia. Dois modelos foram utilizados, um modelo com somente efeitos aditivos e um modelo com efeitos aditivos e efeitos de dominância digênica. Nós observamos uma modesta melhora na performance do modelo preditivo quando estimativas de ploidia e dosagem alélica foram incluídas, indicando que esta é uma possível maneira de aprimorar a seleção genômica em cana-de-açúcar. Os resultados obtidos nos dois estudos podem auxiliar pesquisadores e melhoristas de espécies de propagação vegetativa, abrindo portas para novas pesquisas e indicando as maneiras mais eficientes para implementação da seleção genômica.

Palavras-chave: Seleção genômica; Cana-de-açúcar; Dosagem alélica; Poliploides; Índice de seleção

ABSTRACT

New strategies for implementing of genomic selection in breeding programs of clonally propagated crops

Genomic selection consists of using predicted effects of genetic markers to predict breeding values and/or genotypic values of genotyped individuals. With this approach, selection can be carried based only on those predicted breeding values, reducing the need for further phenotypic evaluations. This represents a great advance in terms of cost and effectiveness of selection in breeding programs of all kinds of crops. In the first chapter of this work, we explore one of the ways genomic selection can be used to increase efficiency when breeding clonally propagated crops for multiple traits. Using stochastic simulations, we show that an economic selection index should be preferred over independent culling. Our results show that the use of genomic selection may render the cost-efficiency benefit of independent culling obsolete when all early generation individuals are genotyped and accurate prediction of all traits becomes available simultaneously. Despite the potential benefits of selecting based on predicted breeding values, for some clonally propagated species the complexity of their genomes limits the implementation of genomic selection in breeding programs. Since including allele dosage information has been shown to improve performance of genomic selection models in autotetraploid species, our objective in the second chapter of this work was to assess the accuracy of genome-wide prediction in the highly complex polyploid sugarcane when incorporating allele dosage information. In this chapter, we expanded GBLUP genomic selection models developed for autotetraploids to include higher levels of ploidy. Two types of model were used, one with additive effects only and one with additive and digenic dominance effects. We observed a modest improvement in the performance of the prediction model when ploidy and allele dosage estimates were included, indicating that this is a possible way of improving genomic selection in sugarcane. The results obtained in both studies can assist researchers and breeders of clonally propagated crops, opening new research opportunities and indicating the most efficient ways to implement genomic selection.

Keywords: Genomic selection; Sugarcane; Allele dosage; Polyploids; Selection Index

1. INTRODUCTION

Genomic selection is an approach that aims to increase the efficiency of selection in a breeding program, especially when selecting for complex traits, which usually requires evaluation of field trials in several locations and years, an expensive and time-consuming process (Heffner *et al.* 2009a). The method consists of using a training population that is both genotyped and phenotyped to predict the effect of genetic markers widely spread throughout the genome. The estimated effects can then be used to predict the phenotype of genotyped individuals in terms of estimated breeding values or estimated genotypic values (Meuwissen *et al.* 2001a). This allows selection to be carried without the need for further phenotypic evaluations, therefore shortening the time needed for selection of superior genotypes. Genomic selection can be implemented in any population of interest and has been successfully applied in several crop breeding programs (Bernardo and Yu 2007a; Heffner *et al.* 2009a; Crossa *et al.* 2010; Resende *et al.* 2012; Duhnen *et al.* 2017).

The breeding schemes for clonally propagated crops generally comprise several sequential steps that are carried over several years. The overall scheme can be simply summarized in generating genetic variation through crosses and subsequently selecting clones in the resulting F_1 progenies, which is done in several stages of selection, until the most promising clones can be released as cultivars, which often are also the candidate parents for the next breeding cycle (Simmonds 1979). Typically, the initial stages of selection include a large number of individuals to be evaluated and, to increase program efficiency, individuals are initially culled based on traits that can be phenotyped at a lower cost and, as the number of individuals decreases and higher-cost phenotyping becomes feasible, selection is performed for other traits in later stages of selection (Grüneberg *et al.* 2009a). In the context of genomic selection, the plant breeder may no longer be forced to cull if the individuals are genotyped. Because accurate prediction of breeding values of genotyped individuals can become available simultaneously for all traits, a selection index could be used instead of independent culling.

The selection index method involves selection for all traits simultaneously based on a linear or non-linear combination of individual traits weighted by their importance for the breeding objective (Hazel and Lush 1942). Theoretically, the selection index is the most efficient method of selection for multiple traits (Hazel and Lush 1942; Young 1961a). A major drawback of independent culling in comparison to selection index is that independent culling, if strictly applied, will not select individuals below the threshold for one single trait despite being exceptional for all other traits, while the use of a selection index makes it possible to retain those individuals (Bernardo 2010). Thus, especially for the selection of parents in breeding programs, the use of a selection index instead of independent culling might lead to higher genetic gains across cycles of selection, particularly when the correlation between traits is unfavorable. In this context, in the first chapter we investigated the gains over several generations of genomic selection in a recurrent selection breeding program using either a selection index or independent culling. We used simulations of recurrent breeding programs to evaluate and compare both strategies with the purpose of quantifying the magnitude of the difference between the different selection methods.

In the second chapter, we focus on the practical deployment of genomic selection in sugarcane breeding programs. Sugarcane cultivars are auto-allopolyploids, with 100 to 130 chromosomes and different number of chromosome copies between homology groups (i.e., aneuploid) (D'Hont *et al.* 1996, 1998). Due to this extremely complex genome structure, the majority of genetic studies in sugarcane use either dominant or single-dosage codominant markers (Wu *et al.* 1992; Huckett and Botha 1995; Besse *et al.* 1998; Nair *et al.* 2002; Gouy *et al.* 2013; Aitken *et al.* 2014; Racedo *et al.* 2016; Balsalobre *et al.* 2017), i.e., polymorphisms that were either detected in a presence/absence fashion or that could only be detected in one chromosome per homology group, without considering information of other allele dosage levels. With the recent possibility of estimating the ploidy and allele dosage of markers (Serang *et al.* 2012; Garcia *et al.* 2013; Mollinari and Serang 2015), markers with higher dosages can be used in studies of polyploid species. Also, given that recent studies have shown that allele dosage information can improve the accuracy of genomic selection models in autotetraploid species (Slater *et al.* 2016, 2016; de Bem Oliveira *et al.* 2018; Hawkins and Yu 2018; Endelman *et al.* 2018), our objective in the second chapter was to assess the accuracy of genomic selection in sugarcane when incorporating allele dosage information.

Overall, we tackled possible ways to improve the implementation of genomic selection in breeding programs of clonally propagated crops in two levels. First, in terms of rearranging breeding schemes, by replacing the use of independent culling for an economic selection index; second, in terms of adapting genotyping techniques and genomic selection models to the complexity of the polyploid sugarcane genome, by estimating both ploidy and allele dosage of markers and incorporating this information in the prediction model.

REFERENCES

Aitken, K. S., M. D. McNeil, S. Hermann, P. C. Bundock, A. Kilian *et al.*, 2014 A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput Diversity Array Technology (DArT) markers. BMC genomics 15: 152–152.

Balsalobre, T. W. A., G. da Silva Pereira, G. R. A. Margarido, R. Gazaffi, F. Z. Barreto *et al.*, 2017 GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18:.

de Bem Oliveira, I., M. F. Resende, F. Ferrao, R. Amadeu, J. Endelman *et al.*, 2018 Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. bioRxiv.

Bennett, G. L., and L. A. Swiger, 1980 Genetic variance and correlation after selection for two traits by index, independent culling levels and extreme selection. Genetics 94: 763–775.

Bernardo, R., 2010 *Breeding for quantitative traits in plants*. Stemma Press, Woodbury.

Bernardo, R., and J. Yu, 2007a Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Science 47: 1082–1090.

Bernardo, R., and J. Yu, 2007b Prospects for Genomewide Selection for Quantitative Traits in Maize All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any informati. Crop Science 47: 1082–1090.

Besse, P., G. Taylor, B. Carroll, N. Berding, D. Burner *et al.*, 1998 Assessing genetic diversity in a sugarcane germplasm collection using an automated AFLP analysis. Genetica 104: 143–153.

Bulmer, M. G., 1971 THE EFFECT OF SELECTION ON GENETIC VARIABILITY. 105: 1–88.

Casler, M. D., and K. P. Vogel, 1999 Accomplishments and Impact from Breeding for Increased Forage Nutritional Value. Crop Science 39: 12–20.

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

Clark, S. A., B. P. Kinghorn, J. M. Hickey, and J. H. J. Van Der Werf, 2013 The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics Selection Evolution 45: 1–8.

Cotterill, P. P., and J. W. James, 1981 Optimising two-stage independent culling selection in tree and animal breeding. Theoretical and Applied Genetics 59:.

Cowling, W., and L. Li, 2018 *Turning the heat up on independent culling in crop breeding.*

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186: 713.

D'Hont, A., L. Grivet, P. Feldmann, J. C. Glaszmann, S. Rao *et al.*, 1996 Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. Molecular and General Genetics MGG 250: 405–413.

D'Hont, A., D. Ison, K. Alix, C. Roux, and J. C. Glaszmann, 1998 Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. Genome 41: 221–225.

Ducrocq, V., and J. J. Colleau, 1989 Optimum truncation points for independent culling level selection on a multivariate normal distribution, with an application to dairy cattle selection. Genetics Selection Evolution 21: 185.

Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres *et al.*, 2017 Genomic Selection for Yield and Seed Protein Content in Soybean: A Study of Breeding Program Data and Assessment of Prediction Accuracy. Crop Science 57: 1325.

Duvick, D. N., and K. G. Cassman, 1999 Post–Green Revolution Trends in Yield Potential of Temperate Maize in the North-Central United States. Crop Science 39: 1622–1630.

Endelman, J. B., C. A. S. Carley, P. C. Bethke, J. J. Coombs, M. E. Clough *et al.*, 2018 Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato. Genetics 209: 77.

Erskine, W., P. C. Williams, and H. Nakkoul, 1985 Genetic and environmental variation in seed yield, seed size and cooking quality of lentil. Field Crops Research 12: 153–161.

Falconer, D. S., T. F. Mackay, and R. Frankham, 1996 Introduction to Quantitative Genetics (4th edn). Trends in Genetics 12: 280.

Garcia, A. A. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva *et al.*, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific Reports 3: 3399.

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 Science 57: 2372–2386.

Gaynor, R. C., G. Gorjanc, and D. L. Wilson AlphaSimR: An R Package for Breeding Program Simulations. Manuscr Prep.

Goddard, M. E., 1983 Selection indices for non-linear profit functions. Theoret. Appl. Genetics 64: 339–344.

Gorjanc, G., R. C. Gaynor, and J. M. Hickey, 2017 Optimal cross selection for long-term genetic gain in two- part programs with rapid recurrent genomic selection. bioRxiv.

Gouy, M., Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal *et al.*, 2013 Experimental assessment of the accuracy of genomic selection in sugarcane. Theoretical and Applied Genetics 126: 2575–2586.

Grüneberg, W., R. Mwanga, M. Andrade, and J. Espinoza, 2009a Selection methods. Part 5: Breeding clonally propagated crops. Plant Breeding and Farmer Participation 275–322.

Grüneberg, W., R. Mwanga, M. Andrade, and J. Espinoza, 2009b Selection methods. Part 5: Breeding clonally propagated crops. Plant Breeding and Farmer Participation 275–322.

Hawkins, C., and L.-X. Yu, 2018 Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. The Crop Journal.

Hazel, L. N., and J. L. Lush, 1942 The Efficiency of Three Methods of Selection. Journal of Heredity 33: 393–399.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009a Genomic Selection for Crop Improvement All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval syst. Crop Science 49: 1–12.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009b Genomic Selection for Crop Improvement All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval syst. Crop Science 49: 1–12.

Hickey, J. M., T. Chiurugwi, I. Mackay, W. Powell, and I. G. S. in C. B. P. W. Participants, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics 49: 1297.

Huckett, B. I., and F. C. Botha, 1995 Stability and potential use of RAPD markers in a sugarcane genealogy. Euphytica 86: 117–125.

Itoh, Y., 1991 Changes in genetic correlations by index selection. Genetics selection evolution GSE 23: 301–308.

Kato, T., and K. Takeda, 1996 Associations among Characters Related to Yield Sink Capacity in Space-Planted Rice. Crop Science 36: 1135–1139.

Kwon, S. H., and J. H. Torrie, 1964 Heritability of and interrekationships among traits of two soybean population.

Meredith, W. R., and R. R. Bridge, 1971 Breakup of Linkage Blocks in Cotton, Gossypium hirsutum L.1. Crop Science 11: 695–698.

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

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

Mollinari, M., and O. Serang, 2015 Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms. Batley J. (eds) Plant Genotyping. Methods in Molecular Biology (Methods and Protocols) 1245:.

Nair, N. V., A. Selvi, T. V. Sreenivasan, and K. N. Pushpalatha, 2002 Molecular diversity in Indian sugarcane cultivars as revealed by randomly amplified DNA polymorphisms. Euphytica 127: 219–225.

Namkoong, G., 1970 Optimum Allocation of Selection Intensity in Two Stages of Truncation Selection. Biometrics 26: 465.

Racedo, J., L. Gutiérrez, M. F. Perera, S. Ostengo, E. M. Pardo *et al.*, 2016 Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. BMC plant biology 16: 142–142.

Resende, M. D. V., M. F. R. Resende, C. P. Sansaloni, C. D. Petroli, A. A. Missiaggia *et al.*, 2012 Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194: 116–128.

Rharrabti, Y., S. Elhani, V. Martos-Núñez, and L. F. García del Moral, 2001 Protein and Lysine Content, Grain Yield, and Other Technological Traits in Durum Wheat under Mediterranean Conditions. Journal of Agricultural and Food Chemistry 49: 3802–3807.

Rotundo, J. L., L. Borrás, M. E. Westgate, and J. H. Orf, 2009 Relationship between assimilate supply per seed during seed filling and soybean seed composition. Field Crops Research 112: 90–96.

Saxton, A. M., 1989 INDCULL Version 3.0: Independent Culling for Two or More Traits. Journal of Heredity 80: 166–167.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. PLOS ONE 7: e30906.

Simmonds, N. W., 1979 *Principles of crop improvement.* Longman., London, K.

Slater, A. T., N. O. I. Cogan, J. W. Forster, B. J. Hayes, and H. D. Daetwyler, 2016 Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. The Plant Genome 9:.

Smedegaard-Petersen, V., and K. Tolstrup, 1985 The Limiting Effect of Disease Resistance on Yield. Annual Review of Phytopathology 23: 475–490.

Smith, H. F., 1936 A Discriminant Function for Plant Selection. Annals of Eugenics 7: 240–250.

Smith, S. P., and R. L. Quaas, 1982 Optimal Truncation Points for Independent Culling-Level Selection Involving Two Traits. Biometrics 38: 975.

Tallis, G. M., 1987 Ancestral covariance and the Bulmer effect. Theoretical and Applied Genetics 73: 815–820.

Triboi, E., P. Martre, C. Girousse, C. Ravel, and A. M. Triboi-Blondel, 2006 Unravelling environmental and genetic relationships between grain yield and nitrogen concentration for wheat. European Journal of Agronomy 25: 108–118.

Venables, W. N., and B. D. Ripley, 2002 *Modern Applied Statistics with S.* Springer, New York.

Woolliams, J. A., P. Berg, B. S. Dagnachew, and T. H. E. Meuwissen, 2015 Genetic contributions and their optimization. Journal of Animal Breeding and Genetics 132: 89–99.

Wu, K. K., W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore *et al.*, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics 83: 294–300.

Xu, S., and W. M. Muir, 1991 Multistage Selection for Genetic Gain by Orthogonal Transformation. Genetics 129: 963–974.

Young, S., 1961a A further examination of the relative efficiency of three methods of selection for genetic gains under less-restricted conditions. Genetical Research Cambridge 2: 106–121.

Young, S., 1961b A further examination of the relative efficiency of three methods of selection for genetic gains under less-restricted conditions. Genetical Research Cambridge 2: 106–121.

Young, S. S. Y., and H. Weiler, 1960 Selection for two correlated traits by independent culling levels. Journal of Genetics 57: 329–338.

2. PLANT BREEDERS SHOULD BE DETERMINING ECONOMIC WEIGHTS FOR A SELECTION INDEX INSTEAD OF USING INDEPENDENT CULLING FOR CHOOSING PARENTS IN BREEDING PROGRAMS WITH GENOMIC SELECTION

ABSTRACT

In the context of genomic selection, we evaluated and compared recurrent selection breeding programs using either index selection or independent culling for selection of parents. We simulated a clonally propagated crop breeding program for 20 cycles of selection using either independent culling or an economic selection index with two unfavourably correlated traits under selection. Cycle time from crossing to selection of parents was kept the same for both strategies. Our results demonstrate that accurate knowledge of the economic importance of traits is essential even when performing independent culling. This is because independent culling achieved its optimum genetic gain when the culling threshold for each trait varied accordingly to the economic importance of the traits. When gains from independent culling were maximised, the efficiency of converting genetic diversity into genetic gain of both selection methods were equivalent. When the same proportion selected of 10% for each trait was used instead of optimal culling levels, index selection was 10%, 128% and 310% more efficient than independent culling when T2 had a relative economic importance of 1.0, 2.5 and 5.0, respectively. Given the complexity of estimating optimal culling levels and the fact that the gains achieved with independent culling are, at most, equivalent to index selection, the use of an economic selection index is recommended for multi-trait genomic selection.

Keywords: economic index; genomic selection; recurrent selection; independent culling

2.1. Introduction

Crop breeding seeks to develop improved cultivars. Besides high yield levels, a successful cultivar in many crops must meet minimal standards for several other traits that are economically important, such as pest and disease resistance and product quality. Traits are often unfavourably correlated with each other (e.g., Kwon and Torrie 1964; Meredith and Bridge 1971; Erskine et al. 1985; Kato and Takeda 1996; Triboi et al. 2006). When traits are antagonistically correlated, selection for one trait causes an undesired economic response in the other trait (Falconer et al. 1996; Bernardo 2010). This makes breeding to simultaneously improve multiple traits complicated.

Independent culling and the use of a selection index are two commonly used methods in plant breeding programs for selecting on multiple traits (Bernardo 2010). Independent culling involves establishing minimum standards (i.e., culling levels) for each trait and only selecting individuals that meet these minimum standards. The thresholds can be set according to a specific selection intensity or a specific value, such as a value relative to an agronomic check. The application of independent culling can be on multiple traits simultaneously or on individual traits sequentially. The selection index method involves selection for all traits simultaneously based on a linear or non-linear combination of individual traits weighted by their importance for the breeding objective (Hazel and Lush 1942).

