frontiers | Research Topics

Genome wide association studies and genomic selection for crop improvement in the era of big data

Edited by
Nunzio D'Agostino, Alison Bentley and Charles Chen

Published in
Frontiers in Genetics
Frontiers in Ecology and Evolution

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8897-6338-2
DOI 10.3389/978-2-8897-6338-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Genome wide association studies and genomic selection for crop improvement in the era of big data

Topic editors

Nunzio D'Agostino — University of Naples Federico II, Italy
Alison Bentley — National Institute of Agricultural Botany (NIAB), United Kingdom
Charles Chen — Oklahoma State University, United States

Citation

D'Agostino, N., Bentley, A., Chen, C., eds. (2023). *Genome wide association studies and genomic selection for crop improvement in the era of big data*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8897-6338-2

Table of contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 20 May 2022
doi: 10.3389/fgene.2022.873060

Editorial: Genome Wide Association Studies and Genomic Selection for Crop Improvement in the Era of Big Data

Alison R. Bentley[1,2]*, *Charles Chen*[3]* and *Nunzio D'Agostino*[4]*

[1]International Wheat and Maize Improvement Center (CIMMYT), El Batan, Mexico, [2]NIAB, Cambridge, United Kingdom, [3]Department of Biochemistry and Molecular Biology, 246 Noble Research Center, Oklahoma State University, Stillwater, OK, United States, [4]Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy

Keywords: allele mining, genetic diversity, genotyping, high-throughput phenotyping (HTPP), genomic estimated breeding values (GEBV), genome-to-phenome

Editorial on the Research Topic

Genome Wide Association Studies and Genomic Selection for Crop Improvement in the Era of Big Data

OPEN ACCESS

Edited and reviewed by:
Aditya Pratap,
Indian Institute of Pulses Research (ICAR), India

***Correspondence:**
Alison R. Bentley
a.bentley@cgiar.org
Charles Chen
charles.chen@okstate.edu
Nunzio D'Agostino
nunzio.dagostino@unina.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *10 February 2022*
Accepted: *04 May 2022*
Published: *20 May 2022*

Citation:
Bentley AR, Chen C and D'Agostino N (2022) Editorial: Genome Wide Association Studies and Genomic Selection for Crop Improvement in the Era of Big Data. Front. Genet. 13:873060. doi: 10.3389/fgene.2022.873060

The exploitation of the genetic diversity of crops is essential for breeding purposes, as the identification of useful/beneficial alleles for target traits within plant genetic resources allows the development of new varieties capable of responding to the challenges of global agriculture (Food and Agriculture Organization of the United Nations, 2010).

Whole genome re-sequencing, genome skimming, fractional genome sequencing strategies, and high-density genotyping arrays enable large-scale assessment of genetic diversity for a wide range of species, including major and "orphan" crops (D'Agostino and Tripodi, 2017; Rasheed et al., 2017). This is however of limited value unless associated with adaptation and functional improvement of crops. Recently, several advances in high-throughput phenotyping have overcome the "phenotyping bottleneck" (Walter et al., 2015; Pieruschka and Schurr, 2019; Song et al., 2021), making available robust phenotypic data points acquired following the precise characterization of the agronomic and physiological attributes of crops. More and more studies are taking advantage of these scientific advances and of data science techniques to uncover the genome-to-phenome relationship and unlock the breeding potential of plant genetic resources. Genome-wide association studies (GWAS) and genomic selection (GS) are powerful data science approaches to investigate marker-trait associations (MTAs) for the basic understanding of simple and complex adaptive and functional traits (Liu and Yan, 2019; Voss-Fels et al., 2019; Varshney et al., 2021). Both approaches accelerate the rate of genetic gain in crops and reduce the breeding cycle in a cost-effective manner.

For this Research Topic we sought high-quality contributions, covering various aspects of genomics-assisted-breeding: increase in yield, improvement of nutritional content and end-use quality of crops, climate-smart agriculture, cropping systems in agriculture. We did not miss to ask for contributions on technical challenges related to the design of GWAS and GS experiments and data analysis.

Enhancing knowledge on (a)biotic stress tolerance of plants has a major impact on crop improvement strategies that aim to develop high yielding varieties in suboptimal environmental conditions.

Odilbekov et al. performed GWAS on a collection of nearly 200 winter wheat accessions to identify loci associated with seedling-stage resistance to *Septoria tritici* blotch (STB) disease, which is

responsible for severe yield losses worldwide. Association tests with different statistical models returned a strong signal on chromosome 1B. Seven genes were identified as the most probable candidate genes for this QTL, as they play a key role in plant immunity and modulate the defense response. Finally, the authors demonstrated that the accuracy of the GS model for STB resistance can be improved when modeling GWAS associated variants as fixed effects.

Thapa et al. performed GWAS on a panel of 257 rice accessions to identify the QTLs and the underlying candidate genes responsible for cold tolerance and cold recovery during the germination phase. Their findings enrich the toolbox available to breeders for the development of new varieties with tolerance to low temperatures.

Hernández and Cortés subjected 78 geo-referenced wild common bean accessions to genotyping-by-sequencing (GBS) and derived three heat stress indices from phenotypic data points. Then, they applied the latest-generation GWAS models under a genome–environment association framework to identify putative loci underlying heat stress adaptation. The goal was to identify new sources of tolerance in the wild gene pool for use in breeding programs.

Increasing of crop yield potential is one of the main goals of breeding. Indeed, producing more with less is the key to feeding the growing world population. Within this motivating context, Zaïm et al., tested in the open field and in different environments four recombinant inbred line populations of durum wheat. GBS and the construction of a consensus linkage map led to the identification of over 30 QTLs for key agronomic traits. Six QTLs were found to be associated with grain yield and thousand kernel weight. The SNP markers anchored to these QTLs were then included as fixed effects into GS models, improving overall accuracy.

Sidhu et al. performed GWAS on a collection of almost 300 winter wheat accessions to determine SNP markers associated with coleoptile length. As a result, the authors identified eight candidate regions within which they found genes possibly involved in determining the target phenotype.

Many articles aimed to improve the predictive accuracy of GS models by considering some variables that influence traits or by proposing innovative technological solutions to fully exploit the genetic variability of plant genetic resources.

The article by Crossa et al. is about the comparison of the genome-based prediction accuracy of four methods: the additive genomic best linear unbiased predictor (GB), non-additive Gaussian kernel (GK), arc-cosine kernel (AK), and Deep Learning (DL). Single-environment and multi-environment G × E models on two real wheat datasets were used for benchmarking. Comparative analysis showed that AK outperformed the remaining methods, as it ensures competitive predictions at low costs in the tuning process.

Olatoye et al. identified main effect and epistatic effect loci of flowering time, maturity, and seed size in cowpea using a MAGIC population. Then, they used the identified quantitative trait nucleotides as fixed effects in parametric, semi-parametric, and non-parametric GS models and demonstrated that *a priori* knowledge of the genetic architecture of a trait improves prediction accuracy.

Allier et al. proposed adjustments to two parameters, namely the expected genetic value in the progeny (V) under a certain constraint on inbreeding (D), of the cross-selection strategy they published earlier. This arises from the need to consider within-family selection in recurrent genomic selection programs. The authors compared their UCPC-based optimal cross-selection strategy with the existing ones and proved that it was more efficient for converting genetic diversity into short- and long-term genetic gains.

Klápště et al. improved genomic predictions for traits with relatively low heritability and poor prediction accuracy by implementing multi-trait models based on the use of a marker-based relationship matrix, instead of classic pedigrees. The models applied to the diameter at breast height (target trait) did not outperform the multivariate model using all genetic markers in the case of the *Pinus radiata* population; conversely the strategy was advantageous in the case of the *Eucalyptus nitens* population, where the target trait had a low/moderate correlation with other heritable traits.

The untapped genetic variation preserved in germplasm banks serves as a source for future food and nutritional security for the globe. However, a major obstacle that prevents the use of bank accessions is the lack of adequate characterization and performance evaluation. In Kehel et al., 789 bread wheat landraces held in-trust at the gene bank of the International Center for Agricultural Research in the Dry Areas were scanned for seed traits and genetically evaluated using 12k DArTSeq SNP markers. Based on cross-validation, predicting untyped seed traits can be as accurate as 74% for seed width. Moreover, when incorporating climatic and environmental variables based on passport data, the prediction accuracy improved by an additional 8%. These findings advocate the advancement in predictive analytics and genomic technologies for identifying potential donors of desirable alleles for genetic introgression.

Considering the long reaction time and the expensive cost in conservation and sustainability of forest resources, Lstibůrek et al. conducted a multi-trait and multi-site large-scale genetic analysis with 4,625 25–35 years old European larch trees grown over 21 reforestation sites across four distinct climatic regions. In this study, the capacity of the marker-based pedigree information was demonstrated by comparing *in situ* heritability estimates. Furthermore, using this approach, a higher genetic response of the selected individuals can be expected for fitness and productivity attributes, suggesting that broad-spectrum climatic genetic evaluation can be an effective guiding principle for reforestation and genetic resource management without the reliance on structured tree breeding methods.

Finally, the last series of articles describes some data resources available for future studies or technical challenges related to the design of GWAS experiments.

Piot et al. analyzed over 1,000 *Populus trichocarpa* genomes to assess genomic diversity and identify rare and common alleles with high confidence for subsequent use in GWAS. Approximately 5% of the variants identified were non-synonymous and could represent rare defective genetic variants hypothetically associated with poplar phenotypic plasticity.

The mini-review by Srivastava et al. is quite different in content, as it provides an overview of the latest development of genetic and genomic resources in pearl millet and their use in GWAS and in the development of GS models for the estimation of GBEVs (genomic estimated breeding values).

The review by Pavan et al. provides advice on how to plan the experiments and choose the most appropriate and cost-effective genotyping method for crop GWAS. It also describes which quality control procedures should be applied on genotypic data points to avoid bias and false signals in genotype-phenotype association tests.

As genomics-driven knowledge advances rapidly and data science techniques for omics data continue to evolve and improve, the combination of the huge amount of genetic and phenotypic data points becomes more and more reachable. We looked at innovative examples whose purpose was to describe the genome-to-phenome connection and causation and to highlight the strengths and weaknesses of popular data mining strategies. We hope that the articles in this Research Topic can give further impetus to this area of research and can help expand the tools available to breeders.

AUTHOR CONTRIBUTIONS

All authors managed the peer review of manuscripts submitted to the Research Topic and contributed to manuscript writing and editing. All authors approved the final version of the editorial.

ACKNOWLEDGMENTS

The editors would like to thank all authors for their outstanding contributions and all reviewers for their valuable work, helpful comments, and suggestions. We hope this Research Topic of articles will be of interest to the plant scientific community.

REFERENCES

D'Agostino, N., and Tripodi, P. (2017). NGS-Based Genotyping, High-Throughput Phenotyping and Genome-Wide Association Studies Laid the Foundations for Next-Generation Breeding in Horticultural Crops. *Diversity* 9, 38. doi:10.3390/d9030038

Food and Agriculture Organization of the United Nations (2010). *Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome: FAO, UK distributor: Stationery Office.

Liu, H.-J., and Yan, J. (2019). Crop Genome-wide Association Study: A Harvest of Biological Relevance. *Plant J.* 97, 8–18. doi:10.1111/tpj.14139

Pieruschka, R., and Schurr, U. (2019). Plant Phenotyping: Past, Present, and Future. *Plant Phenomics* 2019, 7507131. doi:10.34133/2019/7507131

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., et al. (2017). Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Mol. Plant* 10, 1047–1064. doi:10.1016/j.molp.2017.06.008

Song, P., Wang, J., Guo, X., Yang, W., and Zhao, C. (2021). High-Throughput Phenotyping: Breaking through the Bottleneck in Future Crop Breeding. *Crop J.* 9, 633–645. doi:10.1016/j.cj.2021.03.015

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M. E. (2021). Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci.* 26, 631–649. doi:10.1016/j.tplants.2021.03.010

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating Crop Genetic Gains with Genomic Selection. *Theor. Appl. Genet.* 132, 669–686. doi:10.1007/s00122-018-3270-8

Walter, A., Liebisch, F., and Hund, A. (2015). Plant Phenotyping: From Bean Weighing to Image Analysis. *Plant Methods* 11, 14. doi:10.1186/s13007-015-0056-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bentley, Chen and D'Agostino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 July 2019
doi: 10.3389/fgene.2019.00677

Epistasis Detection and Modeling for Genomic Selection in Cowpea (*Vigna unguiculata* L. Walp.)

*Marcus O. Olatoye[1], Zhenbin Hu[2] and Peter O. Aikpokpodion[3]**

[1] Department of Crop Sciences, University of Illinois, Urbana-Champaign, IL, United States, [2] Department of Agronomy, Kansas State University, Manhattan, KS, United States, [3] Department of Genetics and Biotechnology, University of Calabar, Calabar, Nigeria

Genetic architecture reflects the pattern of effects and interaction of genes underlying phenotypic variation. Most mapping and breeding approaches generally consider the additive part of variation but offer limited knowledge on the benefits of epistasis which explains in part the variation observed in traits. In this study, the cowpea multiparent advanced generation inter-cross (MAGIC) population was used to characterize the epistatic genetic architecture of flowering time, maturity, and seed size. In addition, consideration for epistatic genetic architecture in genomic-enabled breeding (GEB) was investigated using parametric, semi-parametric, and non-parametric genomic selection (GS) models. Our results showed that large and moderate effect–sized two-way epistatic interactions underlie the traits examined. Flowering time QTL colocalized with cowpea putative orthologs of *Arabidopsis thaliana* and *Glycine max* genes like *PHYTOCLOCK1* (*PCL1* [Vigun11g157600]) and *PHYTOCHROME A* (*PHY A* [Vigun01g205500]). Flowering time adaptation to long and short photoperiod was found to be controlled by distinct and common main and epistatic loci. Parametric and semi-parametric GS models outperformed non-parametric GS model, while using known quantitative trait nucleotide(s) (QTNs) as fixed effects improved prediction accuracy when traits were controlled by large effect loci. In general, our study demonstrated that prior understanding of the genetic architecture of a trait can help make informed decisions in GEB.

Keywords: cowpea, genetic architecture, epistasis, QTL, genomic-enabled breeding, genomic selection, flowering time, photoperiod

OPEN ACCESS

Edited by:
Charles Chen, Oklahoma State University, United States

Reviewed by:
Eric Von Wettberg, University of Vermont, United States
Inês Fragata, University of Lisbon, Portugal

****Correspondence:***
Peter O. Aikpokpodion
paikpokpodion@unical.edu.ng
paikpokpodion@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *22 March 2019*
Accepted: *27 June 2019*
Published: *30 July 2019*

Citation:
Olatoye MO, Hu Z and Aikpokpodion PO (2019) Epistasis Detection and Modeling for Genomic Selection in Cowpea (Vigna unguiculata L. Walp.). Front. Genet. 10:677. doi: 10.3389/fgene.2019.00677

INTRODUCTION

Asymmetric transgressive variation in quantitative traits is usually controlled by non-additive gene interaction known as epistasis (Rieseberg et al., 1999). Epistasis has been defined as the interaction of alleles at multiple loci (Mathew et al., 2018). The joint effect of alleles at these loci may be lower or higher than the total effects of the loci (Johnson, 2008). In selfing species, epistasis is common due to high level of homozygosity (Volis et al., 2010) and epistatic interactions have been found among loci underlying flowering time in barley (Mathew et al., 2018), rice (Chen et al., 2015; Chen et al., 2018b), and sorghum (Li et al., 2018a). Although, theoretical models and empirical studies involving simulations have suggested the significant role for epistasis in breeding (Melchinger et al., 2007; Volis et al., 2010; Messina et al., 2011; Howard et al., 2014), empirical evidence from practical breeding are limited. In addition, most of the current statistical models cannot efficiently characterize or

account for epistasis (Mackay, 2001; Moore and Williams, 2009; Sun et al., 2012; Mathew et al., 2018). Common quantitative traits mapping approaches are often single-locus analysis techniques. These techniques focus on the additive contribution of genomic loci (Barton and Keightley, 2002), which only explains a fraction of the genetic variation which can lead to missing heritability.

Regardless of the limitations of genomic mapping approaches, characterization of the genetic basis of complex agronomic traits has been beneficial for breeding purposes. For example, markers tagging quantitative trait loci (QTL) have been used in marker-assisted selection (MAS) in breeding programs (Zhang et al., 2003; Pan et al., 2006; Saghai Maroof et al., 2008; Foolad and Panthee, 2012; Massman et al., 2013; Mohamed et al., 2014; Zhao et al., 2014). However, the efficiency of QTL-based MAS approach in breeding is limited. First, the small sample size of bi-parental populations where QTL is detected often results in overestimation of the respective QTL effect sizes, a phenomenon known as Beavis effect (Utz et al., 2000; Xu, 2003; King and Long, 2017). Second, linkage mapping is limited in power to detect small effect loci; thus, only the available large effect loci are used for MAS (Ben-Ari and Lavi, 2012). Third, genetic diversity is limited to the two parents forming the bi-parental population; thus, QTL may not reflect the entire variation responsible for the trait and may not be transferable to other genetic backgrounds (Xu et al., 2017). Multi-parental populations as nested association mapping (NAM) and multiple advanced generation intercross (MAGIC) offer increased power, resolution, reliable estimate of QTL effects, and increased diversity than bi-parentals. Additionally, the MAGIC mapping population presents greater genetic diversity than bi-parentals to identify higher-order epistatic interactions (Mathew et al., 2018).

Notably, MAS is more efficient with traits controlled by few genomic loci than polygenic traits (Bernardo, 2008). In contrast, genomic selection (GS) that employs genome wide markers has been found to be more suited for complex traits, and also having higher response to selection than MAS (Bernardo and Yu, 2007; Wong and Bernardo, 2008; Cerrudo et al., 2018). In GS, a set of genotyped and phenotyped individuals are first used to train a model that estimates the genomic estimated breeding values (GEBVs) of un-phenotyped but genotyped individuals (Jannink et al., 2010). GS models often vary in performance with the genetic architecture of traits. Parametric GS models are known to capture additive genetic effects but are not efficient with epistatic effects due to the computational burden of high-order interactions (Moore and Williams, 2009; Howard et al., 2014). Parametric GS models with incorporated kernels (marker based relationship matrix) for epistasis have recently been developed (Covarrubias-Pazaran, 2016). Semi-parametric and non-parametric GS models capturing epistatic interactions have been developed and implemented in plant breeding (Gianola et al., 2006; Gianola and de los Campos, 2008; De Los Campos et al., 2010). Semi-parametric models as reproducing Kernel Hilbert space (RKHS) reduces parametric space dimensions to efficiently capture epistatic interactions among markers (Jiang and Reif, 2015; de Oliveira Couto et al., 2017). Using simulated data, Howard et al. (2014) showed that semi-parametric and non-parametric GS models can improve prediction accuracies under epistatic genetic architectures. In summary, different models may fit different genetic architectures. In general, GS has been widely studied and applied to major crop species including both cereals and legumes while its applications in orphan crop species has gained increased attention in recent times.

Cowpea (*Vigna unguiculata* L. Walp) is a widely adapted warm-season orphan herbaceous leguminous annual crop and an important source of protein in developing countries (Muchero et al., 2009; Varshney et al., 2012; Boukar et al., 2018; Huynh et al., 2018). Due to its flexibility as a "hungry season crop" (Langyintuo et al., 2003), cowpea is part of the rural families' coping strategies to mitigate the effect of changing climatic conditions. Cowpea's nitrogen fixing and drought tolerance capabilities make it a valuable crop for low-input and smallholder farming systems (Hall et al., 2003; Boukar et al., 2018). Breeding efforts using classical approaches have been made to improve cowpea's tolerance to both biotic (disease and pest) and abiotic (drought and heat) stressors (Hall et al., 2003; Hall, 2004). Advances in applications of next-generation sequencing (NGS) and development of genomic resources (consensus map, draft genome, and multi-parent population) in cowpea have provided the opportunity for the exploration for GEB (Muchero et al., 2009; Boukar et al., 2018; Huynh et al., 2018). MAS and GS have improved genetic gain in soybean (*Glycine max*) (Jarquin et al., 2016; Kurek, 2018; Matei et al., 2018), common bean (*Phaseolus vulgaris*) (Schneider et al., 1997; Yu et al., 2000; Wen et al., 2019), chickpea (Roorkiwal et al., 2016; Li et al., 2018b), pigeonpea (Varshney et al., 2010; Pazhamala et al., 2015), and lentil (Haile et al., 2019). However, cowpea still lags behind major legumes in the area of GEB applications. GEB has the potential to expedite cowpea breeding to ensure food security in developing countries where national breeding programs still depend on labor-intensive and time-consuming classical breeding approaches.

In this study, we used the cowpea MAGIC population to first characterize the genetic architecture (main effect and epistatic effect loci) of flowering time, maturity, and seed size, and second, to evaluate considerations for genetic architecture in genomic-enabled breeding using parametric, semi-parametric, and non-parametric GS models and MAS. Our results showed that flowering time and maturity under short day are both controlled by moderate effect loci, while flowering time under long day and seed size are controlled by large and moderate effect loci. Also, accounting for large effect loci as fixed effects in parametric GS model improved prediction accuracy.

EXPERIMENTAL PROCEDURES

Plant Genetic Resource and Phenotypic Evaluation

This study was performed using publicly available cowpea MAGIC population's phenotypic and genotypic data (Huynh et al., 2018). The MAGIC population was derived from an intercross between eight founders. The F_1s were derived from eight-way intercross between the founders and were subsequently selfed through single-seed descent for eight generations. The F_8 RILs were later genotyped with 51,128 SNPs using the Illumina

Cowpea Consortium Array. A core set of 305 MAGIC RILs were selected and phenotyped (Huynh et al., 2018). The RILs were evaluated under two irrigation regimes.

In this study, the flowering time (FLT), maturity (MAT), and seed size (SS) data were analyzed for environment-by-environment correlations and best linear unbiased predictions (BLUPs). The traits analyzed in this study are: FTFILD (FLT under full irrigation and long day), FTRILD (FLT under restricted irrigation and long day), FTFISD (FLT under full irrigation and short day), FTRISD (FLT under restricted irrigation and short day), FLT_BLUP (BLUP of FLT across environments), MFISD (MAT under full irrigation and short day), MRISD (MAT under restricted irrigation and short day), MAT_BLUP (BLUP of MAT across environments), SSFISD (SS under full irrigation and short day), SSRISD (SS under restricted irrigation and short day), and SS_BLUP(BLUP of SS across environments). In addition, using both genomic and phenotypic data, narrow sense heritability was estimated using *rrBLUP* package in R (Endelman, 2011).

QTL and Epistasis Mapping

QTL mapping was performed for all traits using the stepwise regression model implemented in TASSEL 5.0 standalone version (Bradbury et al., 2007). The approach implements both forward inclusion and backward elimination steps. The model accounts for major effect loci and reduces collinearity among markers. The model was designed for multi-parental populations, and no family term was used in the model since MAGIC population development involved several steps of intercross that reshuffles the genome and minimizes phenotype-genotype covariance. A total of 32,130 SNPs across 305 RILs were used in the analysis. A permutation of 1,000 was used in the analysis.

To characterize the epistatic genetic architecture underlying FLT, MAT, and SS, the Stepwise Procedure for constructing an Additive and Epistatic Multi-Locus model (SPAEML; Chen et al., 2018a) epistasis pipeline implemented in TASSEL 5.0 was used to perform epistasis mapping for phenotypic traits (FTFILD, FTRILD, FTFISD, FTRISD, FT_BLUP, MFISD, MRISD, MT_BLUP, SSFISD, SSRISD, and SS_BLUP). One critical advantage of SPAEML that led to its consideration for this study is its ability to correctly distinguish between additive and epistatic loci. SPAEML source code is available at https://bitbucket.org/wdmetcalf/tassel-5-threaded-model-fitter. The minor allele frequency of each marker was estimated using a custom R script from http://evachan.org/rscripts.html. The additive effect of the marker was estimated as the difference between the mean phenotypic value of two homozygous classes of the alleles of a marker divided by two. The proportion of phenotypic variation explained (PVE) by each marker was estimated by multiplying the R^2 obtained from fitting a regression between the marker and the trait of interest by 100. The regression model for estimating PVE is:

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij} \quad [1]$$

where y_{ij} is the phenotype, μ is the overall mean, γ_i is the term for associated marker/SNP, and ε_{ij} is the residual term. This was implemented using the *lm* function in R.

A set of *a priori* genes ($n = 100$; **Data S1**) was put together from *Arabidopsis thaliana* and *G. max* FLT and SS genes obtained from literature and https://www.mpipz.mpg.de/14637/Arabidopsis_flowering_genes. The cowpea orthologs of these genes were obtained by blasting the *A. thaliana* and *G. max* sequence of the *a priori* genes on the new *Vigna* genome assembly *v.1* on Phytozome (Goodstein et al., 2012). The corresponding cowpea gene with the highest score was selected as a putative ortholog. Colocalizations between the cowpea putative orthologs and associated markers were identified using a custom R script. Only significant marker and *a priori* genes at the same genetic position were reported.

Marker-Assisted Selection Pipeline

In order to evaluate the performance of MAS in cowpea, a custom pipeline was developed in R. Using subbagging approach, 80% of the 305 RILs randomly sampled without replacement was used as the training population, followed by performing a multi-locus GWAS (multi-locus mixed model, MLMM) (Segura et al., 2012) on both genomic and phenotypic data of the training population. The MLMM approach implements stepwise regression involving both forward and backward regressions. This model accounts for major effect loci and reduces the effect of allelic heterogeneity. A K-only model that accounts for a random polygenic term (kinship relationship matrix) was used in the MLMM model. No term for population structure was used in the model since MAGIC population development involved several steps of intercross that reshuffles the genome and minimizes phenotype-genotype covariance. A total of 32,130 SNPs across 305 RILs were used in the GWAS analysis and coded as −1 and 1 for homozygous markers/SNPs and 0 for heterozygous SNPs. Bonferroni correction with $\alpha = 0.05$ was used to determine the cut-off threshold for each trait association (α/total number of markers = 1.6 e-06).

$$y = X\beta + S\propto + Zu + e \quad [2]$$

where y is the vector of phenotypic data, β is a vector of fixed effects other than SNPs, $\propto$ is the vector of SNP effects, u is a vector of polygenic background effects, and e is the vector of residual effects. X, S, and Z are incident matrices of 1s and 0s relating y to β, $\propto$, and u (Yu et al., 2006).

Afterwards, the top three most significant associations were then selected from the genomic data of the training population to train a regression model by fitting the SNPs as predictors in a regression model with the phenotypic information as the response variable. This training model was later used alongside the *predict* function in R to predict the phenotypic information of the validation population (20% that remained after sub-setting the training population). The prediction accuracy of MAS was obtained as the correlation between this predicted phenotypic information and the observed phenotypic information for the validation data.

Genomic Selection Pipeline

In order to evaluate the performance of using known marker/SNP as fixed effects in GS models and to compare

the performance of parametric, semi-parametric, and non-parametric GS models, a custom GS pipeline was developed in R. The GS pipeline was made up of four GS models, which were named as FxRRBLUP (ridge regression BLUP where markers were fitted as both fixed and random effects; parametric), RRBLUP (RRBLUP where markers were only fitted as random effects; parametric), reproducing Kernel Hilbert space (RKHS; semi-parametric), and support vector regression (SVR; non-parametric). First, using subagging approach, 80% of the RILs were randomly sampled without replacement (training population) followed by running MLMM GWAS and selecting the three most significant associations, which were used as fixed effects in the FxRRBLUP. These three SNPs were removed from the rest of SNPs that were fitted as random effects in the FxRRBLUP model. Using a high number of SNPs as fixed effects have been found to increase bias (Rice and Lipka, 2019), as a result, three QTNs were fitted as fixed effects. The RRBLUP, RKHS, and SVR models were fitted simultaneously in the same cycle as FxRRBLUP to ensure unbiased comparison of GS models. Likewise, in order to ensure unbiased comparison between GS and MAS approaches, similar seed numbers were used for the subagging sampling of training populations across 100 cycles for GS and MAS. The validation set was composed of the remaining 20% of the RILs after sampling the 80% (training set). Prediction accuracy in GS was estimated as the Pearson correlation between measured phenotype and GEBVs of the validation population. Also, for FLT, each environment was used as a training population to predict the other three environments.

Ridge Regression BLUP (RRBLUP)

The two RRBLUP models (with and without fixed-effect term) can be described as;

$$\boldsymbol{y} = \mu + \sum_{m=1}^{p} \boldsymbol{Z}_m u_m + \boldsymbol{e} \quad [3]$$

$$\boldsymbol{y} = \mu + \sum_{k=1}^{q} \boldsymbol{X}_k \alpha_k + \sum_{m=1}^{p} \boldsymbol{Z}_m u_m + \boldsymbol{e} \quad [4]$$

where **y** is the vector (n x 1) of observations (phenotypic data), μ is the vector of the general mean, q is the number of selected significant associated markers (q = 3), $\boldsymbol{X_k}$ is the k^{th} column of the design matrix $\boldsymbol{X}$, α is the fixed additive effect associated with markers $k \ldots q$, u random effects term, with $E(u_m) = 0$, $Var(u_m) = \sigma^2_{u_m}$ (variance of marker effect), p is the marker number ($p > n$), $\boldsymbol{Z}_m$ is the m^{th} column of the design matrix $\boldsymbol{Z}$, and u is the vector of random marker effects associated with markers $m \ldots p$. In the model, u random effects term, with $E(u_m) = 0$, $Var(u_m) = \sigma^2_{u_m}$ (variance of marker effect), $Var(\mathbf{e}) = \sigma^2$ (residual variance), $Cov(\mathrm{u}, \mathbf{e}) = 0$, and the ridge parameter λ equals σ^2_e / σ^2_u (Meuwissen et al., 2001; Endelman, 2011; Howard et al., 2014). In this study, RRBLUP with and without fixed effects were implemented using the *mixed.solve* function in *rrBLUP* R package (Endelman, 2011).

Reproducing Kernel Hilbert Space (RKHS)

Semi-parametric models are known to capture interactions among loci. The semi-parametric GS approach used in this study was implemented as Bayesian RKHS in *BLGR* package in R (Perez, 2014), and described as follows:

$$\boldsymbol{y} = 1\mu + \boldsymbol{u} + \varepsilon \quad [5]$$

where $\boldsymbol{y}$ is the vector of phenotype, **1** is a vector of 1's, μ is the mean, $\boldsymbol{u}$ is vector of random effects ~MVN ($\mathbf{0}$, $\mathbf{K}_h\sigma^2_u$), and ε is the random residual vector ~ MVN ($\mathbf{0}, \boldsymbol{I}\sigma^2_\varepsilon$). In Bayesian RKHS, the priors $p(\mu, \boldsymbol{u}, \varepsilon)$ are proportional to the product of density functions MVN ($\mathbf{0}$, $\mathbf{K}_h\sigma^2_u$) and MVN ($\mathbf{0}$, $\boldsymbol{I}\sigma^2_\varepsilon$). The kernel entries matrix ($\mathbf{K}_h$) with a Gaussian kernel uses the squared Euclidean distance between marker genotypes to estimate the degree of relatedness between individuals, and a smoothing parameter (h) multiplies each entry in $\mathbf{K}_h$ by a constant. In the implementation of RKHS, a default smoothing parameter h of 0.5 was used alongside 1,000 burns and 2,500 iterations.

Support Vector Regression (SVR)

Support vector regression method (Vapnik, 1995; Maenhout et al., 2007; Long et al., 2011) was used to implement non-parametric GS approach in this study. The aim of the SVR method is to minimize prediction error by implementing models that minimizes large residuals (Long et al., 2011). Thus, it is also referred to as the "ε-intensive" method. It was implemented in this study using the normal radial function kernel (*rbfdot*) in the *ksvm* function of *kernlab* R package (Karatzoglou et al., 2004).

Parameters Evaluated in GS and MAS

Additional parameters were estimated to further evaluate the performance of GS and MAS models. A regression model was fitted between observed phenotypic information and GEBV of the validation set to obtain both intercept and slope for both GS and MAS in each cycle of prediction. The estimates of the intercept and slope of the regression of the observed phenotypic information on GEBVs are valuable since their deviations from expected values can provide insight into deficiencies in the GS and MAS models (Daetwyler et al., 2013). The bias estimate (slope and intercept) signifies how the range of values in measured and predicted traits differ from each other. In addition, the coincidence index between the observed and GEBVs for both GS and MAS models was evaluated. The coincidence index (Fernandes et al., 2018) evaluates the proportion of individuals with highest trait values (20%) that overlapped between the measured phenotypes and predicted phenotypic trait values for the validation population.