Theoretically, the selection index is the most efficient method of selection for multiple traits (Hazel and Lush 1942; Young 1961b). However, independent culling can achieve nearly equivalent efficiencies using optimised thresholds (Xu and Muir 1991). Independent culling is less efficient, because when strictly applied it will not select individuals below the threshold for only one trait despite being exceptional for all other traits, while the use of a selection index makes it possible to retain those individuals (Bernardo 2010).

When cost is considered, independent culling can be more efficient than a selection index (Xu and Muir 1991). This is because independent culling does not require phenotypes for all individuals and traits at one time, whereas strict application of a selection index requires phenotypes for all traits. This benefit is particularly valuable to plant breeders, because early stages of the breeding program often have a very large number of individuals. Phenotyping all individuals for all traits is likely to be logistically and financially infeasible. For example, some traits have a high measurement cost, such as bread quality in wheat, so that they cannot be measured on a large number of individuals. Further, some traits can only be measured on older plants, such as lifetime production in sugarcane, or on a plot or group basis. Delaying selection until these traits become available would be effectively equivalent to random selection, because the breeder would have to reduce the overall size of the early stage. Thus, practical constraints require at least some use of independent culling on traits that can be phenotyped simply/quickly and at a lower cost in breeding programs utilising phenotypic selection.

The use of genomic selection in plant breeding may render the cost efficiency benefit of independent culling obsolete if all early generation individuals are genotyped. This is because genomic selection allows for accurate prediction of all traits at once (Meuwissen *et al.* 2001b). While genotyping all early generation individuals is not standard in most current breeding programs, it may become so in the future. This is likely to be the case if breeding programs adopt a two-part strategy to breeding that explicitly splits breeding programs into a rapid cycling, genomic selection guided, population improvement part tasked with developing new germplasm and a product development part focused on developing new varieties. Simulations of these breeding programs suggest they can deliver considerably more genetic gain than more conventional breeding programs (Gaynor et al. 2017).

Several studies have already discussed the benefits of incorporating genomic selection strategies into crop breeding programs (Bernardo and Yu 2007b; Heffner *et al.* 2009b; Gaynor *et al.* 2017; Hickey *et al.* 2017). However, to our knowledge, there have been no studies to date that have investigated the gains over several generations in a recurrent selection breeding program using either a selection index or independent culling, at least in the context of genomic selection. We used simulations of recurrent breeding programs to evaluate and compare both strategies for 20 cycles of selection. The purpose of these simulations was to quantify the magnitude of the difference between optimally set independent culling levels and an optimal selection index. The simulations also investigated the sensitivity of independent culling to sub-optimal culling levels.

2.2. Material and Methods

Stochastic simulations of entire breeding programs for multiple traits were used to compare the genetic gains in a breeding program using independent culling levels and a breeding program using an economic selection index for selection of parents. In the independent culling approach, selection was performed for one trait at a time at each stage of selection. A clonally propagated crop species was considered. In breeding programs for clonally propagated species, all the genotypes in the F_1 population are candidate clones to be released as cultivars or used as parents in the next breeding cycle (Grüneberg *et al.* 2009b). The methods were compared using the average of fifty replicates, each replicate consisting of: i) a burn-in phase shared by both strategies so that each strategy had an identical, realistic starting point; and ii) an evaluation phase that simulated future breeding with different breeding strategies. The burn-in phase consisted of 20 years of breeding using independent culling for the selection of parents and the evaluation phase consisted of 20 cycles of selection using either independent culling or index selection.

Genome sequence

For each replicate, a genome consisting of 10 chromosome pairs was simulated for the hypothetical clonally propagated plant species. These chromosomes were assigned a genetic length of 1.43 Morgans and a physical length of 8x10^8 base pairs. Sequences for each chromosome were generated using the Markovian Coalescent Simulator (Chen et al. 2009) and AlphaSimR (Gaynor et al.). Recombination rate was inferred from genome size (i.e. 1.43 Morgans / 8x10^8 base pairs = 1.8x10^{-9} per base pair), and mutation rate was set to 2x10^{-9} per base pair. Effective population size was set to 50, with linear piecewise increases to 1,000 at 100 generations ago, 6,000 at 1,000 generations ago, 12,000 at 10,000 generations ago, and 32,000 at 100,000 generations ago.

Founder genotypes

Simulated genome sequences were used to produce 50 founder genotypes. These founder genotypes served as the initial parents in the burn-in phase. This was accomplished by randomly sampling gametes from the simulated genome to assign as sequences for the founders. Sites that were segregating in the founders' sequences were randomly selected to serve as 1,000 causal loci per chromosome (10,000 across the genome in total). To simulate genetic correlations between traits, the traits were treated as pleiotropic and the additive effects of the causal loci alleles were sampled from a multivariate normal distribution with mean $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and desired values of correlation.

Estimated breeding values

The true genetic value of the simulated traits was determined by the summing of its causal loci allele effects. The matrix **E** with the estimated breeding values of the traits for each individual in the population was obtained according to the formula:

$$\mathbf{E} = \mathbf{Y}\mathbf{P}^{-1}\mathbf{G}$$

Where **Y** is the matrix of phenotypes simulated by adding random error to the true genetic values of the traits, where rows correspond to individuals in the population and columns correspond to traits. The random error was sampled from a multivariate normal distribution with mean $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and zero covariance, with variance values tuned to achieve a target level of accuracy for both traits. **P** is the phenotypic variance-covariance matrix of the traits, and **G** is the genetic variance-covariance matrix for the traits.

Breeding methods

The simulations modelled breeding for two component traits (T1 and T2) that were improved using either independent culling or an economic selection index. With both strategies, an F_1 population of 5,000 individuals was generated by randomly crossing the individuals in the crossing block (Parents). With independent culling, selection was carried out in two stages: a proportion of individuals was selected first based on T1 and then, from this proportion, the parents of the next breeding cycle were selected based on T2. With the selection index approach, the F_1 individuals with the highest values for the index trait were selected as parents of the next breeding cycle. The index trait was the sum of the estimated breeding values for each trait weighted by their economic importance. The number of selected parents (50 parents) and the cycle time from crossing to selection of new parents was kept the same for both strategies, so the comparisons between them reflect only the differences due to the method of selection. For simulation of breeding programs, we used the R package AlphaSimR (Gaynor et al.).

Simulated scenarios

The selection index and independent culling methods were compared in a set of scenarios that aimed to assess the relative performance of the methods under different levels of accuracy of selection, and relative economic importance of T2. A summary of all simulated scenarios is shown in Table 1.

For one set of scenarios we simulated four levels of accuracy (0.3, 0.5, 0.7, and 0.99), assigned the same economic importance for both traits, and an unfavourable initial genetic correlation of -0.5 between traits. In another set of scenarios, we varied the relative economic importance of T2, but fixed selection accuracy to 0.7 and set an unfavourable initial genetic correlation of -0.5 between traits. Here, three levels of relative economic importance were simulated. T1 was given an economic importance of 1.0 and T2 an economic importance of either 1.0, 2.5 or 5.0. For each level of relative economic importance, we simulated: i) scenarios where the proportion selected was the same (10%) when selecting for both traits, and ii) scenarios where the proportions selected were set to achieve optimal culling levels (i.e., optimal independent culling). To achieve optimal culling levels, in each cycle of selection we chose the proportion selected for each trait that maximised the genetic gain. To find the optimal proportions, we fixed the number of parents selected (50 parents) and found the number of individuals to be selected in the first culling stage that maximized parents' economic value (i.e., index trait).

Comparison

The comparisons were made in terms of: i) genetic gain ii) genetic diversity, iii) the efficiency of converting genetic diversity into genetic gain for the index; and iv) genetic correlation between traits. For genetic gain and genetic diversity, we report values based on the individuals in the crossing block (parents) at each cycle of selection. We measured genetic gain as the increment in genetic mean (average of true genetic values) compared to the genetic mean in year 20. We measured genetic diversity with genetic standard deviation and genic standard deviation. We calculated genetic standard deviation as standard deviation of true genetic values. We calculated genic standard deviation as $\sigma_a = \sqrt{2\sum_{i=1}^{n_q} p_i(1-p_i)\alpha_i^2}$, where n_q is the number of causal loci and p_i and α_i are, respectively, allele frequency and allele substitution effect at the *i*-th causal locus.

To measure efficiency, genetic mean and genic standard deviation were standardized to mean zero and unit standard deviation in year 20. We measured efficiency of converting genetic diversity into genetic gain by regressing the achieved genetic mean ($y_t = \left(\mu_{a_t} - \mu_{a_{20}}\right)/\sigma_{a_{20}}^2$) on lost genetic diversity ($x_t = 1 - \sigma_{a_t}/\sigma_{a_{20}}$), i.e., $y_t = \alpha + bx_t + e_t$, where b is efficiency (Gorjanc et al. 2017). We estimated efficiency with robust regression using function rlm() in R (Venables and Ripley 2002).

For genetic correlation, we report the correlation between the true genetic values of T1 and T2. We calculated this metric on the individuals in the F$_1$ population at each cycle of selection.

2.3. Results

Overall the results show that index selection provided consistent genetic gains and was equivalent to independent culling in terms of genetic gains and efficiency when optimal culling levels were used. Index selection performed better than independent culling in scenarios where independent culling levels were suboptimal.

We have structured the description of the results in two parts, corresponding to how the relative performance of the selection methods was affected by: i) the accuracy of selection, and ii) the relative economic importance of traits.

Accuracy of selection

The results show that increases in accuracy accentuated the differences in the genotypes being selected by either independent culling or index selection. This is shown in Fig. 1, where the genotypes selected as parents by each selection method are highlighted. Lower levels of accuracy led to a more diffuse cluster of selected genotypes and, with increasing selection accuracy, the cluster of selected genotypes approached what was expected for each method of selection (Bernardo, 2010).

Fig. 2 shows the change in the genetic correlation between the component traits for both independent culling and index selection over 20 cycles of selection at different levels of accuracy. Both selection methods resulted in the correlation between traits becoming increasingly unfavourable over the cycles of selection. For both methods, the change in the genetic correlation was higher with higher values of accuracy. Compared to independent culling, index selection led to larger changes in the genetic correlation between the two traits. After 20 cycles of selection with accuracy of 0.3, independent culling led to a genetic correlation that was 9% more unfavourable compared to the genetic correlation in cycle 0, while index selection led to a genetic correlation that was 17% more unfavourable compared to the genetic correlation in cycle 0. After 20 cycles of selection with accuracy of 0.99, independent culling led to a genetic correlation that was 29% more unfavourable compared to the genetic correlation in cycle 0, while index selection led to a genetic correlation that was 64% more unfavourable compared to the genetic correlation in cycle 0.

The change of genetic mean in parents for the component traits and the index trait over the cycles of selection using each method is shown in Fig. 3. For both methods, the genetic gains for the component traits and the index trait increased with higher values of accuracy. In general, the selection index method and independent culling with optimal culling levels led to equivalent genetic gains for the component traits and the index trait. Only in the scenario with 0.99 accuracy did index selection lead to a slightly higher genetic gain compared to that achieved with optimal independent culling. For the index trait, after 20 cycles of selection with accuracy of 0.99, index selection had a genetic gain 4% higher than the genetic gain achieved with independent culling.

Table 2 shows the genetic standard deviation of parents in cycle 20 and the loss in genetic standard deviation in cycle 20 compared to the genetic standard deviation in cycle 0 for the component traits and the index trait. The change of genetic diversity in parents for the component traits and the index trait over the cycles of selection using each method is shown in Supplementary material 1 (Fig S1.1). For the component traits, when using index selection, the genetic standard deviation showed an initial increase in the first few cycles of selection followed by a gradual decrease in the subsequent cycles. When using independent culling, the decrease in the genetic standard deviation of the component traits was continual over the cycles of selection. Both of these trends were more obvious with increasing values of accuracy. For all values of accuracy, independent culling led to a higher loss in the genetic standard deviation of the component traits compared to the index selection. For T1 and T2, independent culling with accuracy of 0.3 led to a loss of genetic standard deviation that was 6% and 5% higher than the loss of genetic standard deviation observed for index selection, respectively. With accuracy of 0.99, for T1 and T2 independent culling led to a loss of genetic standard deviation that was 65% and 51% higher than the loss of genetic standard deviation observed for index selection, respectively. For the index trait, both methods led to equivalent values of genetic standard deviation. With accuracies of 0.3 and 0.99, index selection led to a loss in the genetic standard deviation of the index trait that was 3% higher compared to the loss of genetic standard deviation observed using independent culling, respectively.

Table 3 shows the genic standard deviation of parents in cycle 20 and the loss in genic standard deviation in cycle 20 compared to the genic standard deviation in cycle 0 for the component traits and the index trait. The values of genic standard deviation of T1, T2, and the index trait were equivalent. The highest difference between methods in the loss in genic standard deviation was 1% for all values of accuracy, except with accuracy of 0.99. With 0.99 accuracy, for T1, T2 and the index trait, index selection led to a loss in the genic standard deviation that was 3% higher compared to the loss of genic standard deviation observed using independent culling.

Relative economic importance of traits

Fig. 4 shows the efficiency of converting genetic diversity into genetic gain for the index trait when the relative economic importance of T2 varies. Independent culling was compared to index selection using either optimal culling levels or selection with the same proportion of plants selected (10%) for each trait. Index selection had the highest efficiency and most gain for all levels of economic importance. The efficiency and gain for optimal independent culling levels was nearly equivalent to index selection. The efficiency and gain for selecting the same proportion of plants for both traits was worse than index selection for all levels of relative economic importance. Index selection was 10%, 128% and 310% more efficient than independent culling using the same proportion of selected plants for relative economic importance of 1.0, 2.5 and 5.0, respectively.

Fig. 4 also shows the proportion of plant selected for T1 under optimal independent culling over the different levels of economic importance for T2. The mean proportion selected for T1 only varied slightly over the cycles of selection. The means were 29%, 93%, and 99% for relative economic importance of 1.0, 2.5, and 5.0, respectively. The variation about those means was largest with relative economic importance of 1.0 and smallest with relative economic importance of 5.0.

2.4. Discussion

This study evaluated and compared recurrent selection breeding programs that either use index selection or independent culling for the selection of parents by genomic selection. Overall the results show that using index selection is either better or equivalent to independent culling in this context. Index selection outperformed independent culling when sub-optimal culling levels were used. Our results demonstrate that accurately assessing the economic importance of the traits is essential regardless of the method of selection being used.

The main difference between index selection and independent culling is that, when using index selection, genotypes that are exceptional for one of the traits under selection are more likely to be selected even though their performance for other traits is average. This can be seen in Fig. 1, with the cluster of individuals selected as parents with the index method including individuals that are more contrasting for the two traits under selection compared to the individuals selected with independent culling. The main implications of this are in the way each method affects the correlation between traits and the genetic diversity over cycles of recurrent selection. We discuss each of these aspects in the following two sections. In the third section, we discuss how the relative economic importance of the traits can affect the relative performance of the methods. Lastly, we discuss the implications of our results for modern plant breeding programs which deploy genomic selection.

Methods of selection and genetic correlation between traits

The results show that, after only a few cycles of selection, index selection generates F_1 populations with a more unfavourable genetic correlation between traits than the F_1 populations generated by independent culling (Fig 2). An explanation for the faster decrease of the genetic correlation observed with index selection is that the index is a linear combination of component traits. As shown by Bulmer (1971), selection on a linear combination leads to negative covariances between components (i.e., Bulmer effect). Consequently, the same principle applies to the component traits and index selection, with index selection leading to an unfavourable genetic correlation between the component traits (Tallis 1987; Itoh 1991).

In general, genetic gains in multi-trait selection, regardless of the method of selection, are expected to be higher when the correlation between traits is favourable and lower when this correlation is unfavourable (Young 1961b). As index selection generated F_1 populations with more unfavourable genetic correlation between traits than independent culling, the genetic gains for index selection were potentially lower than for independent culling. Nevertheless, despite index selection being carried out under increasingly unfavourable genetic correlations over the cycles, the genetic gains obtained for the index trait were equivalent to the gains obtained using independent culling (Fig. 3).

Unfavourable genetic correlations are the most challenging scenario for breeders. When traits are unfavourably correlated, selection on one trait results in response in an undesired direction for the other trait. When these correlations are due to pleiotropy, they cannot be broken with repeated cycles of recombination. This case is likely pervasive in several crops, e.g., grain yield and protein content in cereal crops (Duvick and Cassman 1999; Rharrabti et al. 2001; Rotundo et al. 2009), quality and disease resistance in forage crops (Casler and Vogel 1999), and yield and disease resistance in barley (Smedegaard-Petersen and Tolstrup 1985). However, the extent of genetic correlation and pleiotropy in these examples is unknown because unfavourable genetic correlations between the traits could also be, at least partly, induced by selection, as demonstrated in this study.

Methods of selection and genetic diversity over cycles of selection

According to Bulmer (1971), reduction in the genetic variance due to selection stems mostly from the build-up of negative linkage disequilibrium between causal loci when selection is performed. This can be seen by comparing genetic and genic variation (Table 2 and Table 3, respectively). Genic variation is a function of the allele frequencies and the allele substitution effect only, and thus is not affected by changes in linkage disequilibrium. The results in Table 3 show that the loss of genic standard deviation of the component traits and index trait are not greatly affected by the method of selection. Also, the method of selection did not greatly affect the trait means, as shown in Fig. 3. This indicates that, in terms of allele frequencies, there was little difference in the parents selected by either independent culling or the selection index method in situations similar to our simulation. Therefore, the difference between the selection methods derives from how they induce and exploit linkage disequilibrium between the causal variants of the component traits. Specifically, as shown in Table 2, independent culling induced a greater degree of negative linkage disequilibrium between the causal variants of the component traits resulting in those traits having less genetic variation. A deviation from this result is expected with more intense selection schemes and more component traits selected in successive stages, which would induce larger changes in allele frequencies due to drift. As a consequence, differences between index selection and independent culling would be accentuated. Cowling and Li (2018) simulated and compared wheat breeding programs using different selection strategies under high and low selection intensities. They observed index selection resulted in higher population coancestry over cycles of selection compared to independent culling, and the difference between methods increased in scenarios with high selection intensity. Their results indicate index selection leads to a higher loss of genic standard deviation.

Somewhat surprisingly, it is possible to make an argument for the superiority of independent culling relative to a selection index on the basis of the differences observed in linkage disequilibrium. This is because independent culling produced populations with nearly equivalent mean performance, but with more consistent performance between individuals, which is demonstrated by the lower variation observed for the component traits. This property could be beneficial from a management perspective if differences in the component traits require variations in management of individuals. Breeding for plant-architecture traits in outbreeding cultivars is a good example where this property might be valuable, as having more uniform plants in the field favours mechanical harvest. However, we believe this property is more of an academic curiosity than something that will have practical application.

For simplicity and ease of implementation, our simulations consider the same genetic architecture for both traits, with both traits being controlled by a high number (10,000) of causal loci with small additive effects. Under different circumstances, such as at least one of the traits being controlled by few causal loci with higher allele substitution effects, different results could be expected. The results for the two-locus model of Bennett and Swiger (1980) show that independent culling tends to eliminate genotypes that are homozygous for alleles with low effect for one of the traits. For one pleiotropic causal locus, when both alleles are favourable for one trait and unfavourable for the other trait, both homozygous genotypes tend to be culled, and independent culling would select the heterozygous genotypes. If heterozygous genotypes were preferred, the fixation of alleles would be slower and, therefore, the loss in genic standard deviation would be lower. Our results indicate that, for highly polygenic traits, differences between methods of selection in the loss of genetic diversity are mostly due to changes in linkage disequilibrium as opposed to distinctive changes in allele frequencies. Therefore, in terms of conserving genetic diversity there was no obvious advantage for either method. Other strategies such as optimal-cross selection should be considered in order to optimize gains while also controlling the loss of genetic diversity over cycles of selection (Clark et al. 2013; Woolliams et al. 2015; Gorjanc et al. 2017; Cowling and Li 2018).

Economic importance of the traits

In general, when using the same selection intensity for both traits, the greater the difference in the economic importance of the traits, the better index selection will perform compared to independent culling (Fig. 4). This happens because there is a combination of selection intensities for each trait that maximizes the genetic gain when performing independent culling (Hazel and Lush 1942). Finding these selection intensities when selecting for two traits in two stages of selection is complex (Young and Weiler 1960; Namkoong 1970; Cotterill and James 1981; Smith and Quaas 1982), and becomes even more complex with increasing number of traits and stages of selection (Saxton 1989; Ducrocq and Colleau 1989; Xu and Muir 1991).

The results in Fig. 4 show that independent culling approaches its maximal gain when a higher selection intensity is used for the trait with higher economic importance and a lower selection intensity is used for the trait with lower economic importance. In fact, when one trait had 5 times the economic importance of the other trait, the optimum was achieved when almost no selection was carried out for the less important trait. These results demonstrate that accurately assessing the economic importance of the traits is essential even when independent culling is performed.

Regardless of the gains achieved with independent culling being maximised, when parents are selected based on an index, equivalent gains are achieved by simply summing the values of the traits weighted by their economic importance. Once the true economic weights of the traits are quantified, index selection is much simpler than independent culling when using these weights for optimizing the genetic gains in a plant breeding program.

Index selection in modern plant breeding programs that use genomic selection

There is little to no evidence suggesting plant breeders use analytical techniques to determine optimal independent culling thresholds and/or constructing selection indices in most plant breeding programs. More likely, the majority of breeders rely on their intuition for setting thresholds and constructing indices. Their decisions are likely guided by the performance of agronomic checks and are prone to fluctuations between seasons and individual breeders. This model has clearly been successful, because plant breeding programs have continued to deliver genetic gain. However, it is likely sub-optimal, and a more analytical approach should be adopted in the future.

The value of a more analytical approach becomes greater as genomic selection is more widely used. The results presented in this paper show a selection index is superior to independent culling when using genomic selection. These results are further supported by earlier theoretical work (Smith 1936; Hazel and Lush 1942; Young 1961b). This indicates a clear preference for implementing selection indices in plant breeding.

The focus of plant breeders should be determining the economic weights for a selection index. In this paper the economic model used to select weights was implicitly assumed to be known and linear. The reality is that true economic model may be unknown to breeders and it is likely non-linear. The presence of a non-linear model does not pose a problem, because linear economic weights can be derived for improving the economic value of germplasm (Goddard 1983). However, this still requires defining the economic model. For this reason, it is our opinion that plant breeders would benefit greatly from an increased emphasis on understanding and quantifying the economics of their species. This information would greatly aid breeders in getting the most out of genomic selection.

2.5. Conclusions

We evaluated and compared recurrent selection breeding programs using either independent culling or index selection for parent selection. The results show that, despite selection being carried out under unfavourable genetic correlations when using the selection index instead of independent culling, equivalent or higher genetic gains were achieved with index selection in all simulated scenarios. In terms of genetic diversity, the differences between methods in the studied system were driven mostly by differences in the generation of linkage disequilibrium between causal loci induced and not differences in allele frequencies. When linkage disequilibrium was not considered, both methods were equivalent in terms of loss of genetic diversity, and the differences between methods in terms of efficiency of converting genetic diversity into genetic gains mostly reflected the differences in the genetic gains obtained with each method. To obtain higher genetic gains, accurately assessing the economic importance of the traits is essential even when independent culling is performed, as optimal culling levels should be determined in order for maximum gain to be achieved. Given that optimal culling levels are complex to estimate, once the economic importance of each trait is known, maximum genetic gains are more easily achieved with index selection. Therefore, the best choice for plant breeding programs is to select parents using an economic selection index.

2.6. Acknowledgements

This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Computational Biology Programme, Grant No. BEX 0043/17-6). The Roslin authors acknowledge the financial support from BBSRC and KWS UK, RAGT Seeds Ltd., Elsoms Wheat Ltd and Limagrain UK for the project "GplusE: Genomic selection and Environment modelling for next generation wheat breeding" (grants BB/L022141/1 and BB/L020467/1).

REFERENCES

Bennett GL, Swiger LA (1980) Genetic variance and correlation after selection for two traits by index, independent culling levels and extreme selection. Genetics 94:763–775

Bernardo R (2010) Breeding for quantitative traits in plants, 1st edn. Stemma Press, Woodbury

Bernardo R, Yu J (2007) Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Science 47:1082–1090. doi: 10.2135/cropsci2006.11.0690

Bulmer MG (1971) The effect of selection on genetic variability. 105:1–88

Casler MD, Vogel KP (1999) Accomplishments and Impact from Breeding for Increased Forage Nutritional Value. Crop Science 39:12–20. doi: 10.2135/cropsci1999.0011183X003900010003x

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome research 19:136–142. doi: 10.1101/gr.083634.108.1

Clark SA, Kinghorn BP, Hickey JM, Van Der Werf JHJ (2013) The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics Selection Evolution 45:1–8. doi: 10.1186/1297-9686-45-44

Cotterill PP, James JW (1981) Optimising two-stage independent culling selection in tree and animal breeding. Theoretical and Applied Genetics 59:. doi: 10.1007/BF00285891

Cowling W, Li L (2018) Turning the heat up on independent culling in crop breeding

Ducrocq V, Colleau JJ (1989) Optimum truncation points for independent culling level selection on a multivariate normal distribution, with an application to dairy cattle selection. Genetics Selection Evolution 21:185. doi: 10.1186/1297-9686-21-2-185

Duvick DN, Cassman KG (1999) Post–Green Revolution Trends in Yield Potential of Temperate Maize in the North-Central United States. Crop Science 39:1622–1630. doi: 10.2135/cropsci1999.3961622x

Erskine W, Williams PC, Nakkoul H (1985) Genetic and environmental variation in seed yield, seed size and cooking quality of lentil. Field Crops Research 12:153–161

Falconer DS, Mackay TF, Frankham R (1996) Introduction to Quantitative Genetics (4th edn). Trends in Genetics 12:280

Gaynor RC, Gorjanc G, Bentley AR, et al (2017) A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. Crop Science 57:2372–2386. doi: 10.2135/cropsci2016.09.0742

Gaynor RC, Gorjanc G, Wilson DL AlphaSimR: An R Package for Breeding Program Simulations. Manuscr Prep

Goddard ME (1983) Selection indices for non-linear profit functions. Theoret Appl Genetics 64:339–344. doi: 10.1007/BF00274177

Gorjanc G, Gaynor RC, Hickey JM (2017) Optimal cross selection for long-term genetic gain in two- part programs with rapid recurrent genomic selection. bioRxiv. doi: 10.1101/227215

Grüneberg W, Mwanga R, Andrade M, Espinoza J (2009) Selection methods. Part 5: Breeding clonally propagated crops. Plant Breeding and Farmer Participation 275–322

Hazel LN, Lush JL (1942) The Efficiency of Three Methods of Selection. Journal of Heredity 33:393–399. doi: 10.1093/oxfordjournals.jhered.a105102

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic Selection for Crop Improvement. Crop Science 49:1–12. doi: 10.2135/cropsci2008.08.0512

Hickey JM, Chiurugwi T, Mackay I, et al (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics 49:1297

Itoh Y (1991) Changes in genetic correlations by index selection. Genetics selection evolution GSE 23:301–308. doi: 10.1186/1297-9686-23-4-301

Kato T, Takeda K (1996) Associations among Characters Related to Yield Sink Capacity in Space-Planted Rice. Crop Science 36:1135–1139. doi: 10.2135/cropsci1996.0011183X003600050011x

Kwon SH, Torrie JH (1964) Heritability of and interrekationships among traits of two soybean population

Meredith WR, Bridge RR (1971) Breakup of Linkage Blocks in Cotton, Gossypium hirsutum L.1. Crop Science 11:695–698. doi: 10.2135/cropsci1971.0011183X001100050027x

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. doi: 11290733

Namkoong G (1970) Optimum Allocation of Selection Intensity in Two Stages of Truncation Selection. Biometrics 26:465. doi: 10.2307/2529102

Rharrabti Y, Elhani S, Martos-Núñez V, García del Moral LF (2001) Protein and Lysine Content, Grain Yield, and Other Technological Traits in Durum Wheat under Mediterranean Conditions. Journal of Agricultural and Food Chemistry 49:3802–3807. doi: 10.1021/jf001139w

Rotundo JL, Borrás L, Westgate ME, Orf JH (2009) Relationship between assimilate supply per seed during seed filling and soybean seed composition. Field Crops Research 112:90–96. doi: https://doi.org/10.1016/j.fcr.2009.02.004

Saxton AM (1989) INDCULL Version 3.0: Independent Culling for Two or More Traits. Journal of Heredity 80:166–167. doi: 10.1093/oxfordjournals.jhered.a110823

Smedegaard-Petersen V, Tolstrup K (1985) The Limiting Effect of Disease Resistance on Yield. Annual Review of Phytopathology 23:475–490. doi: 10.1146/annurev.py.23.090185.002355

Smith HF (1936) A Discriminant Function for Plant Selection. Annals of Eugenics 7:240–250. doi: 10.1111/j.1469-1809.1936.tb02143.x

Smith SP, Quaas RL (1982) Optimal Truncation Points for Independent Culling-Level Selection Involving Two Traits. Biometrics 38:975. doi: 10.2307/2529877

Tallis GM (1987) Ancestral covariance and the Bulmer effect. Theoretical and Applied Genetics 73:815–820. doi: 10.1007/BF00289384

Triboi E, Martre P, Girousse C, et al (2006) Unravelling environmental and genetic relationships between grain yield and nitrogen concentration for wheat. European Journal of Agronomy 25:108–118. doi: 10.1016/j.eja.2006.04.004

Venables WN, Ripley BD (2002) Modern Applied Statistics with S., Fourth Edi. Springer, New York

Woolliams JA, Berg P, Dagnachew BS, Meuwissen THE (2015) Genetic contributions and their optimization. Journal of Animal Breeding and Genetics 132:89–99. doi: 10.1111/jbg.12148

Xu S, Muir WM (1991) Multistage Selection for Genetic Gain by Orthogonal Transformation. Genetics 129:963–974

Young S (1961) A further examination of the relative efficiency of three methods of selection for genetic gains under less-restricted conditions. Genetical Research Cambridge 2:106–121. doi: 10.1017/S0016672300000598

Young SSY, Weiler H (1960) Selection for two correlated traits by independent culling levels. Journal of Genetics 57:329–338. doi: 10.1007/BF02987238

TABLES AND FIGURES

Table 1. Summary of parameters simulated in all comparison scenarios of recurrent selection breeding programs using either independent culling or selection index with two traits

Scenario	Selected Proportion		Genetic correlation	Relative economic importance of Trait 2	Accuracy
	Trait 1	Trait 2			
1	Optimum	Optimum	-0.5	1.0	0.3
2	Optimum	Optimum	-0.5	1.0	0.5
3	Optimum	Optimum	-0.5	1.0	0.9
4	Optimum	Optimum	-0.5	1.0	0.7
5	Optimum	Optimum	-0.5	2.5	0.7
6	Optimum	Optimum	-0.5	5.0	0.7
7	10%	10%	-0.5	1.0	0.7
8	10%	10%	-0.5	2.5	0.7
9	10%	10%	-0.5	5.0	0.7

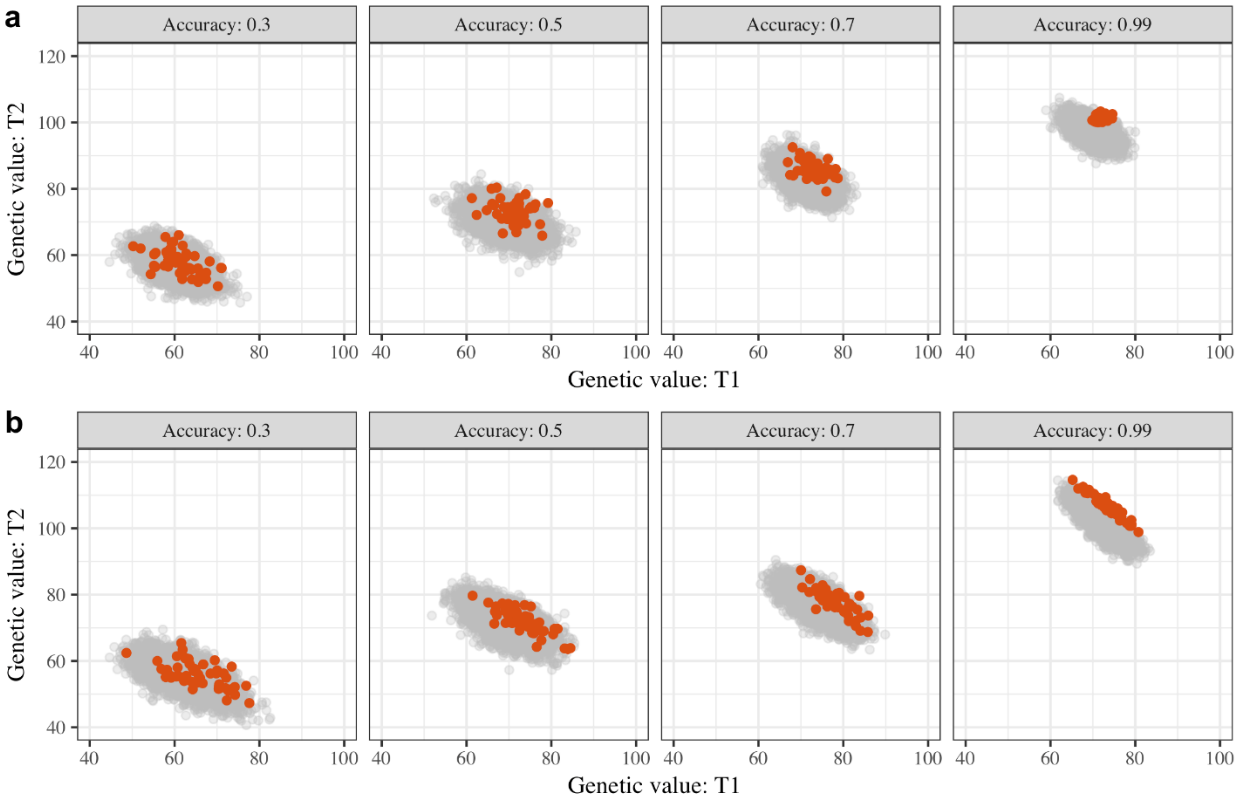

Fig 2. Scatterplots of true genetic values for Trait 1 (T1) and Trait 2 (T2) of the genotypes in the F1 population (grey) and genotypes selected as parents (orange) in the third cycle of selection using either independent culling (**a**) or a selection index (**b**) with different levels of accuracy

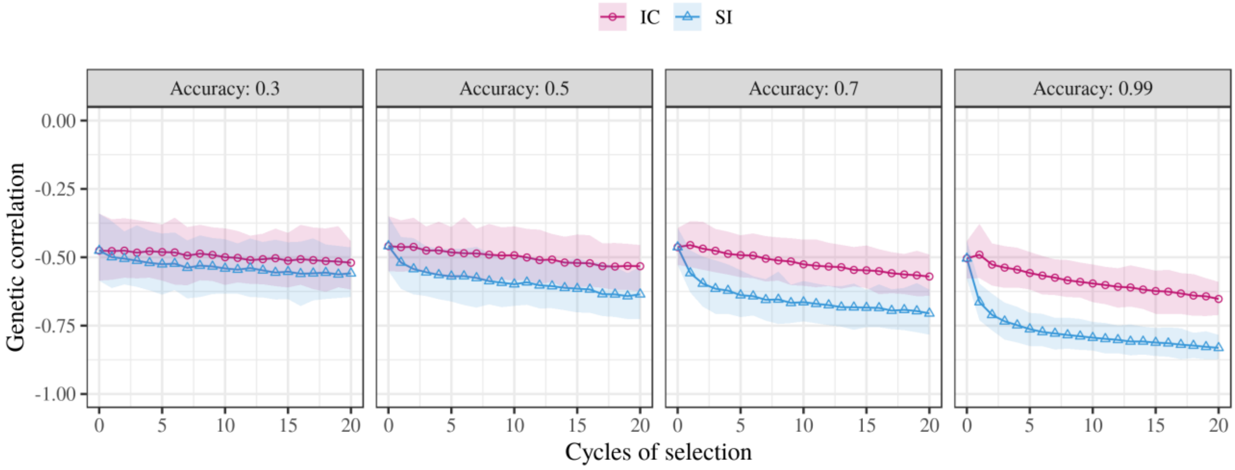

Fig 2. Change in genetic correlation (mean and 95% confidence interval) between traits in the F1 population over 20 cycles of selection using either optimal independent culling (IC) or a selection index (SI) with different levels of accuracy, and Trait 2 relative economic importance of 1.0