RESULTS

Phenotypic and Genotypic Variation in Cowpea

Results showed variation between number of days to 50% flowering under long-day photoperiod and short-day

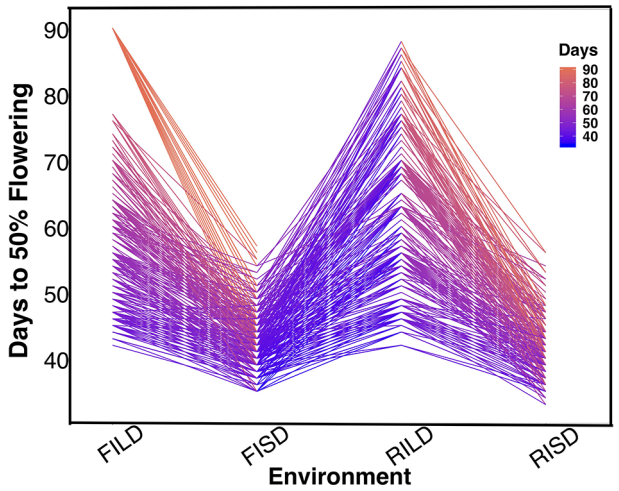


FIGURE 1 | The reaction norm plot for flowering time variation under long-day and short-day periods. Evaluation environments are represented on the *x*-axis (full irrigation and long day [FILD], full irrigation and short day [FISD], restricted irrigation and long day [RILD], and restricted irrigation and short day [RISD]). The number of days to 50% flowering is represented on the *y*-axis.

photoperiod. Days to FLT were higher for RILs under long day than short day (**Figure 1**). Results showed positive high correlations between environments for each trait (**Tables S1** and **S2**). Furthermore, genomic heritability were moderate for the traits ranging between 0.41 under long-day photoperiod to 0.48 for FLT under short-day photoperiod, 0.21 under restricted irrigation to 0.30 under full irrigation for MAT, and 0.39 under restricted irrigation to 0.47 under full irrigation for SS (**Tables S1** and **S2**).

Genetic Architecture of Traits

Main Effect QTL

The cowpea multi-parental advanced generation intercross (MAGIC) population facilitated the characterization of the genetic architecture of FLT, MAT, and SS. In this study, QTL associated with FLT, MAT, and SS were identified using stepwise regression analysis (**Table S3**, **Data S2**). Results showed that 32 QTL (22 unique) in total were associated with FLT traits (FT_BLUP [eight QTLs, explaining 73.2% of phenotypic variation (PV)], FTFILD [five QTL, explaining 66.2% of PV], FTRILD [five QTL explaining 48.6% of PV], FTFISD [eight QTL explaining 52.1% of PV], and FTRISD [six QTL explaining 43.9% of PV]). Each of the total QTL associated with FLT traits explained between 2 and 28% of the phenotypic variation. QTL qVu9:23.36, qVu9:24.77, and qVu9:22.65 (MAF = 0.29, 0.28, and 0.49) explained the largest proportion of variation (28%, 24%, and 19%) with additive effects of 7, 7, and 6 days, respectively. The minor allele frequency (MAF) of the FLT QTL ranges from 0.13 to 0.50. For MAT traits, 13 QTL (11 unique QTL) in total were identified with five QTL (explaining 35.9% of PV) for MAT_BLUP, four QTL (explaining 24.5% of PV) for MFISD, and four QTL (explaining 27.9% of PV) for MRISD. All MAT trait QTL explained between 4.5 to 10% of phenotypic variation and MAF ranges from 0.15 to 0.49.

Furthermore, for SS traits, 10 QTL (seven unique QTL) in total were identified with three QTL (explaining 39.3% of PV) for SS_BLUP, three QTL (explaining 41% of PV) for SSFISD, and four QTL (explaining 39.4% of PV) for SSRISD. QTL qVu8:74.21, qVu8:74.29, and qVu8:76.81 associated with SSFISD, SS_BLUP, and SSRISD explained the largest PV (29%, 25%, and 20%). All SS trait QTL explained between 3 and 29% of PV and with MAF range between 0.21 and 0.49. A pleiotropic QTL qVu8:74.21 (MAF = 0.24) was associated with both MRISD and SSRISD (explained 5% and 29% of PV, respectively). In summary, QTL effects range from small to large for all traits in this study (**Figure 2**).

Two-Way Epistatic Interaction QTL

Currently, there is limited knowledge about what role epistasis plays in phenotypic variation in cowpea. Our results identified epistatic loci underlying FLT, MAT, and SS (**Table S4**, **Data S3**). For FLT traits, there were 42 two-way epistatic interactions at 84 epistatic loci (only 65 loci were unique). Among these are; 20 epistatic loci for FLT_BLUP, 18 epistatic or FTFILD, 12 epistatic loci for FTRILD, 14 epistatic loci for FTFISD, and 20 epistatic loci for FTRISD. Some large effect loci were involved in epistatic interactions in FLT; examples include, QTL qVu9:25.39 (MAF = 0.28, FT_BLUP PVE = 23.5%, FTFILD PVE = 24.5%, FTRILD PVE = 26%) and QTL qVu9:3.46 (MAF = 0.35, FLT_BLUP PVE = 13.5%, FTRILD PVE = 14.1%). For MAT, there were 17 pairwise epistatic interactions across 34 loci (of which 30 were unique). Among the MAT QTL, qVu9:8.37 had the largest effect explaining ~9% of the phenotypic variation. One epistatic interaction overlapped with both FTRISD, MRISD, and MT_BLUP (qVu2:48.05+ qVu9:8.37, MAF = 0.30, and 0.39, respectively). For SS, there were 13 interactions at 26 loci (19 were unique). Only one QTL (qVu8:74.29, MAF = 0.25) had interactions with multiple QTL. The largest effect epistatic QTL associated with the three SS traits (SS_BLUP, SSFISD, and SSRISD) is qVu8:74.29 (MAF0.25). Some QTL were found to overlap among main effect QTL and epistatic effect QTL for FLT (nine QTL), MAT (three QTL), and SS (three QTL) (**Figure S1**).

Main Effect and Epistatic QTL Colocalized with *A priori* Genes

Gene functions can be conserved across species (Huang et al., 2017). In this study, a set of *a priori* genes was compiled from both *A. thaliana* and *G. max*. Both main effect QTL and epistatic QTL colocalized with putative cowpea orthologs of *A. thaliana* and *G. max* FLT and SS genes (**Figures 3–6**, **Figures S2–S11**, **Data S4**) at the same genetic position. However, two genes (*TOE2* and *AHK2*) did not colocalize with the QTL at the same genetic position but were reported due to their proximity and biological relevance. A putative cowpea ortholog (Vigun09g050600) of *A. thaliana* circadian clock gene *phytochrome E* (*PHYE*; AT4G18130) (Aukerman and Sakai, 2003) colocalized with FTFILD QTL (qVu9:22.65; PVE = 19.5%; main effect QTL) at the same genetic position. Also, a putative cowpea ortholog (Vigun07g241700) of *A. thaliana* circadian clock gene *TIME FOR COFFEE (TIC*; AT3G22380) (Hall et al., 2003) colocalized at the same genetic position with FTFISD QTL (qVu7:86.92; PVE = 2.6%; main

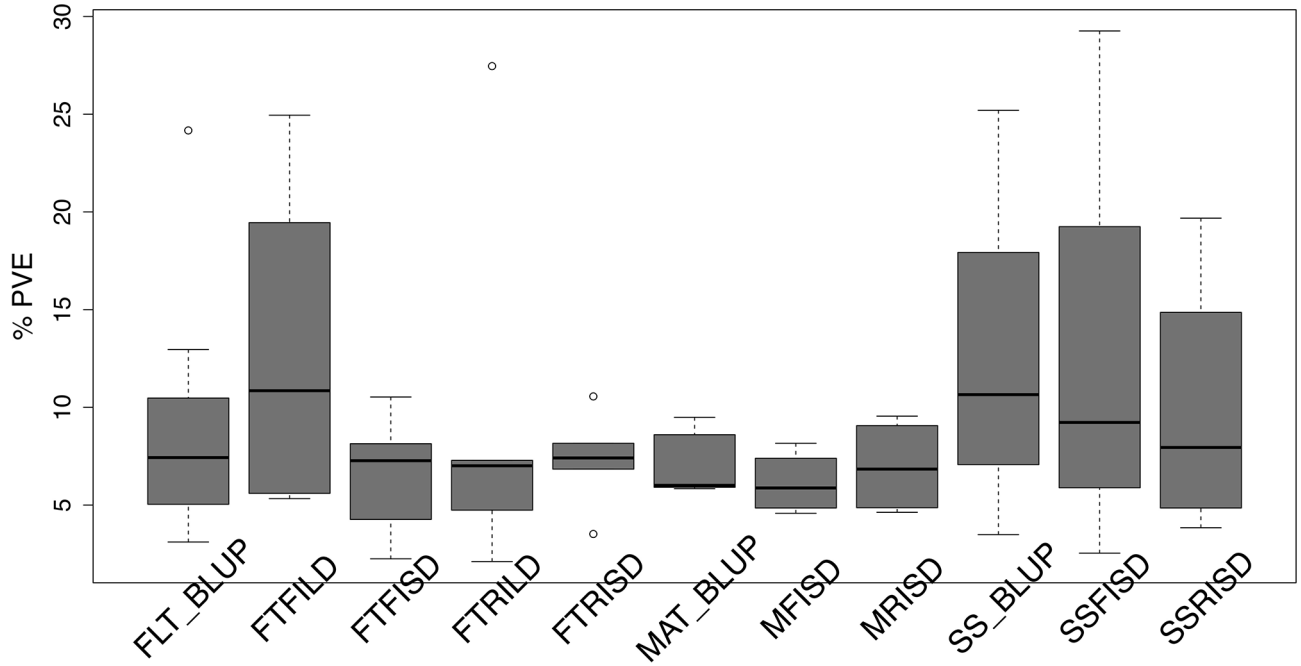


FIGURE 2 | Distribution of effect size of quantitative traits loci (QTL) associated with traits in the cowpea MAGIC population. Box plots of the distribution of proportion of variation explained (PVE) by quantitative traits loci (QTL) associated with best linear unbiased predictions (BLUP) of flowering time across environments (FLT_BLUP), flowering time under full irrigation and long day (FTFILD), flowering time under restricted irrigation and long day (FTRILD), flowering time under full irrigation and short day (FTFISD), flowering time under restricted irrigation and short day (FTRISD), BLUP of maturity across environments (MAT_BLUP) maturity under full irrigation and short day (MFISD), maturity under restricted irrigation and short day (MRISD), BLUP of seed size across environments (SS_BLUP) seed size under full irrigation and short day (SSFISD), and seed size under restricted irrigation and short day (SSRISD).

effect QTL). The cowpea FLT gene (*VuFT*; Vigun06g014600; CowpeaMine *v*.06) colocalized with an epistatic QTL (qVu6:0.68; PVE = 3.5%) associated with FLT_BLUP and FTRILD at the same genetic position. The cowpea ortholog (Vigun11g157600) of *A. thaliana* circadian clock gene *PHYTOCLOCK1* (*PCL1*; AT3G46640) (Hazen et al., 2005) colocalized with an epistatic QTL (qVu11:50.94; PVE = 8–10%) associated with both FTFILD and FTRILD at the same genetic position.

A putative cowpea ortholog (Vigun11g148700) of *A. thaliana* photoperiod gene *TARGET OF EAT2* (*TOE2*; AT5G60120) (Mathieu et al., 2009) was found at a proximity of 0.6cM from a QTL (qVu11:49.06; PVE = 7–11%; main effect QTL) associated with FTFILD, FTRILD, and FLT_BLUP. Some of the *a priori* genes colocalized with some QTL that are both main effect and epistatic QTL. For instance, the cowpea ortholog (Vigun01g205500) of *G. max* FLT gene *phytochrome A* (*PHYA*; Glyma19g41210) (Tardivel et al., 2014) colocalized with a FTFILD QTL (qVu1:66.57; PVE = 5.3%; both main effect and epistatic QTL) at the same genetic position (Data S4). Lastly, a putative cowpea ortholog (Vigun08g217000) of *A. thaliana histidine kinase2* gene (*AHK2*; AT5G35750) (Orozco-Arroyo et al., 2015) was found at a proximity of about 1–2cM from three QTL (qVu8:74.29, qVu8:74.21, qVu8:76.81; PVE = 25%, 29.3%, and 20%, respectively; main effect and epistatic QTL) associated with SS traits SS_BLUP, SSFISD, and SSRISD). In addition, some *a priori* genes were associated with multiple traits. The putative cowpea ortholog (Vigun05g024400) of *A. thaliana* circadian clock gene *CONSTANS* (*CO*; AT5G15840) (Wenkel et al., 2006) colocalized at the same genetic position with a QTL (qVu5:8.5; PVE = 6–8%; both main effect and epistatic QTL) associated with FLT and MAT traits (FLT_BLUP, FTFISD, FTRILD, FTRISD, MAT_BLUP, and MFISD). The putative cowpea ortholog (Vigun09g025800) of *A. thaliana* circadian clock gene *ZEITLUPE* (*ZTL*; AT5G57360) (Somers et al., 2000) colocalized at the same genetic position with a QTL (qVu9:8.37; PVE = 9–11%; both main effect and epistatic QTL) associated with FLT and MAT traits (FTFISD, FTRISD, and MRISD).

GS and MAS for Flowering Time

Prior knowledge about the genetic architecture of a trait can help make informed decisions in breeding. Comparing the performance of GS and MAS models for FLT within each daylength results showed that, under long day length (FTFILD and FTRILD), FxRRBLUP (mean prediction accuracy [mPA] = 0.68, 0.68; mean coincidence index [mCI] = 0.49, 0.40) and MAS (mPA = 0.64, 0.61; mCI = 0.45, 0.37) outperformed RRBLUP (mPA = 0.55, 0.58; mCI = 0.37, 0.35), RKHS (mPA = 0.55, 0.58; mCI = 0.37, 0.36), and SVR (mPA = 0.54, 0.50; mCI = 0.35, 0.28) (**Figures 7** and **8**, **Tables S3** and **4**). For FLT under long day, coincidence index values were higher under full irrigation than under restricted irrigation. For FLT under short day (FTFISD and FTRISD), all GS models outperformed MAS (mPA = 0.33, 0.25; mCI = 0.30, 0.26). Among the GS models, RKHS and RRBLUP had the highest prediction accuracies. However, the coincidence

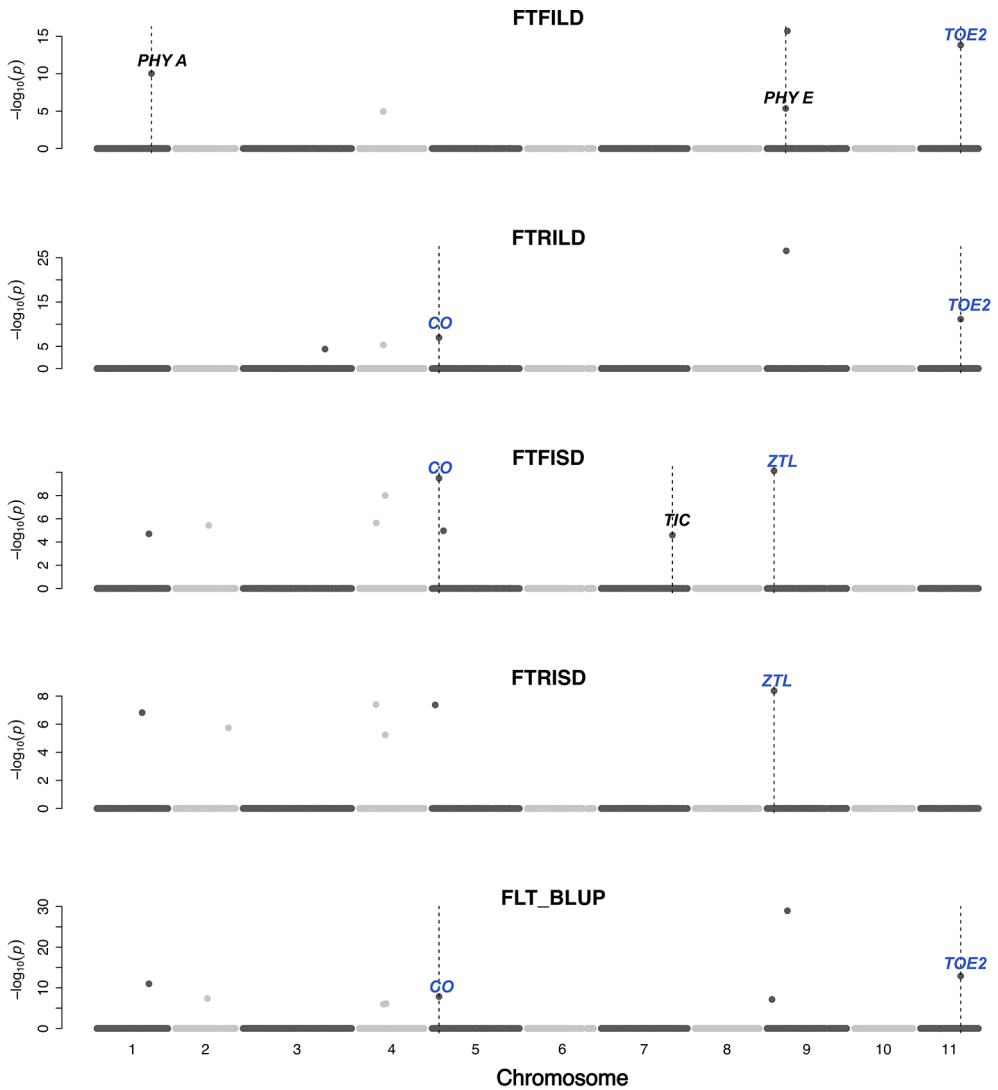


FIGURE 3 | Main QTL plot for flowering time traits in the cowpea MAGIC population. QTL plots for flowering time under full irrigation and long day (FTFILD), flowering time under restricted irrigation and long day (FTRILD), flowering time under full irrigation and short day (FTFISD), flowering time under restricted irrigation and short day (FTRISD), and BLUPs of environments (FLT_BLUP). The chromosome numbers are located on the x-axis and the negative log of the P-values on the y-axis. The genetic position of the colocalization between QTL and *a priori* genes are indicated by broken vertical lines. The texts displayed on the vertical broken lines are the names of *a priori* genes (blue for genes associated with multiple environments or traits, and black for genes associated with single environments or trait).

index of FxRRBLUP was higher than the rest of the GS models for FTRISD. In general, the mean of the slope and intercept for the GS models except SVR were usually close to the expected (1 and 0) (**Figures S12–S13**). MAS also deviated away from the expected slope and intercept (1 and 0) more than the FxRRBLUP, RKHS, and RRBLUP for FTRISD (**Figures S12–S13**). To evaluate the effect of photoperiod and irrigation regime on the performance of training population, each environment (day length and irrigation regime combination) was used as a training population to predict the rest in a di-allele manner. Results showed that prediction accuracy between environments in the same photoperiod was higher than environments in different photoperiod (**Figure S14**). Also, when training populations were under full irrigation, their prediction accuracies were higher than when training populations were under restricted irrigation (**Figure S14**). For FT_BLUP, GS models outperformed MAS except SVR which had the same mPA (0.59) as MAS while FxRRBLUP had the highest mPA and mCI among the GS models (**Figures S15** and **16**). Overall, **Table S7** showed that FxRRBLUP had the best performance in six out of the eight traits by environment combination.

GS and MAS for Maturity and Seed Size

For MAT (MT_BLUP, MFISD, and MRISD), RKHS and RRBLUP had better performance (**Figures 7** and **8**; **Tables S4** and **S5**) than the rest of the models including MAS. All models deviated from the expected slope and intercept estimates, but RRBLUP had

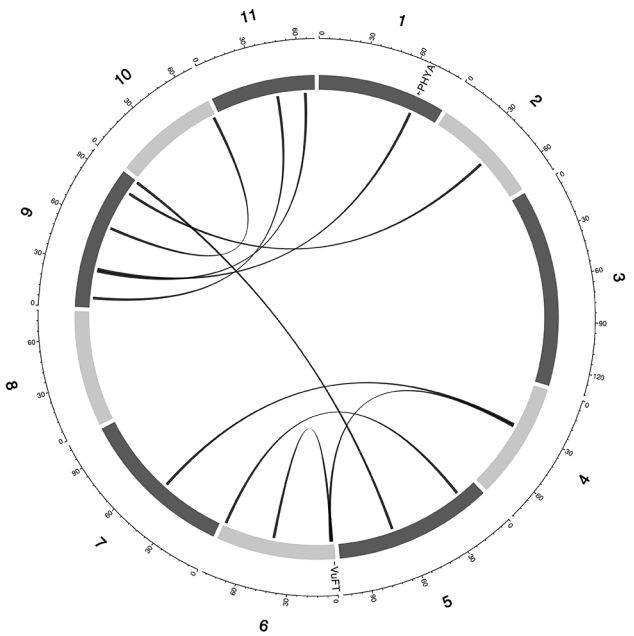


FIGURE 4 | Epistatic QTL for FLT_BLUP for MAGIC population. Chromosomes are shown in shades of gray, two-way interacting loci are connected with black solid lines, and colocalized *a priori* genes are texts between chromosomes and genetic map.

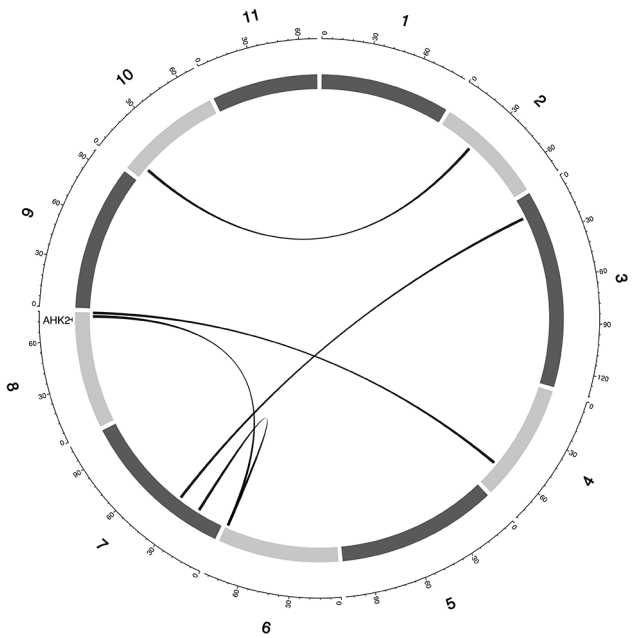


FIGURE 6 | Epistatic QTL for MAT_BLUP in MAGIC population. Chromosomes are shown in shades of gray, two-way interacting loci are connected with black solid lines, and colocalized *a priori* genes are texts between chromosomes and genetic map.

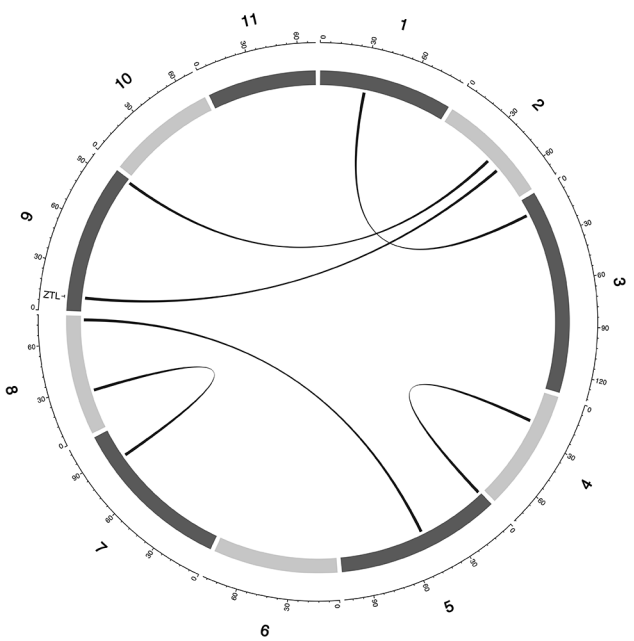


FIGURE 5 | Epistatic QTL for MAT_BLUP in MAGIC population. Chromosomes are shown in shades of gray, two-way interacting loci are connected with black solid lines, and colocalized *a priori* genes are texts between chromosomes and genetic map.

the least deviation for MRISD. For SS, FxRRBLUP had the best performance followed by MAS compared to the rest of the GS models (RKHS, RRBLUP, and SVR) (**Figures 7** and **8**; **Tables S5** and **S6**). GS and MAS models had varying levels of deviation from the expected estimates of slope and intercept. RKHS and RRBLUP were closer to the expected than FxRRBLUP and MAS (**Figures S12–S13**) while SVR had the highest deviation.

DISCUSSION

Epistasis Plays Important Roles in Determining the Genetic Architecture of Agronomic Traits in Cowpea

Multi-parental populations have demonstrated ability to facilitate robust characterization of genetic architecture in terms of genetic effect size, pleiotropy, and epistasis (Buckler et al., 2009; Brown et al., 2011; Peiffer et al., 2014; Bouchet et al., 2017; Mathew et al., 2018). Using the cowpea MAGIC population, this study showed that both additive main QTL and additive × additive epistatic QTL with large and (or) moderate effects underlie FLT, MAT, and SS in cowpea. Although we identified two-way epistatic interactions, results showed that some loci were involved in interactions with more than one independent loci (**Figures 4** and **5** and **Figures S4–11**). This implies the possibility of three-way epistatic interactions underlying some of the traits. Our inability to identify and discuss three-way epistatic interactions is due to the mapping approach used, which only mapped two-way epistatic

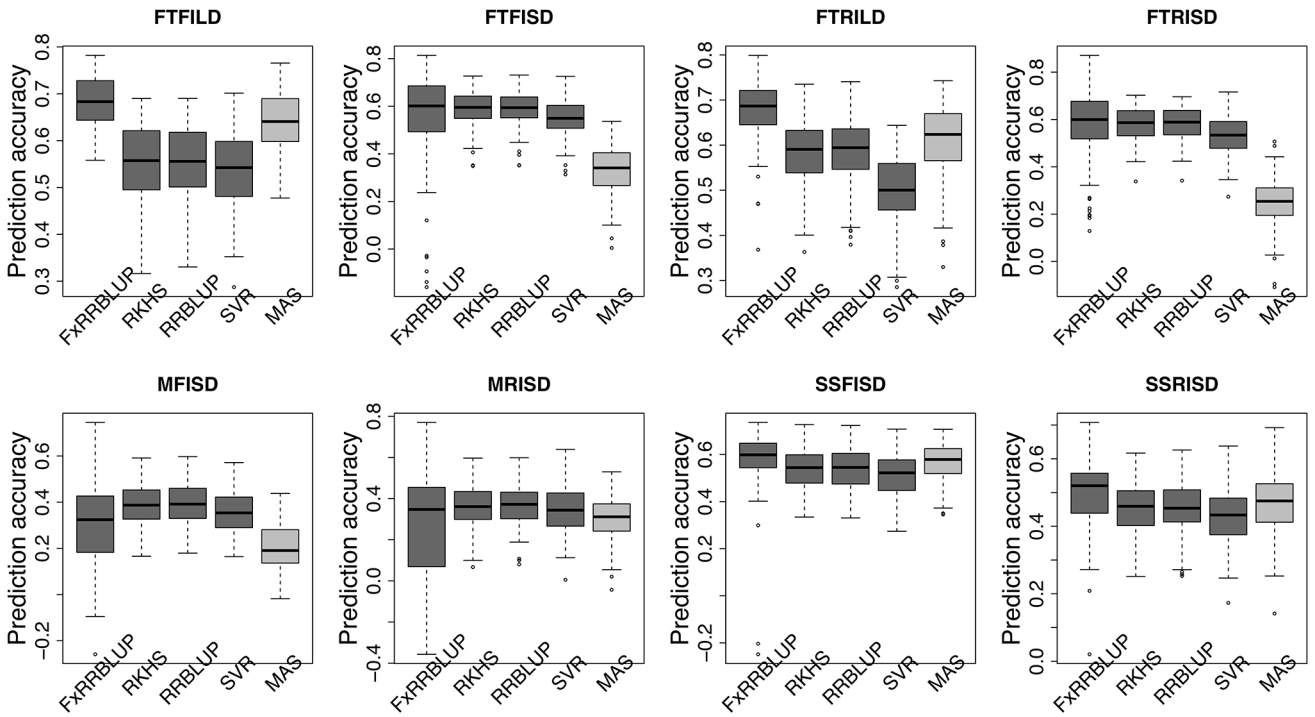


FIGURE 7 | Comparison of prediction accuracy across GS and MAS models. Boxplots in each panel showed the distribution of prediction accuracy values across 100 cycles for FxRRBLUP (ridge regression best linear unbiased prediction: parametric model with fixed effects), RKHS (reproducing Kernel Hilbert space; semi-parametric model), RRBLUP (ridge regression best linear unbiased prediction: parametric model with no fixed effects), SVR (support vector regression: non-parametric model), and MAS (marker-assisted selection) for flowering time under full irrigation and long day (FTFILD), flowering time under restricted irrigation and long day (FTRILD), flowering time under full irrigation and short day (FTFISD), flowering time under restricted irrigation and short day (FTRISD), maturity under full irrigation and short day (MFISD), maturity under restricted irrigation and short day (MRISD), seed size under full irrigation and short day (SSFISD), and seed size under restricted irrigation and short day (SSRISD).

interactions. Three-way epistatic interactions have been found to underlie FLT in the selfing crop specie barley (Mathew et al., 2018). Furthermore, overlaps between main and epistatic loci (**Figure S2**) indicate these to be main effect loci that are involved in epistatic interactions with other loci. However, one caveat that may also be responsible for some of the QTL among the overlaps is the false positive rate of SPEAML. The SPEAML software used for epistasis mapping showed high false positive rate with a sample size of 300 individuals (Chen et al., 2018a). It is possible that some of the overlapped QTL are main QTL that were miscategorized as epistatic loci by SPEAML since our cowpea MAGIC population had 305 RILs.

Distinct and Common Genetic Regulators Underlie Flowering Time

FLT is an important adaptive trait in breeding. Photoperiod impacted days to FLT as observed from the reaction norm plot for cowpea MAGIC FLT data which showed drastic reductions in days to flowering for RILs under short day compared to long days (**Figure 1**). Our mapping results (main effect and epistatic) showed that both unique and common loci underlie FLT variation under long and short photoperiod (**Figure 1**; **Figures S4**–**S8**). Epistatic loci underlie FLT in both selfing (Komeda, 2004; Juenger et al., 2005; Huang et al., 2013; Chen et al., 2018b; Li et al., 2018a; Mathew et al., 2018) and outcrossing (Buckler et al., 2009; Durand et al., 2012) species. In addition, the effect size of FLT loci differs between selfing and out crossing species as QTL effect sizes are large in the former (Lin et al., 1995; Maurer et al., 2015) and small in the later (Buckler et al., 2009). In the present study, the large effects (up to 25% PVE and additive effect of 7 days) of FLT loci were only identified under long-day photoperiod and not under short-day photoperiod (**Figure 2**, **Tables S3** and **S4**). The loci detected under short-day photoperiod were of moderate effects (PVE = 1–10% and maximum additive effect size of 2 days). The large effect size attributed to some of the loci that are unique to FLT adaptation under long photoperiod suggests the possible effect of recent selection ate these loci (Orr, 1998; Orr, 1999; Brown et al., 2011; Dittmar et al., 2016).