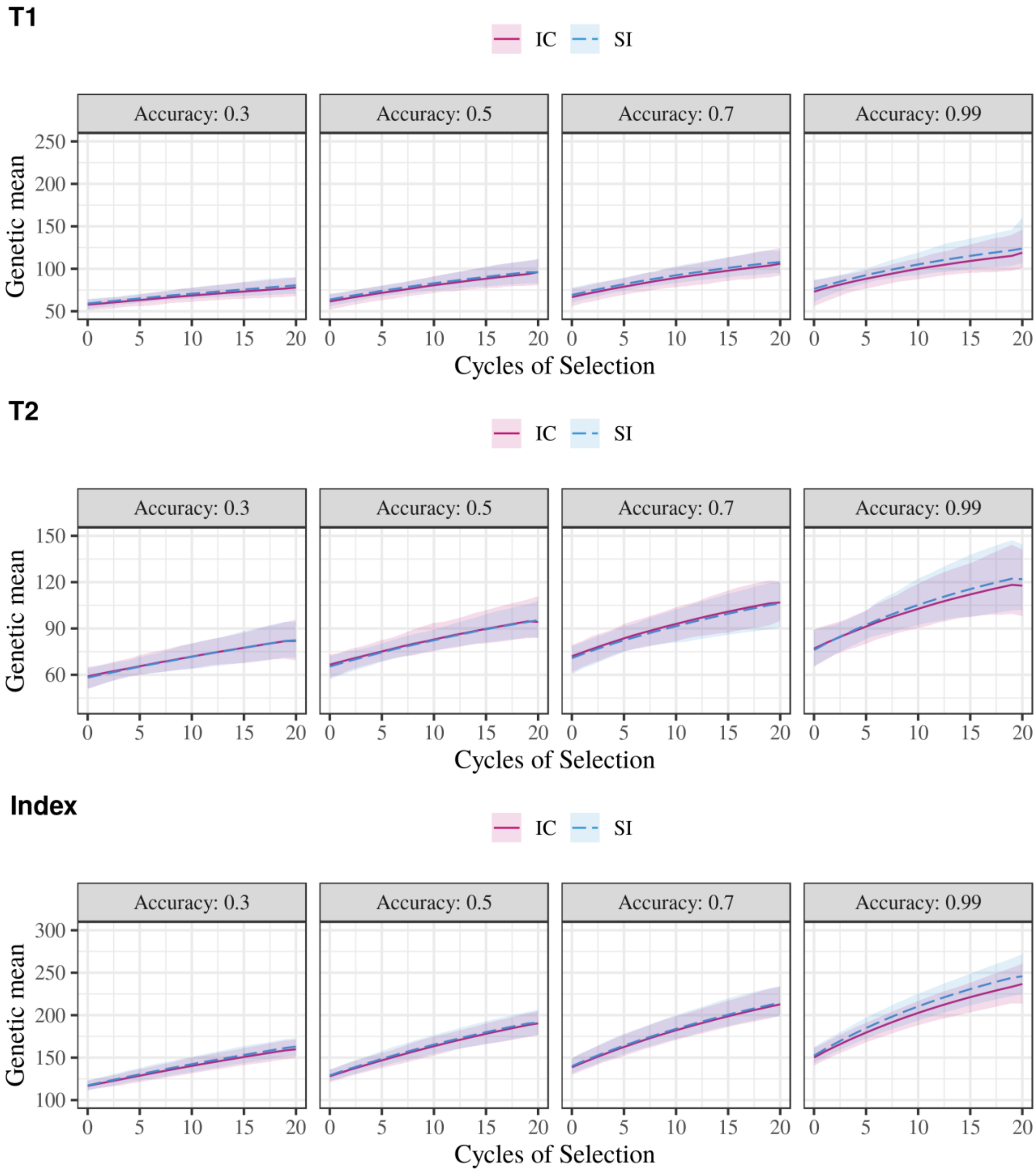

Fig 3. Change in genetic mean for Trait 1 (T1), Trait 2 (T2) and Index Trait (Index) over 20 cycles of selection using either optimal independent culling (IC) or a selection index (SI) with different levels of accuracy, unfavourably correlated traits, and T2 relative economic importance of 1.0

Table 2. Mean genetic standard deviation (Genetic SD) of parents in cycle 20 and loss in genetic standard deviation in cycle 20 in comparison to the genetic standard deviation in cycle 0 (Loss over cycle 0) for trait 1 (T1), trait 2 (T2) and the index trait using either optimal independent culling or index selection with different levels of accuracy, unfavourably correlated traits, and T2 relative economic importance of 1.0

	Independent culling					
	T1		**T2**		**Index trait**	
Accuracy	**Genetic SD (cycle 20)**	**Loss over cycle 0**	**Genetic SD (cycle 20)**	**Loss over cycle 0**	**Genetic SD (cycle 20)**	**Loss over cycle 0**
0.3	3.51 (0.08)*	-17%	3.68 (0.08)	-16%	3.57 (0.06)	-22%
0.5	2.56 (0.06)	-30%	2.45 (0.04)	-28%	2.69 (0.05)	-32%
0.7	1.65 (0.04)	-42%	1.64 (0.03)	-37%	1.88 (0.04)	-45%
0.99	0.45 (0.01)	-68%	0.45 (0.01)	-55%	0.74 (0.02)	-62%
	Index Selection					
	T1		**T2**		**Index trait**	
Accuracy	**Genetic SD (cycle 20)**	**Loss over cycle 0**	**Genetic SD (cycle 20)**	**Loss over cycle 0**	**Genetic SD (cycle 20)**	**Loss over cycle 0**
0.3	3.80 (0.09)	-11%	4.00 (0.09)	-11%	3.66 (0.08)	-19%
0.5	3.19 (0.08)	-17%	3.19 (0.07)	-14%	2.57 (0.06)	-33%
0.7	2.69 (0.06)	-16%	2.60 (0.06)	-18%	1.86 (0.04)	-41%
0.99	1.93 (0.4)	-3%	1.91 (0.04)	-4%	0.51 (0.01)	-59%

* standard errors of the estimates are presented in parenthesis

Table 3. Genic standard deviation (Genic SD) of parents in cycle 20 and loss in genic standard deviation in cycle 20 in comparison to the genic standard deviation in cycle 0 (Loss over cycle 0) for trait 1 (T1), trait 2 (T2) and the index trait using either optimal independent culling or index selection with different levels of accuracy, unfavourably correlated traits, and T2 relative economic importance of 1.0

	Independent culling					
	T1		**T2**		**Index trait**	
Accuracy	**Genic SD (cycle 20)**	**Loss over cycle 0**	**Genic SD (cycle 20)**	**Loss over cycle 0**	**Genic SD (cycle 20)**	**Loss over cycle 0**
0.3	3.94 (0.06)*	-15%	4.11 (0.07)	-15%	4.04 (0.05)	-16%
0.5	3.48 (0.06)	-24%	3.41 (0.05)	-24%	3.44 (0.04)	-25%
0.7	2.94 (0.04)	-34%	2.89 (0.04)	-34%	2.89 (0.04)	-34%
0.99	2.35 (0.04)	-42%	2.35 (0.04)	-42%	2.33 (0.04)	-43%
	Index Selection					
	T1		**T2**		**Index trait**	
Accuracy	**Genic SD (cycle 20)**	**Loss over cycle 0**	**Genic SD (cycle 20)**	**Loss over cycle 0**	**Genic SD (cycle 20)**	**Loss over cycle 0**
0.3	3.92 (0.06)	-16%	4.08 (0.07)	-16%	4.02 (0.05)	-16%
0.5	3.44 (0.06)	-25%	3.37 (0.05)	-25%	3.39 (0.05)	-26%
0.7	2.92 (0.05)	-34%	2.88 (0.05)	-34%	2.87 (0.04)	-35%
0.99	2.21 (0.04)	-45%	2.22 (0.04)	-45%	2.17 (0.03)	-46%

* standard errors of the estimates are presented in parenthesis

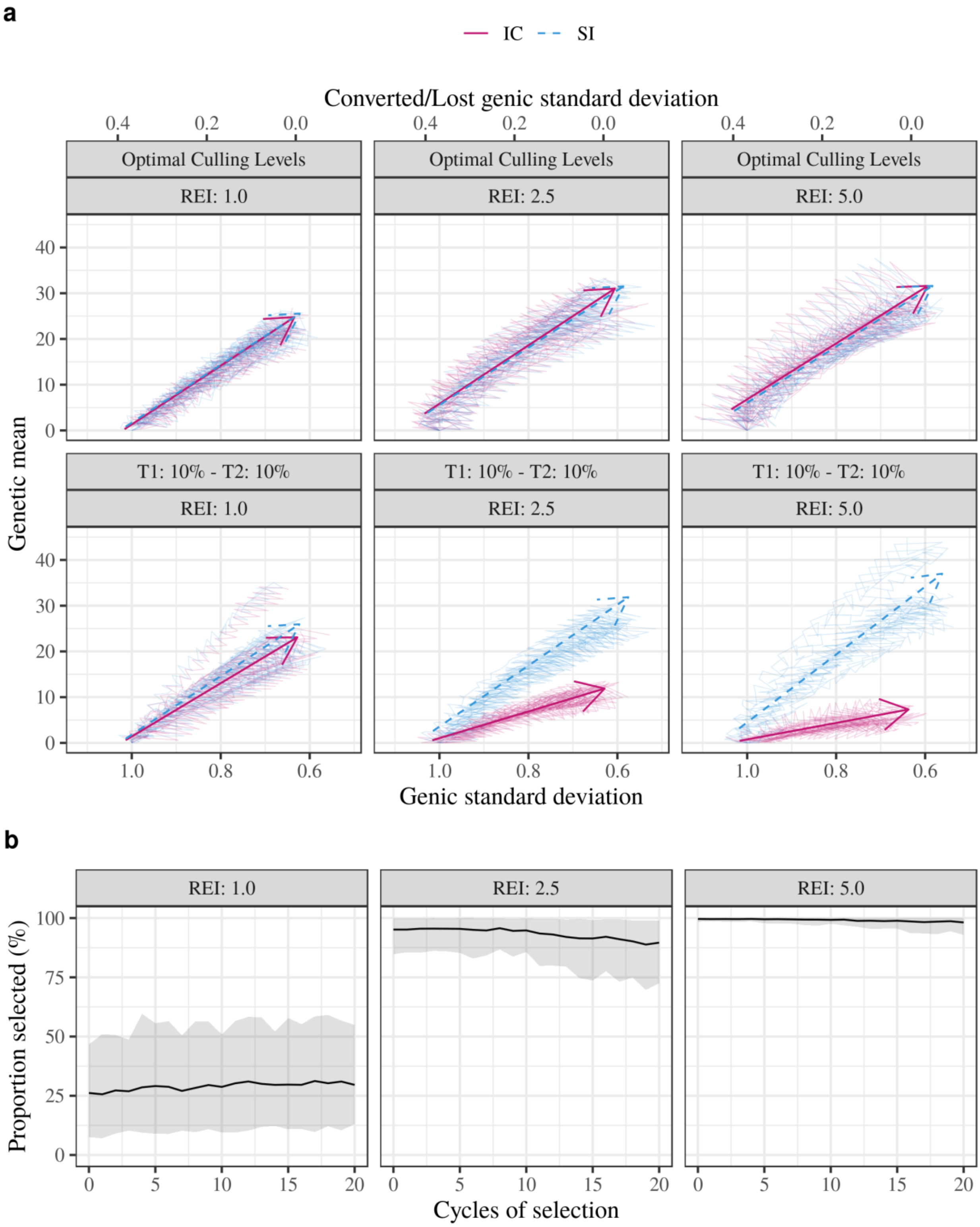

Fig 4. Change of genetic mean and genic standard deviation for the index trait across 20 cycles of selection using either independent culling (IC) or a selection index (SI) under three levels of relative economic importance (REI) and using either the same proportion selected (10%) for Trait 1 (T1) and Trait 2 (T2) or optimal culling levels for each level of relative economic importance of T2 (a); and proportion selected (mean and 95% confidence interval) for T1 used to achieve optimal culling levels over the 20 cycles of selection (b). Traits are unfavourably correlated (-0.5). Individual replicates are shown by thin lines and a mean regression with a time-trend arrow. Values of genetic mean and genic standard deviation shown are standardized to mean zero and unit standard deviation in cycle 0

SUPPLEMENTARY MATERIAL

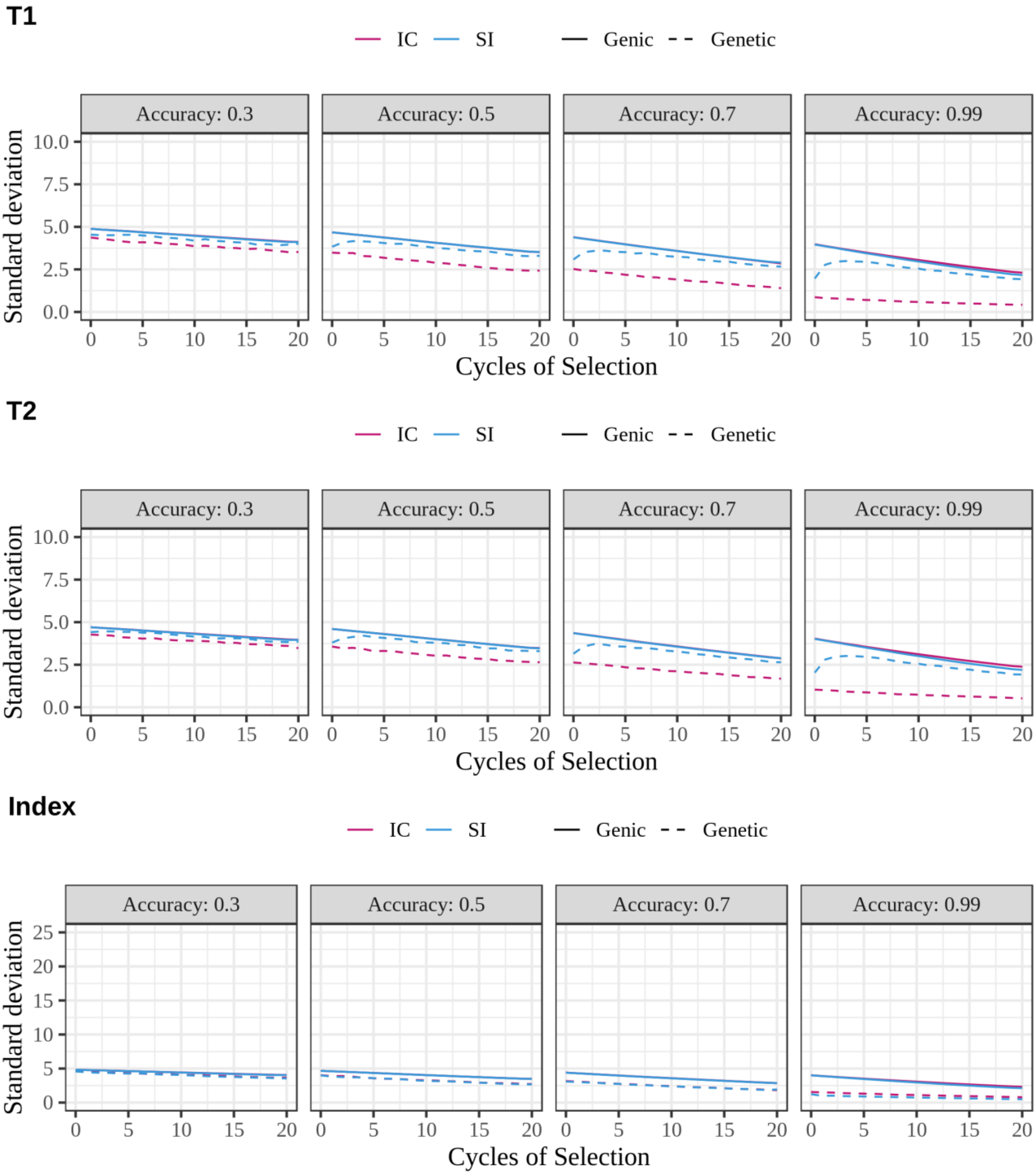

Figure S1. Change in genic and genetic standard deviation for Trait 1 (T1), Trait 2 (T2) and Index Trait (Index) over 20 cycles of selection using either independent culling (IC) or a selection index (SI) with different levels of accuracy, proportion selected of 10%, unfavourably correlated traits, and T2 relative economic importance of 1.0

3. GENOMIC SELECTION WITH ALLELE DOSAGE INFORMATION IN SUGARCANE

ABSTRACT

Modern sugarcane cultivars are derived from interspecific hybrids between *Saccharum officinarum* and *S. spontaneum*, backcrossed with S. *officinarum*. The result is a highly polyploid genome with a varying number of chromosomes in different homology groups. Hence, genomic selection models for sugarcane have to account for different ploidy levels between markers and multiple possible allele dosages. We expanded the methodology used for genomic selection in autotetraploids in order to build covariance matrices of both additive and digenic dominance effects that were subsequently used in GBLUP models. We applied these models using estimates of ploidy and allele dosage of 6,550 polymorphisms obtained through genotyping-by-sequencing of 179 sugarcane genotypes from a biparental F1 progeny. Phenotypes for soluble solids content, sucrose content, fiber percentage, stalk diameter, stalk weight and stalk height were evaluated in two sites, during 2008 and 2009 for the first site and 2012 through 2014 for the second site. We observed low genetic variability and low values of heritability in the progeny, from 0.31 to 0.55. Mean predictive abilities ranged from 0.12 to 0.28 (additive) and from 0.05 to 0.29 (additive + dominant) and did not differ considerably from the mean predictive abilities when allele dosage information was not included in the model. Selection coincidence was higher when allele dosage was included, with a 20% difference from that obtained using diploidized markers. Overall, including estimates of ploidy and allele dosage in the models led to a modest improvement of genomic selection in sugarcane. The improvement is likely to be more evident with training population sets with a higher number of individuals and higher genetic variability.

Keywords: Sugarcane; Genomic selection; Polyploids; Allele dosage; Digenic dominance

3.1. Introduction

Sugarcane (*Saccharum* spp.) accounts for 80% of the sugar production in the world (CIRAD) and has potential to become the main crop for bioenergy production, with the highest mean ethanol yield per hectare and a relatively low production cost (Goldemberg and Guardabassi 2010). With increasing worldwide demand for renewable energy sources, obtaining more productive and robust sugarcane cultivars becomes critical. The main bottleneck in sugarcane breeding programs is the rather long process for selection of cultivars. A traditional sugarcane breeding program is usually divided in several phases of selection, each consisting of large experiments that are usually conducted for more than one crop cycle (Cheavegatti-Gianotto *et al.* 2011; Zhou 2013), taking up to 12 years from the initial crosses until commercial cultivar release (Park *et al.* 2007). In this framework, there exists a pressing need for the development of new strategies that will allow the reduction of experimental costs and time for selection of sugarcane cultivars.

A viable way of achieving improvement in breeding programs in terms of time and costs is through the deployment of genomic selection (Heffner *et al.* 2009). Genomic selection consists of using a representative population that is both genotyped and phenotyped (i.e., the training population) to predict the effect of genetic markers widely spread throughout the genome. The predicted effects are then used to predict the breeding or genotypic value of genotyped individuals (Meuwissen *et al.* 2001). This allows selection to be carried based on predicted breeding values, reducing the need for further costly phenotypic evaluations and shortening the time needed for selection of the best genotypes. Genomic selection has been successfully implemented in several crop breeding programs (Bernardo and Yu 2007; Heffner *et al.* 2009; Crossa *et al.* 2010; Resende *et al.* 2012; Duhnen *et al.* 2017). Although genomic selection could greatly improve sugarcane breeding programs, its implementation demands a relatively large set of genetic markers to be consistently obtained at feasible costs, a process which is severely impaired by the complex genomes observed in the genus *Saccharum*.