Conserved genetic pathways often underlie traits in plant species (Liu et al., 2013; Huang et al., 2017). Examination of colocalizations between *a priori* genes and QTL in this study identified putative cowpea orthologs of *A. thaliana* and *G. max* FLT that may underlie phenotypic variation in cowpea. FLT is affected by photoperiodicity and regulated by a network of

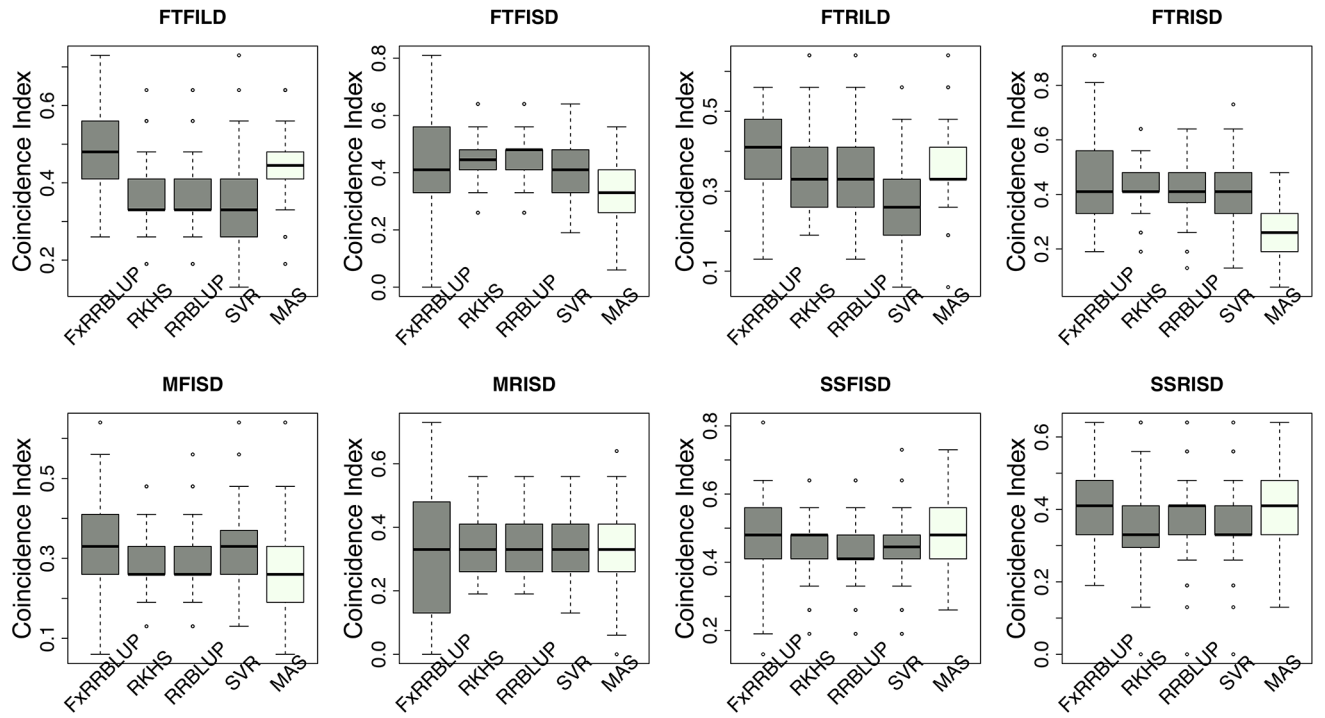


FIGURE 8 | Comparison of coincidence index across GS and MAS models. Boxplots in each panel showed the distribution of coincidence index values across 100 cycles for FxRRBLUP (ridge regression best linear unbiased prediction: parametric model with fixed effects), RKHS (reproducing Kernel Hilbert space; semi-parametric model), RRBLUP (ridge regression best linear unbiased prediction: parametric model with no fixed effects), SVR (support vector regression: non-parametric model), and MAS (marker-assisted selection) for flowering time under full irrigation and long day (FTFILD), flowering time under restricted irrigation and long day (FTRILD), flowering time under full irrigation and short day (FTFISD), flowering time under restricted irrigation and short day (FTRISD), maturity under full irrigation and short day (MFISD), maturity under restricted irrigation and short day (MRISD), seed size under full irrigation and short day (SSFISD), and seed size under restricted irrigation and short day (SSRISD).

genes (Sasaki et al., 2018) involved in floral initiation, circadian clock regulation, and photoreception (Lin, 2002). In addition, certain *a priori* genes were unique to either FLT under long day or short day. For instance, cowpea putative orthologs of photoreceptors (*PHY A* [Vigun01g205500] and *PHY E* [Vigun09g050600]) and circadian clock gene *PHYTOCLOCK1* (*PCL1* [Vigun11g157600]) colocalized with only QTL associated with FLT under long day, while cowpea putative orthologs of circadian clock genes (*Time for Coffee* [*TIC* (Vigun07g241700)] and *Zeitlupe* [*ZTL*]) colocalized with only QTL associated with FLT under short day. However, the cowpea putative ortholog of photoperiod gene *CONSTANS* (*CO* [Vigun05g024400]) colocalized with QTL associated with FLT under both long and short days. Thus, our study suggests that distinct and common genetic regulators control FLT adaptation to both long- and short-day photoperiod in cowpea. Further studies utilizing functional approaches will be helpful to decipher gene regulation patterns under both long- and short-photoperiod in cowpea.

Genetic Basis of Maturity and Seed Size

In this study, the genetic basis of MAT and SS were evaluated under short-day photoperiod only. Our study demonstrated that MAT under short day is controlled by moderate and small effect main and epistatic loci. MAT QTL were found to colocalize with cowpea putative orthologs of *Arabidopsis* circadian clock and photoperiod (*ZTL* [*ZEITLUPE*], *CO* [*CONSTANS*]) genes. One pleiotropic QTL (qVu9:8.37 colocalized with *ZTL* [*ZEITLUPE*]) was found to be associated with both MAT and FLT under restricted irrigation and short-day photoperiod. Pleiotropic QTL between MAT and FLT were also reported in soybean (Kong et al., 2018). This suggest a possible genetic basis for the positive relationship found between MAT and FLT in prior studies (Huynh et al., 2018; Owusu et al., 2018). A major large effect locus explaining up to 29% of the phenotypic variation was found to be associated with SS. This QTL was found at about 2cM from the cowpea ortholog of *Arabidopsis AHK2* SS gene. Further studies, using mapping panels with more diverse founders and more *a priori* genes will be required to identify further genes underlying natural variations in MAT and SS in cowpea.

Genetic Architecture Influenced GS and MAS Performance

GS models differ in their efficiency to capture complex cryptic interactions among genetic markers (de Oliveira Couto et al., 2017).

The traits evaluated in this study are controlled by both main effect and epistatic loci. In this study, comparison among the GS models showed that parametric and semi-parametric GS models outperformed non-parametric GS model for all traits. SVR, a non-parametric model, had the least prediction accuracy and coincidence index and also had the highest bias (**Figures S12** and **S13**). Previous studies have shown that semi-parametric and non-parametric models increased prediction accuracy under epistatic genetic architecture (Howard et al., 2014; Jacquin et al., 2016). In this study, none of semi-parametric and non-parametric models outperformed parametric models (**Figures 6** and **7**). Some of the studies comparing the performance of parametric, semi-parametric, and non-parametric GS models were based on simulations of traits controlled solely by epistatic genetic architectures. Therefore, the performance of the models under simulated combined genetic effects (additive + epistasis) is not well understood. The comparable performance of RKHS to RRBLUP (parametric model) in this study in terms of prediction accuracy, coincidence index, and bias estimates attests to RKHS ability to capture both additive and epistatic interactions (Gianola et al., 2006; Gianola and Van Kaam, 2008; De Los Campos et al., 2010; Gota and Gianola, 2014) for both prediction accuracy and selection of top performing lines. The performance of GS models is often indistinguishable, and RRBLUP has been recommended as an efficient parametric GS model (Heslot et al., 2012; Lipka et al., 2015). SVR had the worst performance with extremely high bias estimates.

Understanding the genetic architecture of agronomic traits can help improve accuracy of genomic predictions (Hayes et al., 2010; Swami, 2010). Our study demonstrated that the effect size of QTL associated with a trait played a role in the performance of GS and MAS models. For instance, for traits controlled by both large and moderate effect loci (FTFILD, FTRILD, SSFISD, and SSRISD), parametric model with known loci as fixed effect (FxRRBLUP) followed by MAS outperformed the rest of the GS models (RRBLUP, RKHS, and SVR). The use of known markers as fixed effects has been shown to increase prediction accuracy (Bernardo, 2014; Spindel et al., 2016) in parametric GS models. For traits that were controlled by moderate effect loci (FTFISD, FTRISD, MFISD, and MTRISD), our results showed that the two parametric GS models (FxRRBLUP and RRBLUP) and semi-parametric (RKHS) had similar prediction accuracy; however, FxRRBLUP had higher bias than RRBLUP and RKHS (**Figure S12–S13**). Furthermore, the performance of MAS in comparison to GS models in this study supported the fact that large effect loci are important influencers of MAS (Bernardo, 2008). For small breeding programs in developing countries, MAS might be a prudent choice over GS for traits controlled by large effects loci in cowpea since GS will require genotyping of more markers than MAS. The large effect loci identified in this study can be transferred to different breeding populations because they were identified in a MAGIC population with wide genetic background (Descalsota et al., 2018; Huynh et al., 2018). Our study thus demonstrates that prior knowledge of the genetic architecture of a trait can help make informed decision about the best GEB method to employ in breeding.

In summary, using the cowpea MAGIC population, our study identified both main QTL and two-way epistatic loci underlying FLT, MAT, and SS. These traits are oligogenic in genetic architecture with QTL effects ranging from small to large sizes. The effect size of the markers/QTL reported in this study may be upwardly biased due to the small size (n = 305) of the cowpea MAGIC population. Thus, studies with higher sample sizes (n > 1,000) will prove more accurate (Xu, 2003; King and Long, 2017). The identified QTL and their colocalized *a priori* genes will serve as stepping stone for future studies considering the molecular characterization of the genes underlying FLT, MAT, and SS in cowpea. Further, we demonstrated that prior knowledge of the genetic architecture of a trait can help make informed decision in GEB. Due to variations observed across photoperiod/environments for FLT, we will recommend the development of photoperiod insensitive lines in cowpea breeding. Also, given that some QTL were identified in specific environments, considerations should be given to field evaluation of mapping populations under contrasting environments that are representative of natural populations' environmental conditions. In addition, the cowpea MAGIC population may not capture all the genetic variation available in cowpea for FLT, MAT, and SS because only eight founders were used for its development. Thus, some of our markers may not be well diagnostic in breeding populations that do not share close ancestry with the cowpea MAGIC founders. Despite this limitation, this study still provides technical details that can be part of considerations for GS and MAS in cowpea breeding.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13827

AUTHOR CONTRIBUTIONS

MO obtained data from UCR; concept by MO and ZH; MO and ZH analyzed the data; MO, ZH, and PA wrote the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

We express our gratitude to Prof. Timothy Close, Prof. Philip Roberts, Dr. Bao-Lam Huynh and their team at the University of California—Riverside, USA for their incredible contributions to cowpea genomics and the privilege to use the cowpea MAGIC population data for this study. The MAGIC population development, phenotyping, and genotyping was supported in large part by grants from the Generation Challenge Program of the Consultative Group on International Agricultural Research, with additional support from the USAID Feed the Future

Innovation Lab for Collaborative Research on Grain Legumes (Cooperative Agreement EDH-A-00-07-00005), the USAID Feed the Future Innovation Lab for Climate Resilient Cowpea (Cooperative Agreement AID-OAA-A-13-00070), and NSF-BREAD (Advancing the Cowpea Genome for Food Security). We also thank Dr. Bode Olukolu, Dr. Sandeep Marla, and Fanna Maina for helping with the manuscript review. Thanks to Joanna, Eleazar, Christy, Grace, Fangfang, and Isimemen for their support.

"This manuscript has been released as a Pre-Print at: (Olatoye et al., 2019);

Olatoye et al. (2019) 'Epistasis detection and modeling for genomic selection in cowpea (Vigna unguiculata. L. Walp.)' *bioRxiv*. Cold Spring Harbor Laboratory, p. 576819. doi: 10.1101/576819."

SUPPLEMENTARY MATERIAL

All the R scripts used for analyses in the study are available at: https://github.com/marcbios/Cowpea.git

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00677/full#supplementary-material

DATA S1 | Candidate gene list

DATA S2 | Main effect QTL list

DATA S3 | Epistatic QTL list

DATA S4 | Genes that colocalized with main and epistatic QTL

REFERENCES

Aukerman, M. J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* 15 (11), 2730–2741. doi: 10.1105/tpc.016238

Barton, N. H, and Keightley, D. (2002). Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3, 11–21. doi: 10.1038/nrg700

Ben-Ari, G., and Lavi, U. (2012). "Marker-assisted selection in plant breeding," in *Plant Biotechnology and Agriculture* (Cambridge, MA: Elsevier), 163–184. doi: 10.1016/B978-0-12-381466-1.00011-0

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47 (3), 1082. doi: 10.2135/cropsci2006.11.0690

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48 (5), 1649–1664. doi: 10.2135/cropsci2008.03.0131

Bernardo, R. (2014). Genomewide Selection when major genes are known. *Crop Sci.* 54 (1), 68. doi: 10.2135/cropsci2013.05.0315

Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., et al. (2017). Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics* 206 (2), 573–585. doi: 10.1534/genetics.116.198499

Boukar, O., Belko, N., Chamarthi, S., Togola, A., Batieno, J., Owusu, E., et al. (2018). Cowpea (Vigna unguiculata): genetics, genomics and breeding. *Plant Breed.* 2018, 1–10. doi: 10.1111/pbr.12589

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Brown, P. J., Upadyayula, N., Mahone, G. S., Tian, F., Bradbury, P. J., Myles, S., et al. (2011). Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* 7 (11), e1002383. doi: 10.1371/journal.pgen.1002383

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., et al. (2009). The genetic architecture of maize flowering time. *Science* 325 (5941), 714–718. doi: 10.1126/science.1174276

Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E. A., Babu, R., et al. (2018). Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front. Plant Sci.* 9, 366. doi: 10.3389/fpls.2018.00366

Chen, A. H., Ge, W., Metcalf, W., Jakobsson, E., Mainzer, L. S., and Lipka, A. E. (2018a). An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers. *Heredity* 122, 660–671. doi: 10.1038/s41437-018-0162-2

Chen, J, Li, X., Cheng, C., Wang, Y., Qin, M., Zhu, H., et al. (2015). Characterization of epistatic interaction of QTLs LH8 and EH3 controlling heading date in rice. *Sci. Rep.* 4 (1), 4263. doi: 10.1038/srep04263

Chen, M., Ahsan, A., Meng, X., Rahaman, M., Chen, H., and Monir, M. (2018b). Identification epistasis loci underlying rice flowering time by controlling population stratification and polygenic effect. *DNA Res.* 26 (2), 119–130. doi: 10.1093/dnares/dsy043

Covarrubias-Pazaran, G. (2016). Genome-Assisted prediction of quantitative traits using the R package *sommer*. *PLoS ONE* 11 (6), 1–15. doi: 10.1371/journal.pone.0156744

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193 (2), 347–365. doi: 10.1534/genetics.112.147983

De Los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* 92 (4), 295–308. doi: 10.1017/S0016672310000285

de Oliveira Couto, E. G., Bandeira e Sousa, M., Jarquín, D., Burgueño, J., Crossa, J., Fritsche-Neto, R., et al. (2017). Genomic-Enabled prediction in maize using kernel models with genotype × environment interaction. *G3: Genes Genomes Genet.* 7 (6), 1995–2014. doi: 10.1534/g3.117.042341

Descalsota, G. I. L., Swamy, B. P. M., Zaw, H., Inabangan-Asilo, M. A., Amparado, A., Mauleon, R., et al. (2018). Genome-wide association mapping in a rice MAGIC Plus population detects QTLs and genes useful for biofortification. *Front. Plant Sci.* 9 (1347), 1–20. doi: 10.3389/fpls.2018.01347

Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A., and Schemske, D. W. (2016). Factors influencing the effect size distribution of adaptive substitutions. *Proc. R. Soc. B Biol. Scie.* 283 (1828), 1–8. doi: 10.1098/rspb.2015.3065

Durand, E., Bouchet, S., Bertin, P., Ressayre, A., Jamin, P., Charcosset, A., et al. (2012). Flowering time in maize: linkage and epistasis at a major effect locus. *Genetics* 190 (4), 1547–1562. doi: 10.1534/genetics.111.136903

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J.* 4 (3), 250. doi: 10.3835/plantgenome2011.08.0024

Fernandes, S. B., Dias, K. O. G., Ferreira, D. F., and Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* 131 (3), 747–755. doi: 10.1007/s00122-017-3033-y

Foolad, M. R., and Panthee, D. R. (2012). Marker-assisted selection in tomato breeding. *Crit. Rev. Plant Sci.* 31 (2), 93–123. doi: 10.1080/07352689.2011.616057

Gianola, D., and de los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genet. Res.* 90 (6), 525–540. doi: 10.1017/S0016672308009890

Gianola, D., and Van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178 (4), 2289–2303. doi: 10.1534/genetics.107.084285

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173 (3), 1761–1776. doi: 10.1534/genetics.105.049510

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40 (Database issue), D1178–D1186. doi: 10.1093/nar/gkr944

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

Haile, T. A., Heidecker, T., Wright, D., Neupane, S., Ramsay, L., Vandenberg, A., et al. (2019). Genomic selection for lentil breeding: empirical evidence. *bioRxiv*, 608406. doi: 10.1101/608406

Hall, A. E. (2004). Breeding for adaptation to drought and heat in cowpea. *Eur. J. Agron.* 21 (4), 447–454. doi: 10.1016/j.eja.2004.07.005

Hall, A. E., Cisse, N., Thiaw, S., Elawad, H. O. A., Ehlers, J. D., Ismail, A. M., et al. (2003). Development of cowpea cultivars and germplasm by the Bean/Cowpea CRSP. *Field Crops Res.* 82 (2–3), 103–134. doi: 10.1016/S0378-4290(03)00033-9

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLoS Genet.* 6 (9), e1001139. doi: 10.1371/journal.pgen.1001139

Hazen, S. P., Schultz, T. F., Pruneda-Paz, J. L., Borevitz, J. O., Ecker, J. R., and Kay, S. A. (2005). LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms. *Proc. Natl. Acad. Sci. U. S. A.* 102 (29), 10387–10392. doi: 10.1073/pnas.0503029102

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52 (1), 146–160. doi: 10.2135/cropsci2011.06.0297

Howard, R. et al. (2016). Evaluation of parametric and nonparametric statistical methods in genomic prediction. Dissertation. Ames (IA): Iowa State University.