Modern sugarcane cultivars are derived from interspecific hybridizations between two highly polyploid species, *S. officinarum* (2n = 80, x = 10) and *S. spontaneum* (2n = 40 to 128, x = 8) (D'Hont *et al.* 1996, 1998). The interspecific hybrids were then successively backcrossed with *S. officinarum*, the so-called noble cane. This culminated in a highly heterozygous, aneuploid genome with 100 to 130 chromosomes, most of which are derived from *S. officinarum*, 10% to 20% from *S. spontaneum*, and approximately 10% from interspecific recombinants (Grivet and Arruda 2002). These biological events resulted in hom(e)ology groups with different ploidy levels and a higher number of heterozygous genotypic classes, which makes estimating genotypic classes a substantially more difficult task (Mollinari and Serang 2015).

To avoid these problems, the majority of genetic studies in sugarcane use either dominant or single-dosage codominant markers (Wu *et al.* 1992; Huckett and Botha 1995; Besse *et al.* 1998; Nair *et al.* 2002; Gouy *et al.* 2013; Aitken *et al.* 2014; Racedo *et al.* 2016; Balsalobre *et al.* 2017), i.e., polymorphisms that were either detected in a presence/absence fashion or that could only be detected in one chromosome per homology group. When using only dominant or single-dosage markers, markers with other allele dosage levels are ignored. However, new tools that leverage the relative allelic abundance of single nucleotide polymorphisms (SNPs) to estimate both their ploidy and allele dosage have allowed markers with higher dosages to be used in sugarcane studies (Garcia *et al.* 2013; Mollinari and Serang 2015).

Garcia *et al.* (2013) showed that the portion of the sugarcane genome effectively explored by single-dosage (simplex) markers can be rather small, indicating that using markers with higher dosages would potentially result in greater coverage and better representation of these polyploid genomes. Moreover, given that recent studies have shown that allele dosage information can improve the accuracy of genomic selection models in autotetraploid species (Slater *et al.* 2016, 2016; de Bem Oliveira *et al.* 2018; Hawkins and Yu 2018; Endelman *et al.* 2018), the objective of our study was to assess the accuracy of genome-wide prediction when incorporating allele dosage information in highly polyploid sugarcane.

3.2. Material and Methods

Genetic material and field experiments

A segregating F_1 progeny of 179 individuals was derived from the crossing of two commercial cultivars, IACSP95-3018 (female) and IACSP93-3046 (male). IACSP95-3018 is a promising clone used as a parent in the breeding program at IAC (Instituto Agronômico de Campinas), and IACSP93-3046 has a high level of sucrose, good tillering and an erect stool habit, being recommended for mechanical harvest.

The first field experiment was set in Sales de Oliveira, SP, Brazil, in 2007. A randomized complete block design with four replicates was used and evaluations were carried in the harvest years of 2008 (plant cane) and 2009 (ratoon cane). The full-sib progeny was then clonally propagated for the second field experiment that was set in Ribeirão Preto, SP, Brazil, in 2011. A randomized complete block design with three replicates was used and evaluations were carried in 2012 (plant cane), 2013 and 2014 (ratoon cane). Both parents were included in each block of the two experiments. All replicates were used to collect phenotypes for stalk diameter (cm), stalk height (cm) and stalk weight (kg) in both experiments. Also, two blocks in each experiment were used to collect phenotypes for soluble solids content (Brix), sucrose content and fiber percentage.

Genotyping

Parents and F$_1$ progeny were genotyped using the genotyping-by-sequencing protocol of Elshire *et al.* (2011). Reduced representation libraries were prepared using the PstI restriction enzyme. PstI is a rare-cutting enzyme, because its restriction site has a length of 6 bp, such that it allows a higher genotyping depth (Poland and Rife 2012). Four lanes containing 96-plex libraries were sequenced using the Illumina GAIIX and, subsequently, another four lanes with the same 96-plex libraries were sequenced using the Illumina NextSeq500 platform. All genotyping-by-sequencing protocols were carried at Center of Molecular Biology and Genetic Engineering, University of Campinas, Campinas, Brazil (CBMEG/UNICAMP).

We performed the variant calling using a modified version of TASSEL-GBS pipeline (Pereira *et al.* 2018). This version provides exact read counts of the alleles at each SNP locus. We used default values in all plugins of the pipeline, except for the MergeDuplicateSNPs plugin, in which we used the argument *callHets* and set the *misMat* argument value to 0.3. These values were chosen in order to allow a greater number of heterozygous SNP loci to be kept in subsequent steps. The sequenced reads were aligned to the methyl-filtrated assembly of the sugarcane genome (Grativol *et al.* 2014), using the software Bowtie2 (Langmead and Salzberg 2012).

We used the read count information of each SNP to estimate their ploidy level and call sample genotypes using the software SuperMASSA (Mollinari and Serang 2015; Pereira *et al.* 2018). Ploidy levels ranging from two to 20 were evaluated and only SNPs with ploidy estimates between six and 14 were kept (Garcia *et al.* 2013). We also filtered for a minimum mean read depth per individual of 50 reads, maximum mean read depth per individual of 500 reads, minimum posterior probability of genotype configuration (argument *p*) of 0.8, minimum posterior probability of each genotype assignment (argument *n*) of 0.5, and minimum call rate of 50%. We then used R package *updog* (Gerard *et al.* 2018) to reestimate the genotypes of the SNPs that met the filtering criteria. The *updog* package has the advantage of accounting for allelic bias, overdispersion and sequencing errors when estimating SNP genotypes, given a predetermined ploidy level. Finally, based on the estimates of SNP genotypes in the parents, we performed a chi-squared segregation test on the population genotype class frequencies, considering a hypergeometric distribution of gametes (Mollinari and Serang 2015). Using the Bonferroni correction, only SNPs with p-values over a 5% threshold were kept.

Phenotypic mixed model analysis

Adjusted phenotypic means (i.e., BLUEs - best linear unbiased estimates) for each individual were obtained using a two-stage analysis (Damesa *et al.* 2017). All analyses were performed using ASReml-R (Butler *et al.* 2009). Stage one consisted of a within-site analysis, where the genotype effect was considered fixed and the remaining effects were considered as random (harvest effects, blocks-within-harvest effects and genotype × harvest interaction effects). The covariance matrix ($\boldsymbol{\Omega_j}$) for the vector of genotype effect estimates ($\boldsymbol{u_j}$) in site *j* was obtained from the inverse of the coefficient matrix of the mixed model equations, returned as *Cfixed* in the asreml object (Endelman *et al.* 2018). Stage two was a joint analysis considering the two sites, using the following linear model:

$$\hat{u}_{ij} = \mu + g_i + s_j + (gs)_{ij} + e_{ij},$$

where $\hat{u}_{ij}$ is the genotype effect estimate obtained in the stage one analysis, the parameter μ is the intercept, g_i is a fixed effect of genotypes, s_j is a random effect of sites, $(gs)_{ij}$ is a random effect for the genotype × site interaction, and the variance of the residual e_{ij} is $(\omega^{ij})^{-1}$, where ω^{ij} is the *i*-th diagonal element of $\mathbf{\Omega_j}^{-1}$ from the stage one analysis (Damesa *et al.* 2017). The BLUEs of the genotypes obtained after this stage were subsequently used to adjust the genomic selection models.

For phenotypic variance partitioning and to obtain estimates of heritability of the traits, the phenotypic values were standardized and the following random linear model was used for each trait:

$$y_{ijkl} = \mu + g_i + s_j + h_k + b_{l(jk)} + (gs)_{ij} + (gh)_{ik} + (gsh)_{ijk} + e_{ijkl},$$

where the parameter μ is the intercept, g_i is the effect of genotypes, s_j is the effect of sites, h_k is the effect of harvest, $b_{l(jk)}$ is the effect of replicates within sites and harvests, $(gs)_{ij}$ is the effect of the genotype × site interaction, $(gh)_{ik}$ is the effect of the genotype × harvest interaction, $(gsh)_{ijk}$ is the effect of the genotype × site × harvest interaction, and e_{ijkl} is the residual effect.

The estimates of heritability were obtained using:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{gs}^2}{s} + \frac{\sigma_{gh}^2}{h} + \frac{\sigma_{gsh}^2}{c_{sh}} + \frac{\sigma_e^2}{r}},$$

where σ_g^2, σ_{gs}^2, σ_{gh}^2, σ_{gsh}^2, and σ_e^2 are the genotypic variance, the variance of the genotype × site interaction, the variance of the genotype × harvest interaction, the variance of the genotype × site × harvest interaction, and the residual variance, respectively. The values s, h, c_{sh}, and r correspond to the number of sites, the number of harvests, the number of combinations of sites and harvests, and the total (combined) number of replicates of both experiments, respectively.

Genomic selection models

We incorporated allele dosage information in our genomic selection models by expanding and adapting the GBLUP methodology for autotetraploid species proposed by Endelman *et al.* (2018). In sugarcane, besides the higher ploidy, the model also has to account for different ploidy levels among SNP loci. In order to achieve that, we expanded the theory by adapting the estimation of both the genomic covariance matrix of the additive values (**G**) and the genomic covariance matrix of digenic dominance values (**D**).

Genomic predictions were obtained using the following linear model:

$$\hat{y}_i = \mu + g_i + e_i,$$

where $\hat{y}_i$ is the BLUE of the *i*-th individual obtained with the two-stage phenotypic analysis, μ is the intercept, g_i is the random effect of genotypes, and e_i is the random residual effect.

We used two covariance structures in the genomic selection model: i) $\mathbf{I}\mathrm{V}_r + \mathbf{G}\mathrm{V}_a$, and ii) $\mathbf{I}\mathrm{V}_r + \mathbf{G}\mathrm{V}_a + \mathbf{D}\mathrm{V}_d$, where V_r is the residual variance, V_a is the additive genetic variance, and V_d is the dominance genetic variance. All analyses were performed using ASReml-R (Butler *et al.* 2009).

Genomic covariance matrix of additive values (G)

Consider a matrix **X** with *n* rows and *m* columns, the rows corresponding to the individuals in the population and the columns corresponding to SNP loci, where each element x_{ij} corresponds to the dosage of the alternative allele for the *j*-th SNP in the *i*-th individual. Because the SNPs have different ploidy levels, the same value of allele dosage for one SNP does not represent the same genotype for other SNPs with different ploidies. For example, for a hexaploid SNP an allele dosage value of six represents a homozygous locus, while for an octoploid SNP the same value represents a heterozygous locus.

To account for the different ploidy levels between SNPs, we used the following formula:

$$\mathbf{Z} = 2\mathbf{X}\mathbf{M}^{-1},$$

where **M** is an $m \times m$ diagonal matrix of ploidy values, such that each diagonal element m_j corresponds to the ploidy of the *j*-th SNP locus. The resulting matrix **Z**, with the same dimensions of **X,** has all its elements varying from 0 to 2, where 0 represents loci that are homozygous for the reference allele and 2 represents loci that are homozygous for the alternative allele, the values in between corresponding to heterozygous loci.

The subsequent steps to obtain **G** are the same as for diploids (VanRaden 2008). If p_j is the frequency of the alternative allele at the *j*-th locus, we can obtain an $n \times m$ matrix **P** where the values in the *j*-th column all correspond to p_j. Subtracting $2\mathbf{P}$ from **Z** results in the matrix **W** of centered genotypes. The **G** matrix is then obtained by the formula:

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}^{\mathrm{T}}}{\sum_j 2p_j\left(1-p_j\right)}$$

Genomic covariance matrix of digenic dominance values (D)

We first introduce the expansion of the digenic dominance values in the autotetraploid model to a hexaploid scenario. Higher ploidy levels can be modeled in a similar fashion. Considering a hexaploid SNP locus with two alleles B and b, the digenic effect for each allele pair can be obtained as demonstrated by Endelman *et al.* (2018), with the following set of equations:

$$\beta_{BB} = q^2\beta$$
$$\beta_{Bb} = -pq\beta$$
$$\beta_{bb} = p^2\beta, \qquad \text{(Eq. 1)}$$

where p is the allele frequency of B, q is the allele frequency of b, with $q = 1 - p$, and β is the digenic dominance effect, such that:

$$\beta = \beta_{BB} - 2\beta_{Bb} + \beta_{bb}.$$

For a hexaploid locus, seven genotypic classes are possible in a population (i.e., allele dosages ranging from 0 to 6). For each genotypic class, different combinations of digenic effects are present. For example, for the genotypic class BBBBbb, there are 6 possible combinations of two B alleles, 8 possible combinations of a B allele with a b allele, and 1 possible combination of two b alleles. By replacing each digenic effect by their corresponding values in (Eq. 1), we obtain the total digenic dominance for each dosage of B. Table 1 shows the combinations of digenic effects and the total digenic dominance for each allele dosage level of a hexaploid locus.

Table 2. Digenic effects and total digenic dominance for each allele dosage level of a hexaploid locus with alleles B and b

Dosage of allele B	Digenic effects	Digenic dominance
6	$15\beta_{BB}$	$\left(15p^2-30p+15\right)\beta$
5	$10\beta_{BB}+5\beta_{Bb}$	$\left(15p^2-25p+10\right)\beta$
4	$6\beta_{BB}+8\beta_{Bb}+\beta_{bb}$	$\left(15p^2-20p+6\right)\beta$
3	$3\beta_{BB}+9\beta_{Bb}+3\beta_{bb}$	$\left(15p^2-15p+3\right)\beta$
2	$\beta_{BB}+8\beta_{Bb}+6\beta_{bb}$	$\left(15p^2-10p+1\right)\beta$
1	$5\beta_{Bb}+10\beta_{bb}$	$\left(15p^2-5p\right)\beta$
0	$15\beta_{bb}$	$\left(15p^2\right)\beta$

The formula to obtain the total digenic dominance for a given biallelic hexaploid locus can then be generalized as:

$$\delta=\left(15p^2-5ap+\frac{1}{2}a(a-1)\right)\beta \ , \qquad \text{(Eq. 2)}$$

where δ is the total digenic dominance and *a* is the allele dosage.

We used the same process described for hexaploid loci to obtain equations for other levels of ploidy. Table 2 shows the generalized formulas to obtain the total digenic dominance for even ploidies from six through 14.

Table 2. Formulas for the total digenic dominance for different levels of ploidy

Ploidy	Total digenic dominance
6	$\left(15p^2-5ap+\frac{1}{2}a(a-1)\right)\beta$
8	$\left(28p^2-7ap+\frac{1}{2}a(a-1)\right)\beta$
10	$\left(45p^2-9ap+\frac{1}{2}a(a-1)\right)\beta$
12	$\left(66p^2-11ap+\frac{1}{2}a(a-1)\right)\beta$
14	$\left(91p^2-13ap+\frac{1}{2}a(a-1)\right)\beta$

The formulas in Table 2 can be generalized as:

$$\mathbf{Q}\beta = \left(\mathbf{P} \odot \mathbf{PC} - \mathbf{P}(\mathbf{M}-1) \odot \mathbf{X} + \frac{1}{2}\mathbf{X} \odot \left(\mathbf{X}-1\right) \right)\beta,$$

where $\odot$ represents the Hadamard product, $\mathbf{C}$ is an $m \times m$ diagonal matrix where each diagonal element c_j corresponds to $\binom{m_j}{2}$, and **P**, **M** and **X** are as previously defined.

Finally, the genomic covariance matrix of digenic dominance values (**D**) was obtained with:

$$\mathbf{D} = \frac{\mathbf{QQ}^{\mathbf{T}}}{\sum_j c_j p_j^2 (1-p_j)^2}.$$

Model and marker set comparisons

We compared two models for the genotype effects, one using only the **G** matrix (G model) and one using both the **G** and **D** matrices (G+D model). We also investigated the effect of using three different sets of genotypic information: i) a fully informative model considering SNP markers with ploidy and allele dosage estimates, ii) diploidized SNP markers, and iii) simplex SNP markers. The diploidized SNP set was obtained by setting the values of all heterozygous loci in matrix **Z** to 1. By doing so, all heterozygous genotypes were effectively merged in a single class, regardless of their dosage. The simplex SNP set was obtained by selecting only loci that had dosage of 0 or 1 for the reference or the alternative allele in both parents. For diploidized and simplex markers, the **G** and **D** matrices were obtained according to the established methodology commonly used for diploids (VanRaden 2008; Vitezica *et al.* 2013).

The models were compared in terms of predictive ability and selection coincidence. For that, 1,000 cross-validation runs were carried, such that in each run 10% of the population was sampled and used as the validation set, while the remaining 90% were used as the training set. We measured predictive ability as the correlation between predicted breeding values and BLUEs of the individuals in the validation set. The selection coincidence was measured as the proportion of coincident individuals selected when using predicted breeding values or BLUEs as selection criteria, with a selection intensity of 50% in the validation set.

Influence of sequencing depth

When performing genotyping-by-sequencing in polyploids, one must account for the fact that high read depths are needed in order to obtain accurate estimates of allele dosage (Uitdewilligen *et al.* 2013; Matias *et al.* 2019). In order to verify how genotyping depth would affect the prediction accuracy of genomic selection in this polyploid scenario, we simulated SNP datasets obtained with fewer sequenced lanes. The simulated datasets were derived from the final SNP set obtained after filtering. For that, we multiplied the read depth of each allele in every locus in the population by a fraction *k*, the value of *k* corresponding to the number of sequenced lanes we were simulating divided by the original number of sequenced lanes (eight). All simulated genotyping depths and the corresponding values of *k* are shown in Table 3.

Table 3. Simulated genotyping depths and fraction *k* of the original read depth used to simulate the read depth in each scenario.

Genotyping depth	***k***
6 lanes	$^3/_4$
4 lanes	$^1/_2$
3 lanes	$^3/_8$
2 lanes	$^1/_4$
1 lane	$^1/_8$

We then reestimated the ploidy and allele dosage of the SNPs in each simulated dataset. The estimates of ploidy were obtained using software SuperMASSA (Mollinari and Serang 2015) with no other filtering criteria other than choosing only SNPs with ploidy estimates between six and 14. The estimates of allele dosage were again obtained using *updog* (Gerard *et al.* 2018). As the genotyping depths decrease, the increasing uncertainty when estimating ploidy and allele dosage could lead to inconsistent results when the process is repeated in the same dataset. To account for this random variation in the estimates of ploidy and allele dosage, this step was replicated 10 times for each simulated dataset. For each simulated replicate, the ploidy and dosage estimates were used to obtain the **G** matrix and perform genomic selection. For each simulated dataset, we measured the mean predictive ability of 1,000 cross-validation runs per replicate. Genomic selection and cross-validation were performed as described in previous sections.

3.3. Results

We were able to obtain a large number of SNPs with estimates of ploidy and allele dosage. However, the genomic selection models showed low prediction ability, and the prediction ability values showed little sensitivity to including ploidy and allele dosage information or dominance effects in the model. On the other hand, the selection coincidence values showed advantage of including ploidy and allele dosage estimates over using diploidized markers. The low values of prediction accuracy were consistent with the low genotypic variation and low to intermediate values of heritability observed in the phenotypic analysis of the progeny.