Howard, R., Carriquiry, A. L., and Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes Genomes Genet.* 4 (6), 1027–1046. doi: 10.1534/g3.114.010298

Huang, P., Jiang, H., Zhu, C., Barry, K., Jenkins, J., Sandor, L., et al. (2017). Sparse panicle1 is required for inflorescence development in Setaria viridis and maize. *Nat. Plants* 3 (5), 17054. doi: 10.1038/nplants.2017.54

Huang, X., Ding, J., Effgen, S., Turck, F., and Koornneef, M. (2013). Multiple loci and genetic interactions involving flowering time genes regulate stem branching among natural variants of Arabidopsis. *New Phytol.* 199, 843–857. doi: 10.1111/nph.12306

Huynh, B. L., Ehlers, J. D., Huang, B. E., Muñoz-Amatriaín, M., Lonardi, S., Santos, J. R. P., et al. (2018). A multi-parent advanced generation inter-cross (MAGIC) population for genetic analysis and improvement of cowpea (*Vigna unguiculata* L. Walp). *Plant J.* 93 (6), 1129–1142. doi: 10.1111/tpj.13827

Jacquin, L., Cao, T.-V., and Ahmadi, N. (2016). A unified and comprehensible view of parametric and kernel methods for genomic prediction with application to rice. *Front. Genet.* 7, 145. doi: 10.3389/fgene.2016.00145

Jannink, J. L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings Funct. Genomics Proteomics* 9 (2), 166–177. doi: 10.1093/bfgp/elq001

Jarquin, D., Specht, J., and Lorenz, A. (2016). Prospects of genomic prediction in the usda soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3: Genes Genomes Genet.* 6 (8), 2329–2341. doi: 10.1534/g3.116.031443

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201 (2), 759–768. doi: 10.1534/genetics.115.177907

Johnson, N. (2008). Sewall wright and the development of shifting balance theory. *Nat. Educ.* 1 (1), 52. doi: 10.1093/rfs/hhx028

Juenger, T. E., Sen, S., Stowe, K. A., and Simms, E. L. (2005) "Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in *Arabidopsis thaliana*" in *Genetics of adaptation*. Ed. R. Mauricio (Dordrecht: Springer), 87–105.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - an S4 package for kernel methods in R. *J. Stat. Software* 11 (9), 1–20. doi: 10.18637/jss.v011.i09

King, E. G., and Long, A. D. (2017). The beavis effect in next-generation mapping panels in drosophila melanogaster. *G3 (Bethesda, Md)* 7 (6), 1643–1652. doi: 10.1534/g3.117.041426

Komeda, Y. (2004). Genetic regulation of time to flower in arabidopsis thaliana. *Annu. Rev. Plant Biol.* 55, 521–556. doi: 10.1146/annurev.arplant.55.031903.141644

Kong, L., Lu, S., Wang, Y., Fang, C., Wang, F., Nan, H., et al. (2018). Quantitative trait locus mapping of flowering time and maturity in soybean using next-generation sequencing-based analysis. *Front. Plant Sci.* 9, 995. doi: 10.3389/fpls.2018.00995

Kurek, A. (2018). Phenotypic and genomic selection for multi-trait improvement in soybean line and variety development. Dissertation. Ames (IA): Iowa State University.

Langyintuo, A. S., Lowenberg-DeBoer, J., Faye, M., Lambert, D., Ibro, G., Moussa, B., et al. (2003). Cowpea supply and demand in west and central africa. *Field Crops Res.* 82 (2–3), 215–231. doi: 10.1016/S0378-4290(03)00039-X

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018a). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci.* 115 (26), 6679–6684. doi: 10.1073/pnas.1718326115

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., et al. (2018b). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* 9, 190. doi: 10.3389/fpls.2018.00190

Lin, C. (2002). Photoreceptors and regulation of flowering time. *Plant Physiol.* 123 (1), 39–50. doi: 10.1104/pp.123.1.39

Lin, Y.-R., Schertz, K. F., and Paterson, A. H. (1995). Comparative analysis of qtls affecting plant height and maturity across the poaceae, in reference to an interspecific Sorghum population. *Genetics* 141, 391–411.

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., et al. (2015). From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118. doi: 10.1016/j.pbi.2015.02.010

Liu, C., Teo, Z. W. N., Bi, Y., Song, S., Xi, W., Yang, X., et al. (2013). A conserved genetic pathway determines inflorescence architecture in arabidopsis and rice. *Dev. Cell* 24 (6), 612–622. doi: 10.1016/j.devcel.2013.02.013

Long, N., Gianola, D., Rosa, G. J. M., and Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123 (7), 1065–1074. doi: 10.1007/s00122-011-1648-y

Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Ann. Rev. Genet.* 35 (1), 303–339. doi: 10.1146/annurev.genet.35.102401.090633

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theor. Appl. Genet.* 115 (7), 1003–1013. doi: 10.1007/s00122-007-0627-9

Massman, J. M., Jung, H.-J. G., and Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci.* 53 (1), 58. doi: 10.2135/cropsci2012.02.0112

Matei, G., Woyann, L. G., Milioli, A. S., de Bem Oliveira, I., Zdziarski, A. D., Zanella, R., et al. (2018). Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Mol. Breed.* 38 (9), 117. doi: 10.1007/s11032-018-0872-4

Mathew, B., Léon, J., Sannemann, W., and Sillanpää, M. J. (2018). Detection of epistasis for flowering time using bayesian multilocus estimation in a barley MAGIC population. *Genetics* 208 (2), 525–536. doi: 10.1534/genetics.117.300546

Mathieu, J., Yant, L. J., Mürdter, F., Küttner, F., and Schmid, M. (2009). Repression of flowering by the miR172 target SMZ. *PLoS Biol.* 7 (7), e1000148. doi: 10.1371/journal.pbio.1000148

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., et al. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16 (1), 290. doi: 10.1186/s12864-015-1459-7

Melchinger, A. E., Utz, H. F., Piepho, H. P., Zeng, Z. B., and Schön, C. C. (2007). The role of epistasis in the manifestation of heterosis: a systems-oriented approach. *Genetics* 177 (3), 1815–1825. doi: 10.1534/genetics.107.077537

Messina, C. D., Podlich, D., Dong, Z., Samples, M., and Cooper, M. (2011). Yield-trait performance landscapes: from theory to application in breeding maize for drought tolerance. *J. Exp. Bot.* 62 (3), 855–868. doi: 10.1093/jxb/erq329

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

Mohamed, A., Ali, R., Elhassan, O., Suliman, E., Mugoya, C., Masiga, C. W., et al. (2014). First products of DNA marker-assisted selection in sorghum released for cultivation by farmers in sub-saharan Africa. *J. Plant Sci. Mol. Breed.* 3 (1), 3. doi: 10.7243/2050-2389-3-3

Moore, J. H., and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *Am. J. Human Genet.* 85 (3), 309–320. doi: 10.1016/j.ajhg.2009.08.006

Muchero, W., Diop, N. N., Bhat, P. R., Fenton, R. D., Wanamaker, S., Pottorff, M., et al. (2009). A consensus genetic map of cowpea [Vigna unguiculata (L) Walp]. and synteny based on EST-derived SNPs. *Proc. Natl. Acad. Sci. U. S. A.* 106 (43), 18159–18164. doi: 10.1073/pnas.0905886106

Okogbenin, E., Porto, M. C. M., Egesi, C., Mba, C., Espinosa, E., Santos, L. G., et al. (2007). Marker-assisted introgression of resistance to cassava mosaic disease into latin American germplasm for the genetic improvement of cassava in Africa. *Crop Sci.* 47 (5), 1895–1904. doi: 10.2135/cropsci2006.10.0688

Olatoye, M. O., Hu, Z., and Aikpokpodion, P. O. (2019). Epistasis detection and modeling for genomic selection in cowpea (*Vigna unguiculata*. L. Walp). *bioRxiv*, 576819. doi: 10.1101/576819

Orozco-Arroyo, G., Paolo, D., Ezquer, I., and Colombo, L. (2015). Networks controlling seed size in Arabidopsis. *Plant Reprod.* 28 (1), 17–32. doi: 10.1007/s00497-015-0255-5

Orr, H. A. (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52 (4), 935–949. doi: 10.1111/j.1558-5646.1998.tb01823.x

Orr, H. A. (1999). The evolutionary genetics of adaptation: a simulation study. *Genet. Res. Camb.* 74, 207–214. doi: 10.1017/S0016672399004164

Owusu, E. Y., Akromah, R., Denwar, N. N., Adjebeng-Danquah, J., Kusi, F., and Haruna, M. (2018). Inheritance of early maturity in some cowpea *(Vigna unguiculata* (L). Walp). Genotypes under rain fed conditions in Northern Ghana. *Adv. Agric.* 2018, 8930259.1–10. doi: 10.1155/2018/8930259

Pan, Y. B., Tew, T. L., Schnell, R. J., Viator, R. P., Richard, E. P., Grisham, M. P., et al. (2006). Microsatellite DNA marker-assisted selection of Saccharum spontaneum cytoplasm-derived germplasm. *Sugar Tech.* 8 (1), 23–29. doi: 10.1007/BF02943737

Pazhamala, L., Saxena, R. K., Singh, V. K., Sameerkumar, C. V., Kumar, V., Sinha, P., et al. (2015). Genomics-assisted breeding for boosting crop improvement in pigeonpea (*Cajanus cajan*). *Front. Plant Sci.* 6, 50. doi: 10.3389/fpls.2015.00050

Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., et al. (2014). The genetic architecture of maize height. *Genetics* 196 (4), 1337–1356. doi: 10.1534/genetics.113.159152

Perez, P. (2014). BGLR : a statistical package for whole genome regression and prediction. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Rice, B., and Lipka, A. E. (2019). Evaluation of rr-blup genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *Plant Genome* 12 (1), 0. doi: 10.3835/plantgenome2018.07.0052

Rieseberg, L. H., Archer, M. A., and Wayne, R. K. (1999). Transgressive segregation, adaptation and speciation. *Heredity* 83 (4), 363–372. doi: 10.1038/sj.hdy.6886170

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., et al. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* 7, 1666. doi: 10.3389/fpls.2016.01666

Saghai Maroof, M. A., Jeong, S. C., Gunduz, I., Tucker, D. M., Buss, G. R., and Tolin, S. A. (2008). Pyramiding of soybean mosaic virus resistance genes by marker-assisted selection. *Crop Sci.* 48 (2), 517. doi: 10.2135/cropsci2007.08.0479

Sasaki, E., Frommlet, F., and Nordborg, M. (2018). GWAS with heterogeneous data: estimating the fraction of phenotypic variation mediated by gene expression data. *Genes Genom Genet* 8 (9), 3059–3068. doi: 10.1534/g3.118.200571

Schneider, K. A., Brothers, M. E., and Kelly, J. D. (1997). Marker-assisted selection to improve drought resistance in common bean. *Crop Sci.* 37 (1), 51. doi: 10.2135/cropsci1997.0011183X003700010008x

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genet.* 44 (7), 825–830. doi: 10.1038/ng.2314

Somers, D. E., Schultz, T. F., Milnamow, M., and Kay, S. A. (2000). ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell.* 101 (3), 319–329. doi: 10.1016/S0092-8674(00)80841-7

Spindel, J., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., et al. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Sun, X., Ma, P., and Mumm, R. H. (2012). Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. *PLoS ONE* 7 (11), e50604. doi: 10.1371/journal.pone.0050604

Swami, M. (2010). Complex traits: using genetic architecture to improve predictions. *Nat. Rev. Genet.* 11 (11), 748. doi: 10.1038/nrg2888

Tardivel, A., Sonah, H., Belzile, F., and O'Donoughue, L. S. (2014). Rapid identification of alleles at the soybean maturity gene E3 using genotyping by sequencing and a haplotype-based approach. *Plant Genome* 7 (2), 0. doi: 10.3835/plantgenome2013.10.0034

Utz, H. F., Melchinger, A. E., and Schön, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154 (4), 1839–1849. doi: 10.2307/1403680

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. 1st edn Jordan, M., Lauritzen S.L., Lawless, J.F., Nair, V., editors. New York: Springer. doi: 10.1007/978-1-4757-3264-1

Varshney, R. K., Penmetsa, R. V., Dutta, S., Kulwal, P. L., Saxena, R. K., Datta, S., et al. (2010). Pigeonpea genomics initiative (PGI): an international effort to improve crop productivity of pigeonpea (Cajanus cajan L). *Mol. Breed.* 26 (3), 393–408. doi: 10.1007/s11032-009-9327-2

Varshney, R. K., Ribaut, J. M., Buckler, E. S., Tuberosa, R., Rafalski, A. J., and Langridge, P. (2012). Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* 30 (12), 1172–1176. doi: 10.1038/nbt.2440

Volis, S., Shulgina, I., Zaretsky, M., and Koren, O. (2010). Epistasis in natural populations of a predominantly selfing plant. *Heredity* 106, 300–309. doi: 10.1038/hdy.2010.79

Wen, L., Chang, H.-X., Brown, P. J., Domier, L. L., and Hartman, G. L. (2019). Genome-wide association and genomic prediction identifies soybean cyst nematode resistance in common bean including a syntenic region to soybean Rhg1 locus. *Hortic. Res.* 6 (1), 9. doi: 10.1038/s41438-018-0085-3

Wenkel, S., Turck, F., Singer, K., Gissot, L., Gourrierec, J. Le, Samach, A., et al. (2006). Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene *CONSTANS*. *Am. Soc. Plant Biol.* 12 (12), 2473–2484. doi: 10.1105/tpc.12.12.2473

Wong, C. K., and Bernardo, R. (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.* 116 (6), 815–824. doi: 10.1007/s00122-008-0715-5

Xu, S. (2003). Theoretical basis of the beavis effect. *Genetics* 165 (4), 2259–2268.