In the following we present our results in three sections. First, we present the results we obtained by genotyping the progeny and parents. Second, we present the results we obtained with the phenotypic data analysis of the progeny. Finally, we present the results we obtained using different genomic selection models with different genotypic datasets.

Genotyping

The distribution of mean read depth per individual of the SNPs we identified with TASSEL-GBS is shown in Fig. 1. Overall, we identified 187,224 SNPs, most of which had mean read depths close to zero. A total of 6,550 SNPs were kept after filtering for mean read depth, posterior probability of genotypes and ploidy estimates, call rate, and segregation distortion in the progeny. The mean read depth per SNP per sample in the filtered set is shown in Fig. 2. A total of 11 individuals had a mean read depth of zero and were considered not genotyped, thus being used in phenotypic analyses but not for genomic selection. The overall mean read depth of the individuals in the F_1 progeny was of 165. The parents were genotyped at a much higher depth, with mean read depths per SNP of 1,508 and 1,850 reads for IACSP95-3018 and IACSP93-3046, respectively.

A summary of ploidy and allele dosage estimates of the SNPs in the filtered set is shown in Fig. 3. Additionally, a summary of ploidy and allele dosage estimates of SNPs that did not pass the segregation distortion test in the population is shown in Fig. S1 of the supplementary material. The majority of the SNPs had ploidy estimates of ten (31.18%) and eight (28.93%), followed by 17.88% of SNPs with ploidy estimates of 12, 15.59% with an estimated ploidy of six, and 6.43% with ploidy 14. Within each ploidy level, most of the genotypes were either homozygous for the reference allele or had only one copy of the reference allele (that is, were in nulliplex or simplex configuration), with allele dosages of zero and one accounting for more than 50% of the total number of genotype calls for ploidy levels from 6 to 12. For ploidy 14, dosage estimates were more evenly distributed among different levels, but there was still an excess of dosages equal to zero and one. When considering only the parents, 4,362 out of the 6,550 SNPs we obtained had dosage of 0 or 1 for the reference or the alternative allele.

Phenotypic analyses

In general, the genotypic variance had a relatively small or intermediate magnitude for all of the traits, with traits accordingly showing either small or intermediate heritability values. Fig. 4 shows the partitioning of the phenotypic variance into its main components. Variance components that are not shown had variance estimates very close to zero. The residual variance had a large magnitude for all of the traits, corresponding to 36%, 35%, 49%, 58%, 48% and 34% of the phenotypic variation observed for Brix, sucrose content, fiber percentage, stalk diameter, stalk weight and stalk height, respectively.

Most of the phenotypic variation of traits Brix and sucrose content was due to the effect of sites, with the variance due to this component respectively corresponding to 44% and 37% of the observed phenotypic variation. A small effect of sites was observed for stalk height, with the variance due to this component corresponding to 10% of the total variation. The distribution and correlation between phenotypic measurements per experimental site are shown in Fig. S2 of the supplementary material. The effect of harvests was intermediate for traits sucrose content, fiber percentage, stalk weight and stalk height, with the variance due to the harvest component corresponding to approximately 15% of the phenotypic variation observed for these traits. For traits Brix and stalk diameter, the variance due to the harvest component corresponded to 6% and 4% of the observed phenotypic variation, respectively. The effect of replicates within sites and harvests had a large magnitude for traits stalk height and stalk weight, with the variance due to this component corresponding to 30% and 24% of the observed phenotypic variation, respectively. For traits Brix, sucrose content, fiber percentage and stalk diameter, the variance due to the effect of replicates within sites and harvests corresponded to 7%, 10%, 18% and 9% of the observed phenotypic variation, respectively.

The effect of genotypes had an intermediate magnitude for stalk diameter and a small magnitude for the other traits, corresponding to 3%, 3%, 7%, 13%, 5% and 3% of the phenotypic variation observed for Brix, sucrose content, fiber percentage, stalk diameter, stalk weight and stalk height, respectively. The genotypes × site interaction effect had an intermediate magnitude for traits fiber percentage, stalk diameter and stalk weight, with the variance due to the interaction component corresponding to, respectively, 13%, 15% and 10% of the observed phenotypic variation. For traits Brix, sucrose content and stalk height the variance due to the interaction component corresponded to 4%, 2% and 6% of the observed phenotypic variation, respectively. The heritability values for traits Brix, sucrose content, fiber percentage, stalk diameter, stalk weight, and stalk height were of 0.31, 0.35, 0.37, 0.55, 0.41, and 0.36, respectively.

Genomic selection

Overall, the predictive abilities of the genomic selection models were low, regardless of the model or marker set utilized. Fig. 5 shows the distribution of the predictive ability values over different cross-validation runs of the G and G+D models when using all the makers with full ploidy and allele dosage information, using diploidized makers, and using only simplex markers.

For Brix, the G model using ploidy and allele dosage estimates showed the highest mean predictive ability (0.24), with a mean predictive ability higher than the mean predictive ability of the corresponding G+D model (0.21), and higher than the mean predictive abilities when using diploidized markers (0.18 for the G model and 0.19 for the G+D model) and using simplex markers (0.21 for the G model and 0.18 for the G+D model). A similar pattern was observed for stalk height, where the G model using ploidy and allele dosage estimates had a mean predictive ability of 0.22, the full ploidy G+D model had a mean predictive ability of 0.19, and when using diploidized or simplex markers, the mean predictive ability did not exceed 0.18 for any of the two models.

For sucrose content, the G+D model had lower mean predictive abilities in comparison to the additive G model for all sets of markers, and the mean predictive abilities of the G model did not differ considerably between sets of markers. We observed a different pattern for stalk diameter, because the mean predictive ability of the G model when using ploidy and allele dosage estimates (0.18) was slightly lower than the mean predictive ability when using diploidized markers or simplex markers (0.20). With regard to the G+D model, the mean predictive abilities were equivalent for all sets of markers. A more marked difference between models was noticeable for fiber percentage, because when using ploidy and allele dosage estimates the mean predictive ability of the G+D model (0.05) was much lower than for the G model (0.12). This, in turn, was lower than the mean predictive ability when using diploidized markers (0.15 for the G and G+D models) and equal to the mean predictive ability of both the additive and additive + dominance models when using simplex markers. Lastly, for stalk weight, the mean predictive abilities were the highest among all traits, and the values did not differ significantly between models or sets of markers (ranging from 0.28 to 0.29).

The selection coincidence varied slightly over different traits, but for all six traits the same pattern was observed, with a much lower mean selection coincidence when using diploidized markers than those obtained when using ploidy and allele dosage estimates or simplex markers. Table 4 shows the mean selection coincidence of the G model and the G+D model with the three contrasting marker sets. The mean selection coincidence of the G model when using ploidy and allele dosage estimates was equal to 0.62, 0.61, 0.57, 0.56, 0.62, and 0.59 for traits Brix, sucrose content, fiber content, stalk diameter, stalk weight, and stalk height, respectively. When using diploidized markers, the mean selection coincidence of the G model dropped to 0.32, 0.35, 0.37, 0.34, 0.38, and 0.40 for the same traits. Finally, when using simplex markers, the corresponding mean selection coincidences of the G model were 0.61, 0.61, 0.58, 0.55, 0.61, and 0.60. For all traits, the mean selection coincidences when using the G model and the G+D model were nearly equivalent regardless of the marker set used. The largest difference between the models was observed when using simplex markers for prediction of stalk height, when the mean selection coincidence of the G and G+D models were 0.60 and 0.57, respectively.

Table 4. Mean selection coincidence of genomic selection models when considering additive effects only (G) and considering additive and digenic dominance effects (G+D). Both models are compared when using markers with ploidy and allele dosage estimates (Full ploidy), diploidized markers and using only simplex markers. The values are shown for traits soluble solids content (Brix), sucrose content (Pol), fiber percentage (Fiber), stalk diameter (Diam), stalk weight (Weight) and stalk height (Height). Standard errors of the means had very low magnitude (≤ 0.006) and are not shown.

Model	**Markers**	**Mean selection coincidence**					
		Brix	**Diam**	**Fiber**	**Height**	**Pol**	**Weight**
G	Full Ploidy	0.62	0.56	0.57	0.59	0.61	0.62
G	Diploidized	0.32	0.34	0.37	0.4	0.35	0.38
G	Simplex	0.61	0.55	0.58	0.6	0.61	0.61
G+D	Full Ploidy	0.61	0.56	0.55	0.58	0.6	0.63
G+D	Diploidized	0.31	0.35	0.36	0.4	0.34	0.38
G+D	Simplex	0.59	0.54	0.58	0.57	0.6	0.62

The results obtained by downsampling the allele read depths did not show a clear pattern of decreasing predictive abilities as the read depth decreased. Fig. 6 shows the distribution of the predictive ability values over different cross-validation runs of the full ploidy G model when using SNPs with their original read depth and when using five simulated SNP sets with different (lower) levels of genotyping depth. Brix was the only trait for which the G model with the original genotyping depth (8 lanes) had the highest mean predictive ability value when compared to the other SNP sets. The mean predictive ability for Brix was of 0.25 when using the original genotyping depth and dropped to 0.23, 0.24, 0.23, 0.23 and 0.20 for the simulated sequentially decreasing genotyping depths.

For traits sucrose content and stalk height, the highest mean predictive ability value was achieved with the simulated genotyping depth of 4 lanes, with decreasing mean predictive abilities values as the simulated genotyping depths decreased further. The mean predictive ability for sucrose content was of 0.18, 0.18, 0.19, 0.17, 0.15, and 0.12 for the original genotyping depth, and simulated genotyping depths of 6, 4, 3, 2 and 1 lane, respectively. The mean predictive ability for stalk height was of 0.20, 0.21, 0.22, 0.20, 0.18, and 0.17 for the original genotyping depth, and for the simulated sequentially decreasing genotyping depths, respectively.

For stalk weight, the mean predictive ability was the lowest when using the lowest simulated genotyping depth, but for the other levels of genotyping depth there was no consistent variation trend. The mean predictive ability for stalk weight was of 0.26, 0.27, 0.26, 0.24, 0.25, and 0.21 from the highest to the lowest sequencing depth. For fiber content, conversely, the lowest simulated genotyping depth led to the highest mean predictive ability, but otherwise there was no apparent pattern, as was the case for stalk weight. The mean predictive abilities for fiber content were 0.10, 0.12, 0.11, 0.10, 0.13, and 0.18 from the highest to the lowest sequencing depth. For stalk diameter also no consistent variation trend was observed, with the mean predictive ability being 0.18, 0.20, 0.17, 0.15, 0.19, and 0.17 from the highest to the lowest sequencing depth.

3.4. Discussion

Our results help demonstrate the potential of using genotyping-by-sequencing to improve genomic studies in sugarcane, as the technique allows identifying a large number of SNPs with estimates of ploidy and allele dosage. Our results show that these estimates could potentially improve genomic selection in sugarcane, but also highlight that for this to be thoroughly achieved, good quality data is required for model training.

In the following we present our discussion in two sections. First, we discuss the genotyping results we obtained. More specifically, we focus on how our results relate to current knowledge of the sugarcane genome and also how they relate to what is expected from a sugarcane biparental progeny. Second, we discuss the results we obtained implementing genomic selection in sugarcane, giving perspectives on how to improve the prediction ability and achieve the full potential of including allele dosage information in the model.

Genotyping

We were able to identify 6,550 SNPs with high mean read depths, high posterior probability of genotypes and ploidy estimates, and which were segregating in the population accordingly to expected (Fig. 1). Also, despite the low uniformity across samples, the overall genotyping depth in the population was high (Fig 2). This was possible due to the genotyping-by-sequencing protocol we used. The use of a rare-cutting enzyme, the low number of samples per sequenced lane and high sequencing depth (8 sequenced lanes) were likely the factors that guaranteed the high genotyping depth we observed. These results indicate that genotyping-by-sequencing can represent a substantial advance for genetic studies in sugarcane, as also demonstrated by Balsalobre *et al.* (2017). Our SNP set exceeds in number of markers most recent genetic studies in sugarcane (Bundock *et al.* 2009; Gouy *et al.* 2013; Costa *et al.* 2016; Yang *et al.* 2017; Gutierrez *et al.* 2018), and is also more informative than sets of dominant or single-dosage markers, which disregard information on other possible allele dosages.

The ploidy estimates we obtained in the filtered dataset showed a predominance of markers with ploidies 8 and 10, intermediate proportions of markers with ploidies 6 and 12 and a small proportion of markers with ploidy 14 (Fig. 3). The high proportion of octoploid markers is consistent with previous estimates obtained using biparental progenies (Garcia *et al.* 2013; Balsalobre *et al.* 2017). The results reported by Garcia *et al.* (2013) considered SNPs for the same progeny used here, but were obtained through the Sequenom iPLEX MassARRAY platform (Gabriel *et al.* 2009), which is a highly reliable genotyping method. The genomes of modern sugarcane cultivars are mostly composed of *S. officinarum* (2n = 80, x = 10) chromosomes, with contributions from *S. spontaneum* (2n = 40 to 128, x = 8) (D'Hont et al. 1996, 1998). Studies with S. spontaneum show that the species displays a wide range of chromosome numbers, with the five major cytotypes having 8, 10, 12, 14 and 16 sets of eight chromosomes (D'Hont et al. 1998). In short, our results also agree with the overall ploidy levels expected for both species. However, to date little is known about the behavior of sugarcane chromosomes during meiosis, and the ploidy estimate of a single locus might represent chromosome segments from both species. Therefore, is not possible to assign variants as representing specifically either *S. officinarum* or *S. spontaneum* based on their estimated ploidy levels alone.

The high relative proportion of markers with ploidies 12 and 14 in the raw SNP dataset is likely to be an over-representation. Before filtering out markers that were not segregating in the population according to expected Mendelian proportions, the distribution of markers across ploidy values was more uniform (Fig. S1). Genotyping-by-sequencing is likely to produce inconsistencies in the number of sites sequenced per sample (Heffelfinger *et al.* 2014) and in the number of reads per site (Jiang *et al.* 2016), which can result in errors when obtaining genotype estimates, especially for higher ploidy levels (Garcia *et al.* 2013). Also, incorrect genotype calls at higher ploidy levels may have led to more deviations from the expected segregation ratio and, thus, to more high-ploidy markers being filtered out.

A large proportion (67%) of the SNP calls corresponded to single-dosage genotypes. A possible explanation is that the parents used to generate the biparental population are commercial cultivars. In breeding programs, cultivars are obtained after an extensive process of selection for several traits, which could lead to fixation of favorable alleles and, in consequence, predominance of either homozygous genotypes for favorable alleles or heterozygotes that have only one copy of deleterious alleles. Our results are also consistent with those observed by

Garcia *et al.* (2013) using the same biparental population. They observed that, for ploidies between six and 12, there was a higher proportion of single-dosage markers in comparison to the proportion observed for higher ploidy values. As shown in Fig. 3 and Fig. S1, for markers with an estimated ploidy of 14 the proportion of single-dosage markers was smaller. The results obtained by Balsalobre *et al.* (2017), also with a biparental population, showed 56% of their SNP set to be single-dosage genotype calls.

Genomic selection

When implementing genomic selection in breeding programs, the main factors that affect the prediction ability of the models are the heritability of the traits, the size of the training set used to estimate marker effects and the marker density used to genotype the population (Combs and Bernardo 2013; Lian *et al.* 2014). Our results show that, for all evaluated traits, the prediction abilities of the genomic selection models were low regardless of using the whole SNP set or only simplex markers (Fig. 5). This indicates that, with simplex markers representing more than 50% of the dataset, the increased marker density by using SNPs with higher dosages may not have resulted in a higher number of markers in linkage disequilibrium to quantitative trait loci. An alternative hypothesis could be that the heritability of the traits and training population size are playing a bigger role in the performance of the genomic selection models.

The phenotypic variance partitioning results showed that, for all traits, most of the variation observed in the field experiments did not stem from differences between the individuals in the F_1 progeny, as the variance components associated to the effect of genotypes and genotype × environment interactions had low magnitude in comparison to other experimental sources of variation. The low magnitude of the genotypic variance indicates that the parents used for crossing were not contrasting enough for the traits evaluated in this study, especially for Brix, sucrose content and stalk height. Both parents are cultivars developed by the sugarcane breeding program at IAC, IACSP95-3018 and IACSP93-3046. The first two numerical digits in their identifiers indicate the year when these cultivars were first originated from crosses and the selection process began, respectively 1995 and 1993. The fact that both parents correspond to elite material from not-so-distant time points of the same breeding program could be the reason for the low genotypic variability generated by their crossing.

These low values of genotypic variability resulted in low to intermediate values of heritability, which in turn are usually associated with lower values of predictive ability (Combs and Bernardo 2013; Lian *et al.* 2014). For all of the traits we evaluated, several studies have reported higher heritabilities when analyzing data from sugarcane cultivar trials (Milligan *et al.* 1990; Gravois and Milligan 1992; Tena *et al.* 2016). This indicates that implementing genomic selection in sugarcane is likely to be more advantageous than our results suggest. We note, however, that the prediction abilities we observed for each trait did not increase with increasing values of heritability, which were all relatively similar. Other factors such as the different genetic architecture of the traits (i.e., distribution and effects of quantitative trait loci across the genome) might be influencing more strongly the differences between the predictive performance of the models.

The small training population size used in this study might also be playing a key role in the low values of predictive ability we observed. This is particularly suggested by the reduction, for most of the traits, of the predictive ability when including the digenic dominance effects in comparison to only using the allele dosage information to estimate additive effects. Including digenic dominance effects result in estimating three additional parameters (Eq. 1), thus requiring more observations for accurate estimates to be obtained (Button *et al.* 2013). With a small population size, the estimates of dominance effects were likely not accurate, and the predictive ability of the model decreased.

We hypothesize that these two factors combined (low heritabilities and small training population size) are probably the reason for observing no substantial difference between the predictive ability of genomic selection models using markers with allele dosage information or diploidized markers. We also believe that these factors were the main reason our downsampling results did not show the expected trend of lower sequencing depth leading to lower predictive abilities (Fig. 6). Under these limitations, the benefits of using high-quality allele dosage estimates could be masked by the low predictive power of the model. Our results demonstrate that expending more resources for sequencing to improve estimates of allele dosage may not always provide a return in investment if the phenotypic data for training the model is not adequate.

Even though the different genotypic datasets had little impact on the predictive abilities, the advantages of including allele dosage estimates can be more clearly seen when looking at the mean values of selection coincidence, which were higher in comparison to those achieved when using diploidized markers, with a difference in selection coincidence of approximately 20% (Table 4). This indicates that, ultimately, including allele dosage information can improve the deployment of genomic selection in sugarcane breeding programs.

Furthermore, to our knowledge genomic selection studies including allele dosage information were limited to the autotetraploid framework until now. Our study expands the theory to higher ploidy levels and, therefore, could be valuable for breeding programs implementing genomic selection on other crops with higher ploidies, such as sweet potato and some ornamental flowers and forage crops (Soltis *et al.* 2014).

3.5. Conclusion

Overall, including estimates of ploidy and allele dosage of the SNPs led to a modest improvement of genomic selection models in sugarcane. The improvement we observed is likely to be more evident with larger training population sets that also display higher genetic variability, which would allow the models to have more precision to accurately estimate both the additive and the digenic dominance effects.

REFERENCES

Aitken, K. S., M. D. McNeil, S. Hermann, P. C. Bundock, A. Kilian *et al.*, 2014 A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput Diversity Array Technology (DArT) markers. BMC genomics 15: 152–152.

Balsalobre, T. W. A., G. da Silva Pereira, G. R. A. Margarido, R. Gazaffi, F. Z. Barreto *et al.*, 2017 GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18:.

de Bem Oliveira, I., M. F. Resende, F. Ferrao, R. Amadeu, J. Endelman *et al.*, 2018 Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. bioRxiv.

Bernardo, R., and J. Yu, 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Science 47: 1082–1090.

Besse, P., G. Taylor, B. Carroll, N. Berding, D. Burner *et al.*, 1998 Assessing genetic diversity in a sugarcane germplasm collection using an automated AFLP analysis. Genetica 104: 143–153.

Bundock, P. C., F. G. Eliott, G. Ablett, A. D. Benson, R. E. Casu *et al.*, 2009 Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnology Journal 7: 347–354.

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel, 2009 ASReml-R reference manual. 160.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint *et al.*, 2013 Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14: 365.

Cheavegatti-Gianotto, A., H. M. C. de Abreu, P. Arruda, J. C. Bespalhok Filho, W. L. Burnquist *et al.*, 2011 Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. Tropical Plant Biology 4: 62–89.

Centre de coopération internationale en recherche agronomique pour le développement (CIRAD). https://www.cirad.fr/en/our-research/tropical-supply-chains/sugarcane/context-and-issues

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

Costa, E. A., C. O. Anoni, M. C. Mancini, F. R. C. Santos, T. G. Marconi *et al.*, 2016 QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. Euphytica 211: 1–16.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186: 713.

Damesa, T. M., J. Möhring, M. Worku, and H.-P. Piepho, 2017 One Step at a Time: Stage-Wise Analysis of a Series of Experiments. Agronomy Journal 109: 845–857.

D'Hont, A., L. Grivet, P. Feldmann, J. C. Glaszmann, S. Rao *et al.*, 1996 Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. Molecular and General Genetics MGG 250: 405–413.

D'Hont, A., D. Ison, K. Alix, C. Roux, and J. C. Glaszmann, 1998 Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. Genome 41: 221–225.

Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres *et al.*, 2017 Genomic Selection for Yield and Seed Protein Content in Soybean: A Study of Breeding Program Data and Assessment of Prediction Accuracy. Crop Science 57: 1325.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE 6: e19379.

Endelman, J. B., C. A. S. Carley, P. C. Bethke, J. J. Coombs, M. E. Clough *et al.*, 2018 Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato. Genetics 209: 77.

Gabriel, S., L. Ziaugra, and D. Tabbaa, 2009 SNP Genotyping Using the Sequenom MassARRAY iPLEX Platform. Current Protocols in Human Genetics 60: 2.12.1-2.12.18.

Garcia, A. A. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva *et al.*, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific Reports 3: 3399.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping Polyploids from Messy Sequencing Data. Genetics 210: 789.

Goldemberg, J., and P. Guardabassi, 2010 The potential for first-generation ethanol production from sugarcane. Biofuels, Bioproducts and Biorefining 4: 17–24.

Gouy, M., Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal *et al.*, 2013 Experimental assessment of the accuracy of genomic selection in sugarcane. Theoretical and Applied Genetics 126: 2575–2586.

Grativol, C., M. Regulski, M. Bertalan, W. R. McCombie, F. R. Da Silva *et al.*, 2014 Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus Saccharum. Plant Journal 79: 162–172.

Gravois, K. A., and S. B. Milligan, 1992 Genetic Relationships between Fiber and Sugarcane Yield Components. 32: 62–67.

Grivet, L., and P. Arruda, 2002 Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology 5: 122–127.

Gutierrez, A. F., J. W. Hoy, C. A. Kimbeng, and N. Baisakh, 2018 Identification of Genomic Regions Controlling Leaf Scald Resistance in Sugarcane Using a Bi-parental Mapping Population and Selective Genotyping by Sequencing. Frontiers in Plant Science 9: 877.

Hawkins, C., and L.-X. Yu, 2018 Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. The Crop Journal.

Heffelfinger, C., C. A. Fragoso, M. A. Moreno, J. D. Overton, J. P. Mottinger *et al.*, 2014 Flexible and scalable genotyping-by-sequencing strategies for population studies. BMC genomics 15: 979–979.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval syst. Crop Science 49: 1–12.

Huckett, B. I., and F. C. Botha, 1995 Stability and potential use of RAPD markers in a sugarcane genealogy. Euphytica 86: 117–125.

Jiang, Z., H. Wang, J. J. Michal, X. Zhou, B. Liu *et al.*, 2016 Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. International journal of biological sciences 12: 100–108.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2014 Genomewide Prediction Accuracy within 969 Maize Biparental Populations. Crop Science 54: 1514.

Matias, F., K. Meireles, S. Nagamatsu, S. Barrios, C. do Valle *et al.*, 2019 Expected Genotype Quality and Diploidized Marker Data from Genotyping-by-Sequencing of <em>Urochloa</em> spp. Tetraploids. bioRxiv 525618.

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: 1819–1829.

Milligan, S. B., K. A. Gravois, K. P. Bischoff, and F. A. Martin, 1990 Crop Effects on Broad-Sense Heritabilities and Genetic Variances of Sugarcane Yield Components. Crop Science 30: 344.

Mollinari, M., and O. Serang, 2015 Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms. Batley J. (eds) Plant Genotyping. Methods in Molecular Biology (Methods and Protocols) 1245:.

Nair, N. V., A. Selvi, T. V. Sreenivasan, and K. N. Pushpalatha, 2002 Molecular diversity in Indian sugarcane cultivars as revealed by randomly amplified DNA polymorphisms. Euphytica 127: 219–225.

Park, S., P. Jackson, N. Berding, and G. Inmam-Bamber, 2007 Conventional breeding practices within the Australian sugarcane breeding program. 29: 10.

Pereira, G. S., A. A. F. Garcia, and G. R. A. Margarido, 2018 A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC bioinformatics 19: 398–398.

Poland, J. A., and T. W. Rife, 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5: 92.

Racedo, J., L. Gutiérrez, M. F. Perera, S. Ostengo, E. M. Pardo *et al.*, 2016 Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. BMC plant biology 16: 142–142.

Resende, M. D. V., M. F. R. Resende, C. P. Sansaloni, C. D. Petroli, A. A. Missiaggia *et al.*, 2012 Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194: 116–128.

Slater, A. T., N. O. I. Cogan, J. W. Forster, B. J. Hayes, and H. D. Daetwyler, 2016 Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. The Plant Genome 9:.

Soltis, D. E., C. J. Visger, and P. S. Soltis, 2014 The polyploidy revolution then…and now: Stebbins revisited. American Journal of Botany 101: 1057–1078.

Tena, E., F. Mekbib, and A. Ayana, 2016 Heritability and Correlation among Sugarcane (&lt;i&gt;Saccharum&lt;/i&gt; spp.) Yield and Some Agronomic and Sugar Quality Traits in Ethiopia. American Journal of Plant Sciences 07: 1453–1477.

Uitdewilligen, J. G. A. M. L., A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser *et al.*, 2013 A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. PLOS ONE 8: e62355.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91: 4414–4423.

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195: 1223.

Wu, K. K., W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore *et al.*, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics 83: 294–300.

Yang, X., S. Sood, N. Glynn, M. S. Islam, J. Comstock *et al.*, 2017 Constructing high-density genetic maps for polyploid sugarcane (Saccharum spp.) and identifying quantitative trait loci controlling brown rust resistance. Molecular Breeding 37:.

Zhou, M., 2013 Conventional Sugarcane Breeding in South Africa: Progress and Future Prospects. American Journal of Plant Sciences 04: 189–197.

FIGURES

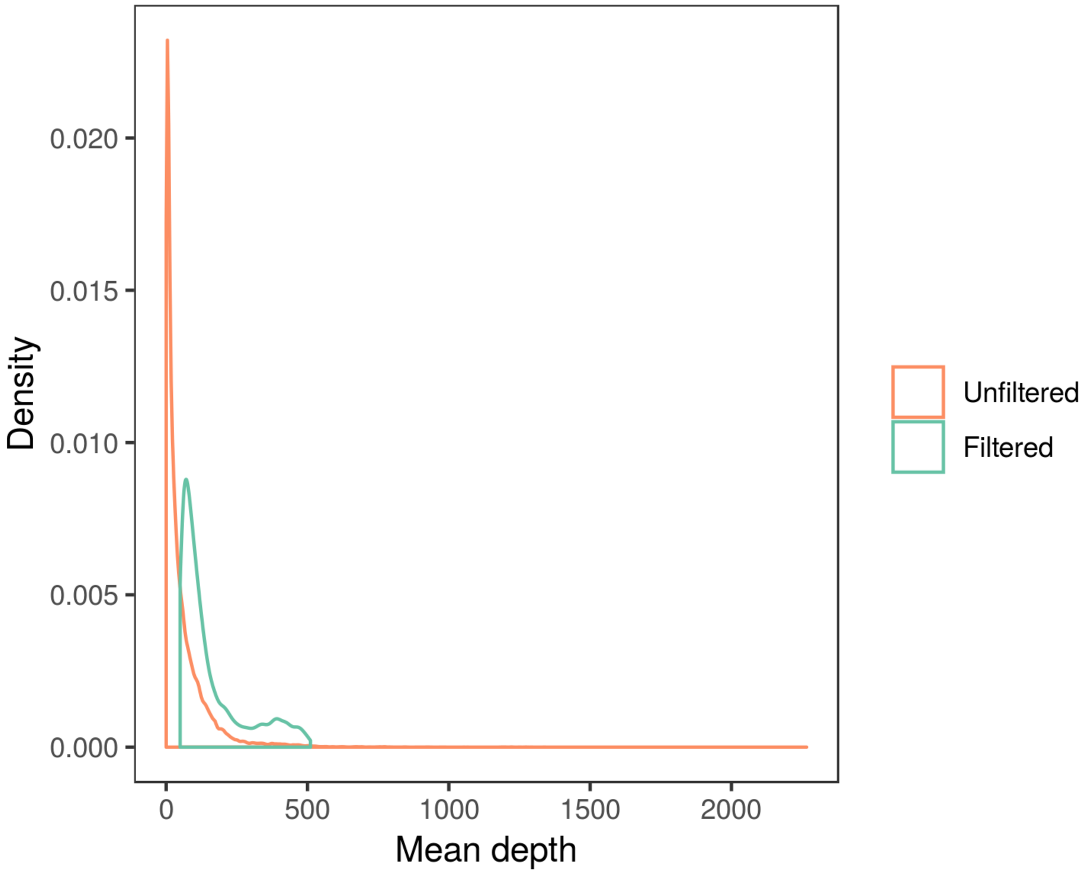

Figure 1. Mean sample read depth per SNP. The orange density curve corresponds to the raw SNP set without any filtering (187,224 SNPs) and the blue density curve corresponds to the SNPs kept after filtering (6,550 polymorphic sites).

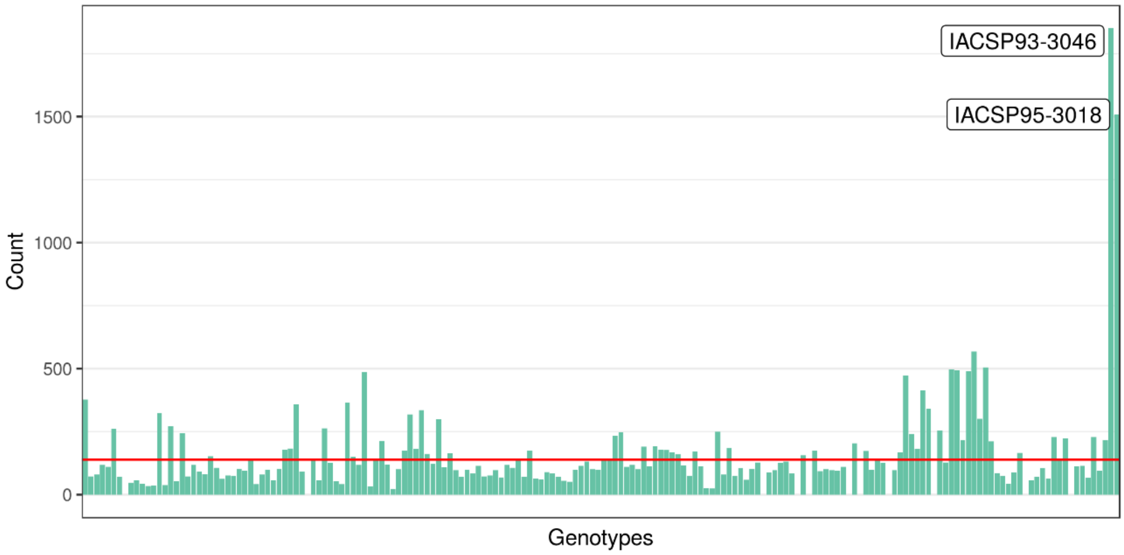

Figure 2. Mean SNP read depth per sample. Both parents (IACSP93-3046 and IACSP95-3018) are highlighted. Eleven individuals were entirely missed in the genotyping, with a mean read depth of zero. The red line indicates the overall mean read depth per SNP in the F_1 progeny.

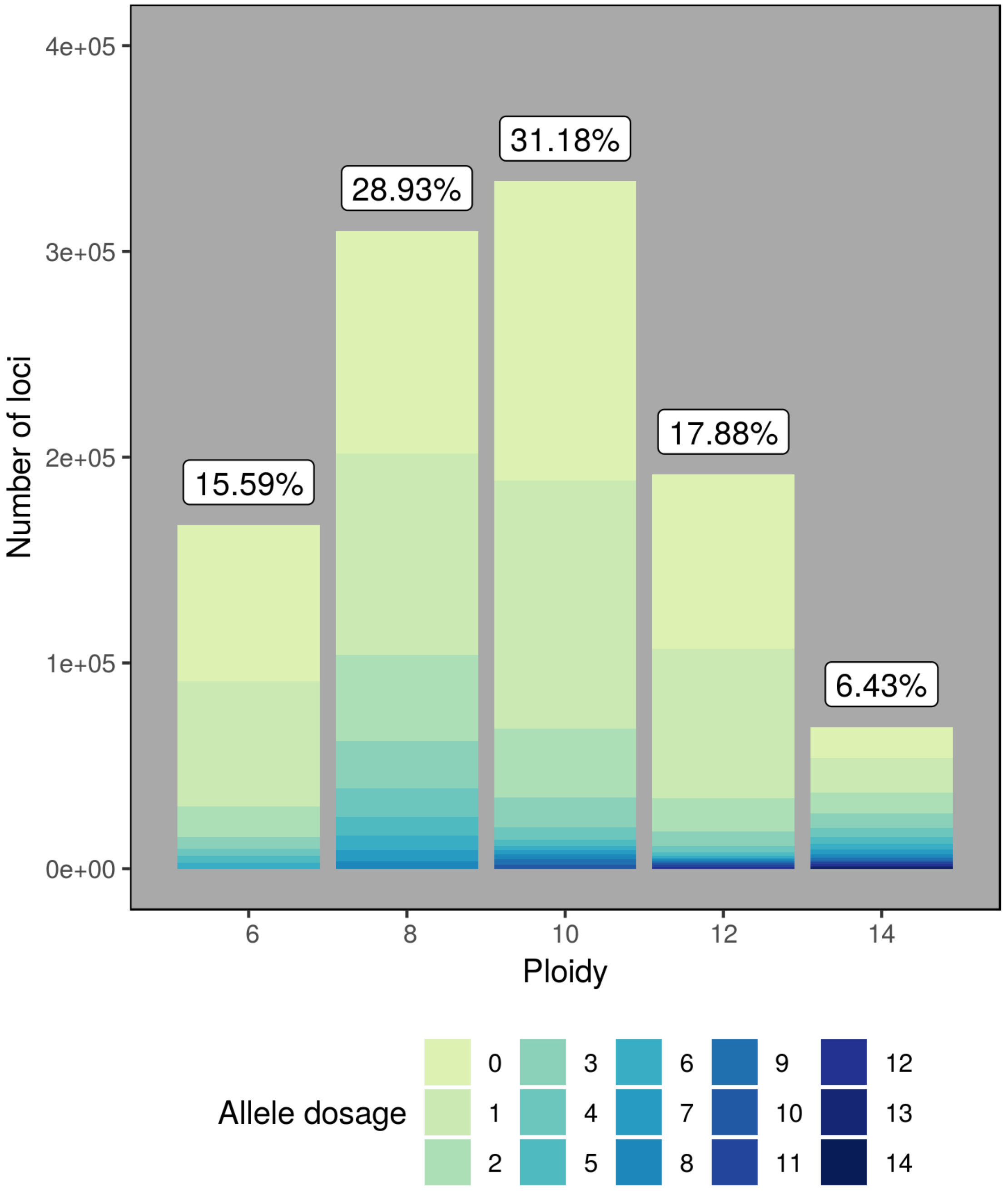

Figure 3. Summary of the estimates of ploidy and allele dosage for 170 samples and 6,550 filtered SNPs. The bars show the total number of genotypes per ploidy level, and different values of allele dosage are shown by different colours. For each ploidy level, the corresponding percentages of the total number of genotypes are shown above the bars.