Xu, Y., Li, P., Yang, Z., and Xu, C. (2017). Genetic mapping of quantitative trait loci in crops. *Crop J.* 5 (2), 175–184. doi: 10.1016/j.cj.2016.06.003

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi: 10.1038/ng1702

Yu, K., Park, S. J., and Poysa, V. (2000). Marker-assisted selection of common beans for resistance to common bacterial blight: efficacy and economics. *Plant Breed.* 119 (5), 411–415. doi: 10.1046/j.1439-0523.2000.00514.x

Zhang, T., Yuan, Y., Yu, J., Guo, W., and Kohel, R. J. (2003). Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. *Theor. Appl. Genet.* 106 (2), 262–268. doi: 10.1007/s00122-002-1101-3

Zhao, Y., Mette, M. F., Gowda, M., Longin, C. F. H., and Reif, J. C. (2014). Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity* 112 (6), 638–645. doi: 10.1038/hdy.2014.1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Olatoye, Hu and Aikpokpodion. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fgene.2019.01006

Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection

Antoine Allier[1,2], Christina Lehermeier[2], Alain Charcosset[1], Laurence Moreau[1] and Simon Teyssèdre[2]*

[1] GQE-Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France,
[2] Genetics and Analytics Unit, RAGT2n, Druelle, France

The implementation of genomic selection in recurrent breeding programs raises the concern that a higher inbreeding rate could compromise the long-term genetic gain. An optimized mating strategy that maximizes the performance in progeny and maintains diversity for long-term genetic gain is therefore essential. The optimal cross-selection approach aims at identifying the optimal set of crosses that maximizes the expected genetic value in the progeny under a constraint on genetic diversity in the progeny. Optimal cross-selection usually does not account for within-family selection, i.e., the fact that only a selected fraction of each family is used as parents of the next generation. In this study, we consider within-family variance accounting for linkage disequilibrium between quantitative trait loci to predict the expected mean performance and the expected genetic diversity in the selected progeny of a set of crosses. These predictions rely on the usefulness criterion parental contribution (UCPC) method. We compared UCPC-based optimal cross-selection and the optimal cross-selection approach in a long-term simulated recurrent genomic selection breeding program considering overlapping generations. UCPC-based optimal cross-selection proved to be more efficient to convert the genetic diversity into short- and long-term genetic gains than optimal cross-selection. We also showed that, using the UCPC-based optimal cross-selection, the long-term genetic gain can be increased with only a limited reduction of the short-term commercial genetic gain.

Keywords: genomic prediction, optimal cross-selection, usefulness criterion, parental contributions, genetic diversity, Bulmer effect

OPEN ACCESS

Edited by:
Charles Chen, Oklahoma State University, United States

Reviewed by:
Changwei Shao, Yellow Sea Fisheries Research Institute (CAFS), China
Zibei Lin, La Trobe University, Australia

****Correspondence:***
Antoine Allier
antoine.allier@inra.fr

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *10 May 2019*
Accepted: *20 September 2019*
Published: *29 October 2019*

Citation:
Allier A, Lehermeier C, Charcosset A, Moreau L and Teyssèdre S (2019) Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection. Front. Genet. 10:1006. doi: 10.3389/fgene.2019.01006

INTRODUCTION

Successful breeding requires strategies that balance immediate genetic gain with the maintenance of population diversity to sustain long-term progress (Jannink, 2010). At each selection cycle, plant breeders are facing the choice of new parental lines and the way in which these are mated, to improve the mean population performance and generate the genetic variation on which selection will act. As breeding programs from different companies compete for short-term gain, breeders tend to use intensively the most performant individuals sometimes at the expense of genetic diversity (Rauf et al., 2010; Gerke et al., 2015; Allier et al., 2019a). The identification of the crossing plan that maximizes the performance in progeny and limits diversity reduction for long-term genetic gain is essential.

Historically, breeders used to select the best individuals based on phenotypic observations, considered as a proxy of their breeding value, i.e., the expected value of their progeny. In order to better estimate the breeding value of individuals, phenotypic selection has been complemented by pedigree-based prediction of breeding values (Henderson, 1984; Piepho et al., 2008) and more recently by genomic prediction of breeding values (Meuwissen et al., 2001), taking advantage of the availability of cheap high-density genotyping. In genomic selection (GS), a model calibrated on phenotype and genotype information of a training population is used to predict genomic estimated breeding values (GEBVs) from genome-wide marker information. A truncation selection is commonly applied on GEBVs, and the selected individuals are intercrossed to create the next generation. The interest of GS is due to the acceleration of selection progress by shortening generation interval, the increase in selection intensity, and the increase in accuracy (Hayes et al., 2010; Daetwyler et al., 2013; Heslot et al., 2015). As a consequence, compared to phenotypic selection, GS is expected to accelerate the loss of genetic diversity due to the rapid fixation of genomic regions with large effects, but also the higher probability to select individuals that are the closest to the training population and are therefore predicted more accurately (Clark et al., 2011; Pszczola et al., 2012). As a result, it has been shown in an experimental study (Rutkoski et al., 2015) and by stochastic simulations (Jannink, 2010; Lin et al., 2016) that GS increases the loss of diversity compared to phenotypic selection. Thus, the optimization of mating strategies in GS breeding programs is a critical area of theoretical and applied research.

Several approaches have been suggested to balance the short- and long-term genetic gain while selecting crosses in GS. In line with Kinghorn, (2011), Pryce et al. (2012), and Akdemir and Isidro-Sánchez (2016), the selection of a set of crosses requires two components: (i) a cross-selection index (CSI) that measures the interest of a set of crosses and (ii) an algorithm to find the set of crosses that maximizes the CSI.

The CSI may consider crosses individually; i.e., the interest of a cross does not depend on the other crosses in the selected set. In classical recurrent GS, candidates with the highest GEBVs are selected and intercrossed to maximize the expected progeny mean in the next generation. In this case, the CSI is simply the mean of parental GEBVs. However, such an approach maximizes neither the expected response to selection in the progeny, which involves genetic variance generated by Mendelian segregation within each family, nor the long-term genetic gain. Alternative measures of the interest of a cross have been proposed to account for parent complementarity, based on within cross variability and expected response to selection. Daetwyler et al. (2015) proposed the optimal haploid value (OHV) that accounts for the complementarity between parents of a cross for predefined haplotype segments. Using stochastic simulations, the authors observed that OHV selection yielded higher long-term genetic gain and preserved greater amount of genetic diversity than truncation GS. However, OHV accounts for neither the position of quantitative trait loci (QTLs) nor the linkage disequilibrium between QTLs (Lehermeier et al., 2017b; Müller et al., 2018). Schnell and Utz (1975) proposed the usefulness criterion (UC) of a cross to evaluate the expected response to selection in its progeny. The UC of a cross accounts for the progeny mean (μ) that is the mean of parental GEBVs and the progeny standard deviation (σ) the selection intensity (i) and the selection accuracy (h): $UC = \mu + ih\sigma$. Zhong and Jannink (2007) proposed to predict progeny variance using estimated QTL effects, accounting for linkage between loci. Genome-wide marker effects have also been considered to predict the progeny variance with computationally intensive stochastic simulations (e.g., Mohammadi et al., 2015). Recently, an unbiased predictor of progeny variance (σ^2) has been derived in Lehermeier et al. (2017b) for two-way crosses and extended in Allier et al. (2019b) for multiparental crosses implying up to four parents. Lehermeier et al. (2017b) observed that using UC as a CSI increased the short-term genetic gain compared to using OHV or mean parental GEBV. Similar results have been obtained by simulations by Müller et al. (2018), considering the expected maximum haploid breeding value (EMBV) that is akin to the UC for normally distributed and fully additive traits.

Alternatively, one can consider a more holistic CSI for which the interest of a cross depends on the other selected crosses. This is the case in optimal contribution selection (Wray and Goddard, 1994; Meuwissen, 1997; Woolliams et al., 2015), where a set of candidate parents is evaluated as a whole regarding the expected short-term gain and the associated risk on loosing long-term gain. Optimal contribution selection aims at identifying the optimal contributions (c) of candidate parents to the next generation obtained by random mating, in order to maximize the expected genetic value in the progeny (V) under a certain constraint on inbreeding (D). Optimal cross-selection, further referred as OCS, is an extension of the optimal contribution selection to deliver a crossing plan that maximizes V by considering additional constraints on the allocation of mates in crosses to limit D (Kinghorn et al., 2009; Kinghorn, 2011; Akdemir and Isidro-Sánchez, 2016; Gorjanc et al., 2018; Akdemir et al., 2018). In GS, the expected genetic value in progeny (V) to be maximized is the mean of parental GEBV ($\boldsymbol{a}$) weighted by parental contributions $\boldsymbol{c}$, i.e $\boldsymbol{c'a}$, and the constraint on inbreeding (D) to be minimized is $\boldsymbol{c'Kc}$ with $\boldsymbol{K}$ a genomic coancestry matrix. Differential evolutionary algorithms have been proposed to obtain optimal solutions for $\boldsymbol{c}$ and the crossing plan (Storn and Price, 1997; Kinghorn et al., 2009; Kinghorn, 2011). Optimal contribution selection is commonly used in animal breeding (Woolliams et al. 2015) and is increasingly adopted in plant breeding (Akdemir and Isidro-Sánchez, 2016; De Beukelaer et al., 2017; Lin et al., 2017; Gorjanc et al., 2018; Akdemir et al., 2018).

In plant breeding, one typically has larger biparental families than in animal breeding. Especially with GS, the selection intensity within-family can be largely increased so that plant breeders capitalize much more on the segregation variance within families than animal breeders. In previous works, the genetic gain (V) and constraint (D) have been defined at the level of the progeny before within-family selection. Exceptions are the work of Shepherd and Kinghorn (1998) and Akdemir and Isidro-Sánchez (2016); Akdemir et al. (2018), who added a term to V accounting for within cross variance assuming linkage equilibrium between QTLs. To our knowledge, no previous

study considered linkage disequilibrium (LD) between QTLs. Furthermore, as observed in historical wheat data (Fradgley et al., 2019) and using simulations in a maize context (Allier et al., 2019b), within-family selection also affects the effective contribution of parents to the next generation. This likely biases the prediction of inbreeding/diversity in the next generation, which to our knowledge has not been considered in previous studies.

In this study, we propose to adjust V and D terms so that within-family selection of the candidate parents for the next generation is accounted for. We propose to use the usefulness criterion parental contribution (UCPC) approach (Allier et al., 2019b) that enables to predict the expected mean performance of the selected fraction of progeny and to predict the contribution of parents to the selected fraction of progeny. We compared our OCS strategy based on UCPC with other cross-selection strategies, in a long-term simulated recurrent GS breeding program involving overlapping generations (**Figure 1A**). Our objectives were to demonstrate (1) the interest of UCPC to predict the genetic diversity in the selected fraction of progeny and (2) the interest of accounting for within-family selection in OCS for both short- and long-term genetic gains.

MATERIALS AND METHODS

Simulated Breeding Program

Breeding Program

We simulated a breeding program to compare the effect of different CSIs on short- and long-term genetic gain in a realistic breeding context considering overlapping and connected generations (i.e., cohorts) and the use of doubled haploid (DH) technology to derive progeny (**Figure 1A**). We considered that the process to derive DH progeny from a cross and to phenotype and genotype DH lines takes 3 years. Furthermore, we considered as candidate parents of a cohort T the selected fraction of DH progeny of the three last available cohorts, i.e., T-3, T-4 and T-5 (**Figures 1A**, **B**).

Each simulation replicate started from a population of 40 founders sampled among 57 Iodent maize genotypes from the Amaizing project (Rio et al., 2019; Allier et al., 2019b). We sampled 1,000 biallelic QTLs among the 40,478 high-quality single-nucleotide polymorphisms (SNPs) from the Illumina MaizeSNP50 BeadChip (Ganal et al. 2011), with consensus genetic positions from Giraud et al. (2014). The sampling process obeyed two constrains: a QTL minor allele frequency ≥ 0.2 and a distance between two consecutive QTLs ≥ 0.2 cM. Each QTL was assigned an additive effect sampled from a Gaussian distribution with a mean of zero and a variance of 0.05, and the favorable allele was attributed at random to one of the two SNP alleles.

We initiated a virtual breeding program starting from the founder genotypes with a burn-in period of 20 years that mimicked recurrent phenotypic selection. Burn-in started by randomly crossing the 40 founders into 20 biparental families, i.e., two-way crosses, during the first 3 years to initiate three overlapping cohorts. In each cohort, 80 DH progeny genotypes per cross were simulated. Phenotypes were simulated considering the genotype at QTLs, an error variance corresponding to a trait repeatability of 0.4 in the founder population and no genotype by environment interactions. For phenotyping, every individual was evaluated in four environments in 1 year. Since no secondary trait was considered and sufficient seed production for extensive progeny testing was assumed, we simulated a unique within-family selection of the 5% best progeny (i.e., 4 DHs) that is a common selection

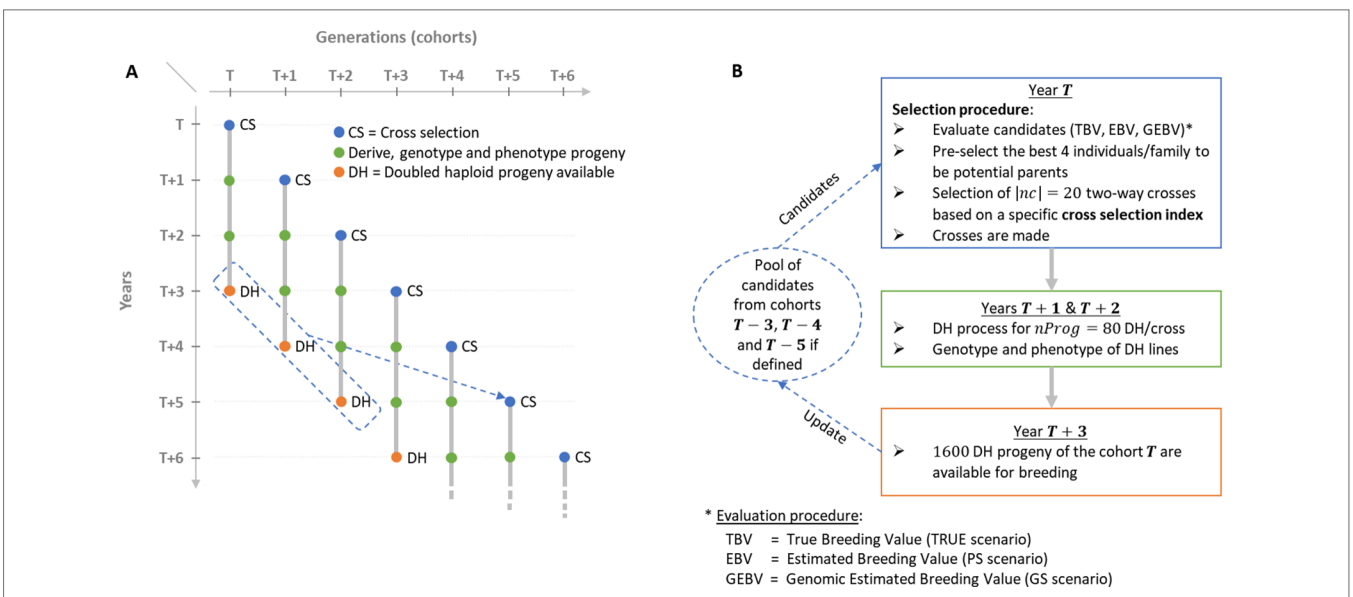


FIGURE 1 | Schematic view of the simulated breeding program. **(A)** Overall view of the breeding program and overlapping cohorts. **(B)** Life cycle of a given post burn-in cohort T depending on the scenario considered (TRUE with 1,000 known QTL effects, PS in absence of genomic information or GS with 2,000 noncausal SNPs estimated effects).

intensity in maize breeding. During burn-in, we first considered within-family phenotypic selection and then used the 50 DHs with the largest phenotypic mean as potential parents of the next cohort. These were randomly mated, i.e., without any constraint on parental contributions, to generate 20 biparental families of 80 DH lines. After 20 years of burn-in, this created extensive linkage disequilibrium as often observed in elite plant breeding programs (e.g., Van Inghelandt et al., 2011). We then compared different CSIs for 60 years of recurrent GS using DH technology (**Figure 1**). As in burn-in, each cohort T was generated by 20 two-way crosses ($|nc|$=20) of 80 DH progeny each ($nProg$ = 80). Candidate parents of cohort T were selected from the available DH of the three cohorts: T=3, T-4, and T-5 (**Figures 1A**, **B**). Per family, the 4 DH lines (i.e., 5%) with the largest breeding values, detailed in "Evaluation scenario" section, were considered as potential parents, yielding 4 DH lines/family × 20 families/cohort × 3 cohorts = 240 potential parents. Considering these N = 240 potential parents, $N(N\text{-}1)/2$ = 28,680 two-way crosses are possible. The set of $|nc|$ = 20 two-way crosses among these 28,680 candidate crosses was defined using different CSI detailed in the following sections. This simulated scheme yielded overlapping and connected cohorts as it is standard in practical plant breeding (**Figure 1A**). A detailed description of the simulated breeding program and the material is provided in **Supplementary Material (File S1)**.