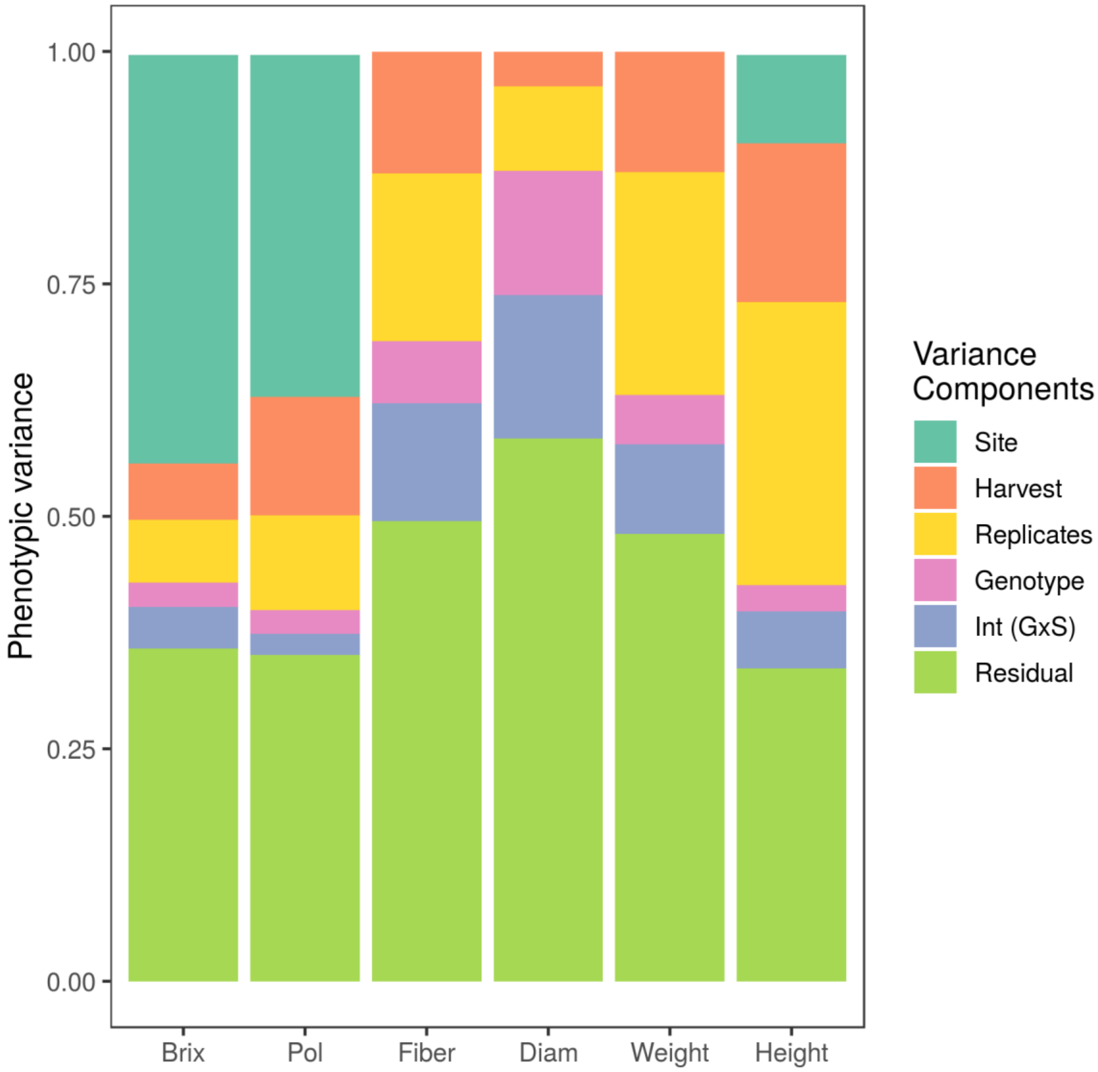

Figure 4. Phenotypic variance partitioning for soluble solids content (Brix), sucrose content (Pol), fiber percentage (Fiber), stalk diameter (Diam), stalk weight (Weight), and stalk height (Height). Contributions of variances due to the effect of sites, harvests, replicates, genotypes, genotype × sites interaction (GxS), and residual variance are shown.

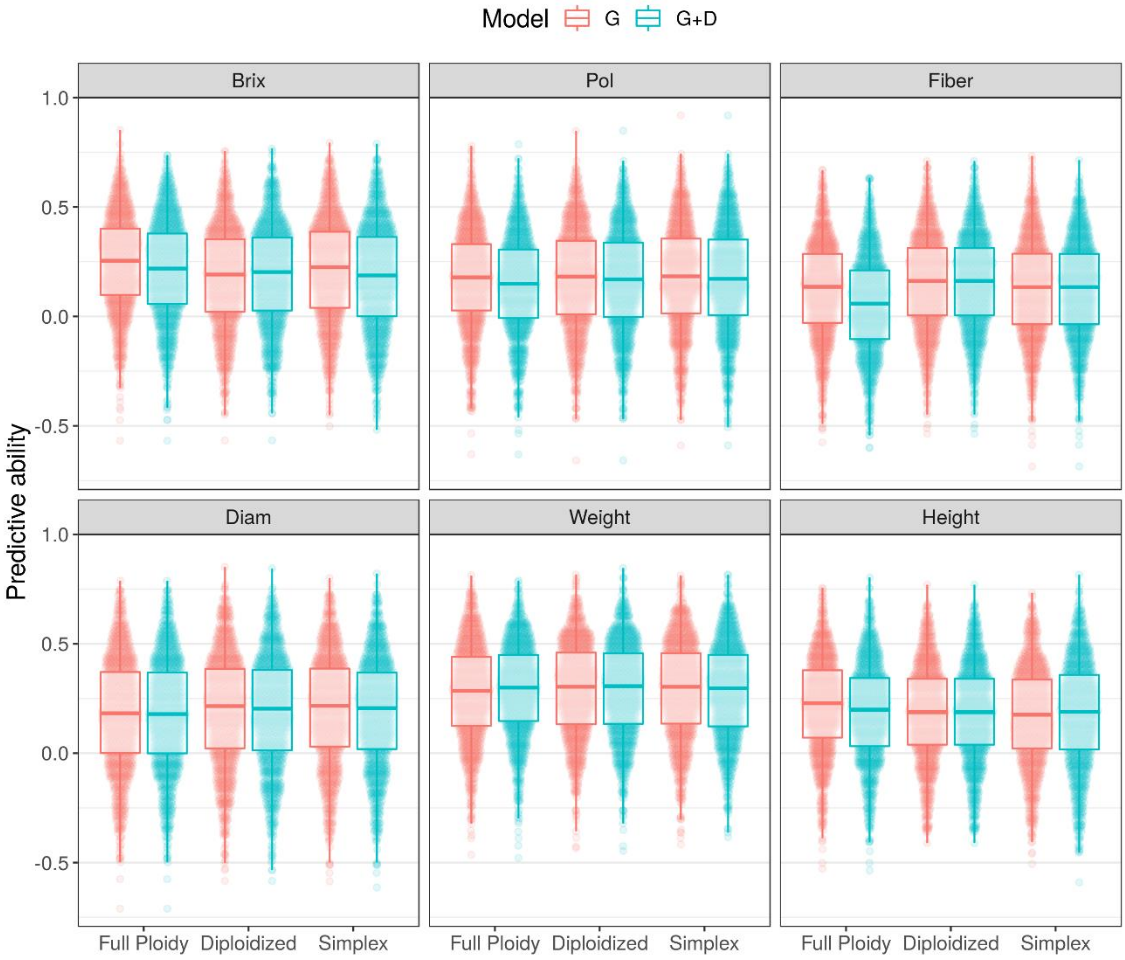

Figure 5. Distribution of the predictive ability values over different cross-validation runs of genomic selection when considering additive effects only (G) and considering additive and digenic dominance effects (G+D). Both models are compared when using markers with ploidy and allele dosage estimates (Full ploidy), diploidized markers, and using only simplex markers. The values are shown for traits soluble solids content (Brix), sucrose content (Pol), fiber percentage (Fiber), stalk diameter (Diam), stalk weight (Weight) and stalk height (Height)

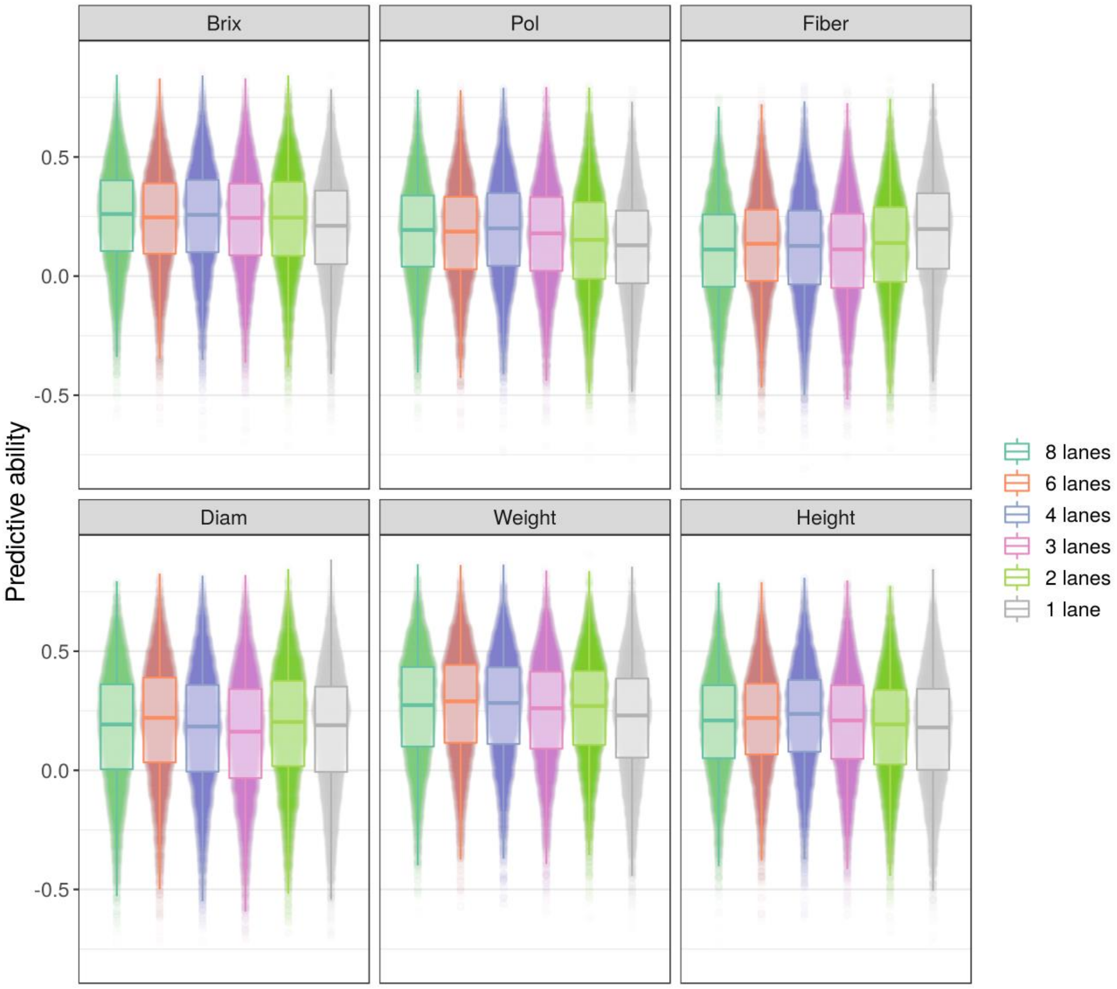

Figure 6. Distribution of the predictive ability values over different cross-validation runs of genomic selection using the SNP dataset with its original genotyping depth (8 lanes with 96x libraries) and using simulated SNP datasets with lower genotyping depths, Simulations included scenarios equivalent to the sequencing of 6 lanes, 4 lanes, 3 lanes, 2 lanes and 1 lane. The genomic selection model included additive effects only and the marker set used included ploidy and allele dosage estimates. The values are shown for traits soluble solids content (Brix), sucrose content (Pol), fiber percentage (Fiber), stalk diameter (Diam), stalk weight (Weight) and stalk height (Height).

SUPPLEMENTARY MATERIAL

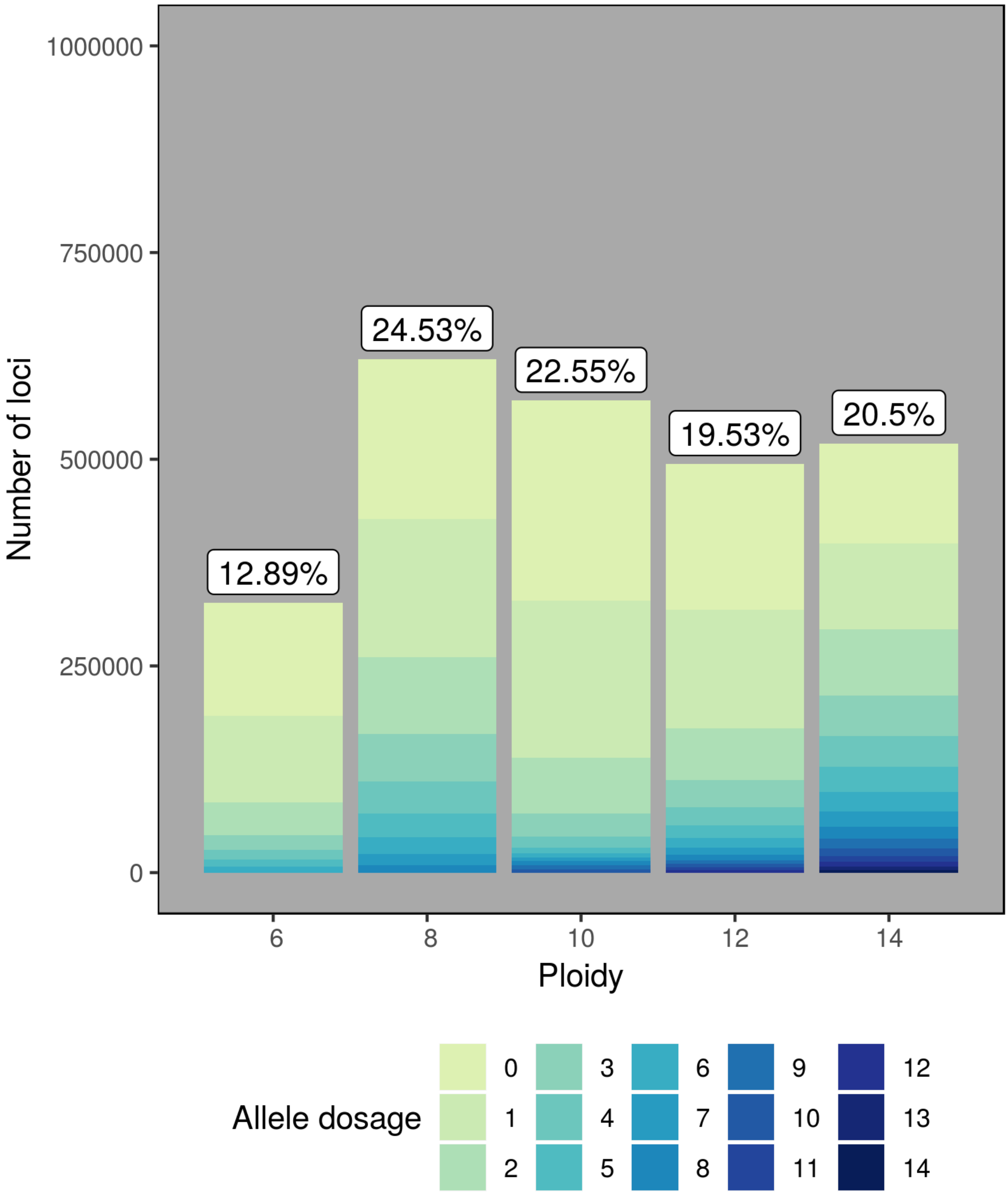

Figure S1. Summary of the estimates of ploidy and allele dosage for 170 samples and 15,466 SNPs not filtered for segregation distortion in the population. The bars show the number of SNP loci per value of ploidy, and different values of allele dosage are shown by different colours. For each ploidy value, the correspondent percentages of the total number of loci are shown above the bars.

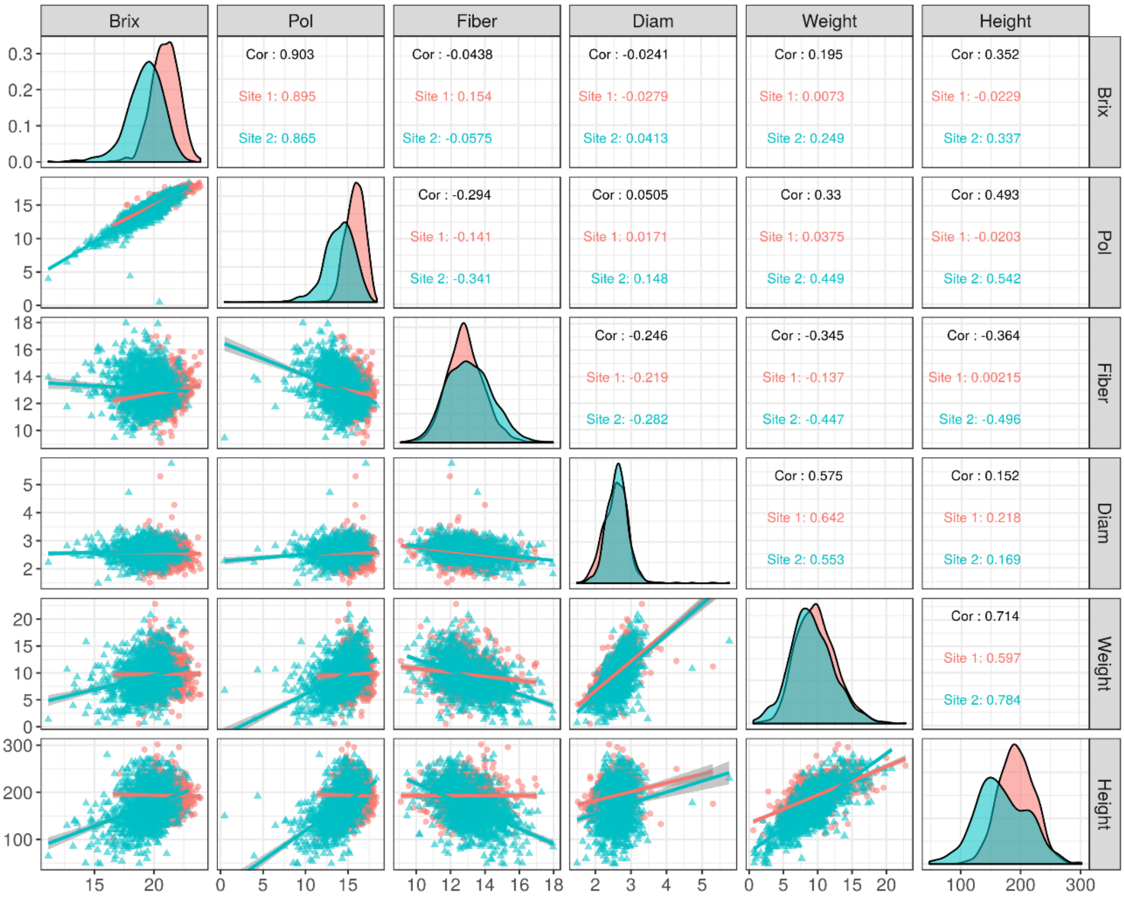

Figure S*2*. Distribution and correlation between measurements of phenotypes for traits soluble solids content (Brix), sucrose content (Pol), fiber percentage (Fiber), stalk diameter (Diam), stalk weight (Weight), and stalk height (Height) in the experimental site of Sales de Oliveira (Orange) and in the experimental site of Ribeirão Preto (Blue).