Evaluation Scenarios

We considered different scenarios for genome-wide marker effects and progeny evaluation. In order to eliminate the uncertainty caused by the estimation of marker effects, we first compared several CSI assuming that we have access to the positions and effects of the 1,000 QTLs (referred to as TRUE scenario). For a representative subset of the CSI showing differentiated results in the TRUE scenario, we also considered a more realistic scenario where the effects of QTLs are unknown and selection was based on the effects of 2,000 noncausal SNPs randomly sampled over the genome. In this scenario, marker effects were obtained by back-solving (Wang et al., 2012) a G-BLUP model fitted using blupf-90 AI-REML solver (Misztal, 2008). This scenario was referred to as GS scenario, and marker effects used to predict the CSI were estimated every year with all candidate parents that were phenotyped and genotyped. The progeny were selected on their GEBV considering their phenotypes and their genotypes at noncausal SNPs. As a benchmark, we also considered a phenotypic selection scenario where progeny were selected based on their phenotypic mean (PS scenario). For details on the evaluation models, see **Supplementary Material (File S1)**. In the following, for sake of clarity, we present the different cross-selection strategies considering selection based on known QTL effects and positions (TRUE scenario). In GS scenario, QTL effects and positions were replaced by estimated marker effects and positions.

Cross-Selection Strategies

Optimal Cross-Selection Not Accounting for Within-Family Selection

Considering N homozygote candidate parents, $N(N\text{-}1)/2$ two-way crosses are possible. We define a crossing plan $\boldsymbol{nc}$ as a set of $|nc|$ crosses out of possible two-way crosses, giving the index of selected crosses, i.e., with the i^{th} element $nc(i)\in[1,N(N\text{-}1)/2]$. The $(N \times 1)$ dimensional vector of candidate parents contributions $\boldsymbol{c}$ is defined as

$$\boldsymbol{c} = \frac{1}{|nc|}\left(\boldsymbol{Z_1 c_1} + \boldsymbol{Z_2 c_2}\right), \tag{1}$$

where $\boldsymbol{Z}_1$ (respectively $\boldsymbol{Z}_2$) is a $(N \times |nc|)$ dimensional design matrix that links each N candidate parent to the first (respectively second) parent in the set of crosses $\boldsymbol{nc}$, $\boldsymbol{c}_1$ (respectively, $\boldsymbol{c}_2$) is a $(|nc| \times 1)$ dimensional vector containing the contributions of the first (respectively, second) parent to progeny, i.e., a vector of 0.5 when assuming no selection within crosses.

The $(N \times 1)$ dimensional vector of candidate parents true breeding values is $\boldsymbol{a} = \boldsymbol{X}\beta_T$ where $\boldsymbol{X} = (x_1,\ldots,\boldsymbol{x}_N)'$ is the $(N \times m)$ dimensional matrix of known parental genotypes at m biallelic QTLs, where $\boldsymbol{x}_p$ denotes the $(m \times 1)$ dimensional genotype vector of parent $p\in[1,N]$ with the j^{th} element coded as 1 or −1 for the genotypes AA or aa at QTL j. β_T is the $(m \times 1)$ dimensional vector of known additive QTL effects for the quantitative agronomic performance trait considered. The genetic gain $V(\boldsymbol{nc})$ for this set of two-way crosses is defined as the expected mean performance in the DH progeny:

$$V(\boldsymbol{nc}) = \boldsymbol{c'a}\,. \tag{2}$$

We define the constraint on diversity (D) as the mean expected genetic diversity in DH progeny (He, Nei, 1973):

$$D\left(\boldsymbol{nc}\right) = 1 - \boldsymbol{c'Kc}\,, \tag{3}$$

where $\boldsymbol{K} = \frac{1}{2}\left(\frac{1}{m}\boldsymbol{XX'} + 1\right)$ is the $(N \times N)$ dimensional identity by state (IBS) coancestry matrix between the N candidates. **Supplementary Material (File S2)** details the relationship between the IBS coancestry among parents ($\boldsymbol{K}$), the parental contributions to progeny ($\boldsymbol{c}$) and the mean expected heterozygosity in progeny $He = \frac{1}{m}\sum_{j=1}^{m} 2p_j\left(1-p_j\right)$ where p_j the frequency of the genotypes AA at QTL j in the progeny.

Accounting for Within-Family Selection in OCS

In the OCS, as defined above, the progeny derived from the $\boldsymbol{nc}$ crosses are all expected to contribute to the next generation. We propose to consider $V(\boldsymbol{nc})$ and $D(\boldsymbol{nc})$ terms accounting for the fact that only a selected fraction of each family will be candidate for the next generation (e.g., 5% per family in our simulation study). For this, we apply the UCPC approach proposed by Allier et al. (2019b) for two-way crosses and extend its use to evaluate the interest of a set $\boldsymbol{nc}$ of two-way crosses after selection in progeny.

UCPC for Two-Way Crosses

Two inbred lines P_1 and P_2 are considered as parental lines for a candidate cross $P_1 \times P_2$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2)'$ denotes their genotyping

matrix. Following Lehermeier et al. (2017b), the DH progeny mean and progeny variance of the performance in the progeny before selection can be computed as follows:

$$\mu_T = 0.5\left(\boldsymbol{x}'_1\boldsymbol{\beta}_T + \boldsymbol{x}'_2\boldsymbol{\beta}_T\right), \tag{4a}$$

$$\sigma_T^2 = \boldsymbol{\beta}_T'\ \Sigma\ \boldsymbol{\beta}_T, \tag{4b}$$

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and β_T were defined previously, and Σ is the $(m\times m)$ -dimensional variance covariance matrix of QTL genotypes in DH progeny defined in Lehermeier et al. (2017b).

To follow parental contributions, we consider P_1 parental contribution as a normally distributed trait (Allier et al., 2019b). As we only consider two-way crosses and biallelic QTLs, we can simplify for computational reasons the formulas by using IBS parental contributions computed for polymorphic QTLs between P_1 and P_2 instead of using identity-by-descent parental contributions (Allier et al., 2019b). We define the $(m\times 1)$ -dimensional vector β_{C1} to follow P_1 genome contribution at QTLs as $\boldsymbol{\beta}_{C1} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)'\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)}$. We compute the mean of P_1 contribution in the progeny before selection $\mu_{C1}=0.5(\boldsymbol{x}'_1\beta_{C1}+\boldsymbol{x}'_2\beta_{C1}+1)$. The progeny variance σ^2_{C1} for P_1 contribution in the progeny before selection is computed using Eq. 4b by replacing β_T by β_{C1} The progeny mean for P_2 contribution is then defined as $\mu_{C2} = 1\text{-}\mu_{C1}$.

Following Allier et al. (2019b), we compute the covariance between the performance and P_1 contribution in progeny as follows:

$$\sigma_{T,\ C1} = \boldsymbol{\beta}'_T\Sigma\ \boldsymbol{\beta}_{C1}. \tag{5}$$

The expected mean performance of the selected fraction of progeny, i.e., UC (Schnell and Utz, 1975), of the cross $P_1\times P_2$ is as follows:

$$UC^{(i)} = \mu_T + ih\sigma_T, \tag{6}$$

where i is the within-family selection intensity, and the exponent (i) in UC expresses the dependency of UC on the selection intensity i. We considered a selection accuracy h=1 as in Zhong and Jannink (2007), which holds when selecting on true breeding values in TRUE scenario. As discussed further, we also considered h = 1 when selecting crosses based on UCPC in GS scenario. The correlated responses to selection on P_1 and P_2 genome contributions in the selected fraction of progeny are as follows (Falconer and Mackay, 1996):

$$c_1^{(i)} = \mu_{C1} + i\frac{\sigma_{T,C1}}{\sigma_T} \text{ and } c_2^{(i)} = 1 - c_1^{(i)}. \tag{7}$$

Cross-Selection Based on UCPC

Accounting for within-family selection intensity i, the genetic gain term $V^{(i)}(\boldsymbol{nc})$ for a set of two-way crosses $\boldsymbol{nc}$ is defined as the expected performance in the selected fraction of progeny:

$$V^{(i)}\left(\boldsymbol{nc}\right) = \frac{1}{\left|\boldsymbol{nc}\right|}\sum\nolimits_{j\in\boldsymbol{nc}} UC^{(i)}(j). \tag{8}$$

The constraint on diversity $D^{(i)}(\boldsymbol{nc})$ in the selected progeny is defined as follows:

$$D^{(i)}\left(\boldsymbol{nc}\right) = 1 - \boldsymbol{c}^{(i)\prime}\boldsymbol{K}\boldsymbol{c}^{(i)}, \tag{9}$$

where $\boldsymbol{c}^{(i)}$ is defined like $\boldsymbol{c}$ in Eq. 1 but accounting for within-family selection by replacing the ante-selection parental contributions $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ by the post-selection parental contributions $\boldsymbol{c}_1^{(i)}$ and $\boldsymbol{c}_2^{(i)}$ (Eq. 7), respectively. Note that considering the absence of selection in progeny, i.e., i = 0, yields $V^{(i=0)}(\boldsymbol{nc})$ being the mean of parent breeding values (Eq. 2) and $D^{(i=0)}(\boldsymbol{nc})$ being the expected diversity in progeny before selection (Eq. 3), which is equivalent to optimal cross-selection as proposed by Gorjanc et al. (2018). The R code (R Core Team, 2017) to evaluate a set of crosses as presented in the UCPC-based optimal cross-selection is provided in **Supplementary Material (File S3)**.

Multiobjective Optimization Framework

In practice, one does not evaluate only one set of crosses but several ones in order to find the optimal set of crosses to reach a specified target that is a function of $V^{(i)}(\boldsymbol{nc})$ and $D^{(i)}(\boldsymbol{nc})$. We use the ε-constraint method (Haimes et al., 1971; Gorjanc and Hickey, 2018) to solve the multiobjective optimization problem:

$$\begin{aligned} &\max_{\mathbf{nc}}\ V^{(i)}\left(\boldsymbol{nc}\right) \\ &\text{with } D^{(i)}\left(\boldsymbol{nc}\right) \geq He\left(t\right), \end{aligned} \tag{10}$$

where $He(t)$, $\forall t\in[0,t^*]$ is the minimal diversity constraint at time t. A differential evolutionary (DE) algorithm was implemented to find the set of $\boldsymbol{nc}$ crosses that is a Pareto-optimal solution of Eq. 10 (Storn and Price, 1997; Kinghorn et al., 2009; Kinghorn, 2011). DE is an optimization process inspired by natural selection. It started from an initial population of 7,170 random candidate solutions that are improved during 1,000 iterations through mutation (random changes in candidate solutions), recombination (exchanges between candidate solutions), and selection (every iteration a candidate solution was replaced by its mutated and recombined version if superior). The direct consideration of $He(t)$ in the optimization allows to control the decrease in genetic diversity similarly to what was suggested for controlling inbreeding rate in animal breeding (Woolliams et al., 1998, Woolliams et al., 2015). The loss of diversity along time is controlled by the targeted diversity trajectory, i.e., $He(t)$,

$\forall t \in [0,t^*]$, where $t^* \in \mathbb{N}^*$ is the time horizon when the genetic diversity $He(t^*) = He^*$ should be reached. In this study, $He(t)$ is defined as follows:

$$He(t) = \begin{cases} He^0 + \left(\frac{t}{t^*}\right)^s \left(He^* - He^0\right), \forall\ t \in [\![0, t^*]\!], \\ He^*, \forall\ t > t^* \end{cases} \quad (11)$$

where He^0 is the initial diversity at $t = 0$, and s is a shape parameter with $s = 1$ for a linear trajectory. **Figure 2** gives an illustration of alternative trajectories that can be defined using Eq. 11.

Cross-Selection Indices

We considered different cross-selection approaches varying in the within-family selection intensity (i) in $V^{(i)}(\boldsymbol{nc})$, $D^{(i)}(\boldsymbol{nc})$ (Eq. 10) and in the targeted diversity trajectory $He(t)$ (Eq. 11). We first considered as a benchmark the absence of constraint $D^{(i)}(\boldsymbol{nc})$, i.e., $He(t) = 0$, $\forall t$. We defined two alternative CSIs PM (parental mean) and UC, respectively considering $V^{(i=0)}(\boldsymbol{nc})$ and $V^{(i=2.06)}(\boldsymbol{nc})$, with $i = 2.06$ corresponding to the selection of the 5% most performant progeny per family. PM is equivalent to cross the best candidates together without accounting for within cross variance, while UC is defined as crossing candidates based on the expected mean performance of the 5% selected fraction of progeny. Note that the absence of constraint on diversity also means the absence of constraint on parental contributions. To compare optimal cross-selection accounting or not for within-family selection, we considered three linear diversity trajectories (Eq. 11) with $He^* = \{0.01, 0.10, 0.15\}$ that should be reached in $t^* = 60$ years. We defined the OCS methods, further referred to as OCS-He*, with $V^{(i=0)}(\boldsymbol{nc})$ and $D^{(i=0)}(\boldsymbol{nc})$. We defined the UCPC cross-selection methods, further referred to as UCPC-He*, with $V^{(i=2.06)}(\boldsymbol{nc})$ and $D^{(i=2.06)}(\boldsymbol{nc})$. The eight CSIs considered are summarized in **Table 1**.

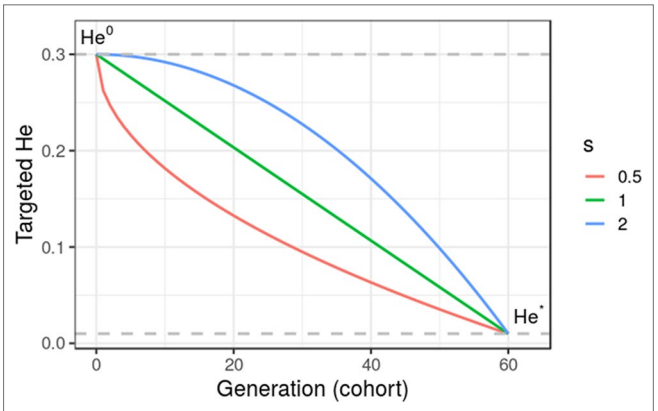


FIGURE 2 | Targeted diversity trajectories for three different shape parameters (s = 1, linear trajectory; s = 2, quadratic trajectory; and s = 0.5, inverse quadratic trajectory) for fixed initial diversity ($He^0 = 0.3$) at generation 0 and targeted diversity ($He^* = 0.01$) at generation 60 ($t^* = 60$). We considered in this study only linear trajectories (s = 1).

Simulation 1: Interest of UCPC to Predict the Diversity in the Selected Fraction of Progeny

Simulation 1 aimed at evaluating the interest to account for the effect of selection on parental contributions, i.e., post-selection parental contributions (using UCPC), compared to ignore selection, i.e., ante-selection parental contributions (similarly as in OCS), to predict the genetic diversity (He) in the selected fraction of progeny of a set of 20 crosses (using Eqs. 9 and 3, respectively). We considered a within-family selection intensity corresponding to selecting the 5% most performant progeny. We used the same genotypes, genetic map, and known QTL effects as for the first simulation replicate of the PM CSI in the TRUE scenario (**Table 1**). We extracted the simulated genotypes of 240 DH candidate parents of the first post burn-in cohort (further referred as E1) and of 240 DH candidate parents of the 20th post burn-in cohort (further referred as E2). Due to the selection process, E1 showed a higher diversity and lower performance compared to E2. We randomly generated 300 sets of 20 two-way crosses: 100 sets of intrageneration E1 crosses (E1 × E1), 100 sets of intrageneration E2 crosses (E2 × E2), and 100 sets of intergeneration and intrageneration crosses randomly sampled (E1 × E2, E1 × E1, E2 × E2). We derived 80 DH progeny per cross and predicted the ante- and post-selection parental contributions to evaluate the post-selection genetic diversity (He) for each set of crosses. We estimated the empirical post-selection diversity for each set of crosses and compared predicted and empirical values considering the mean prediction error as the mean of the difference between predicted He and empirical post-selection He, and the prediction accuracy as the squared correlation between predicted He and empirical post-selection He.

Simulation 2: Comparison of Different Csis

We ran 10 independent simulation replicates of all eight CSI summarized in **Table 1** for 60 years post burn-in considering known effects at the 1,000 QTLs (TRUE scenario). We also compared in 10 independent simulation replicates the CSI: PM,

TABLE 1 | Summary of tested cross-selection indices (CSI) in TRUE scenario defined for a set of crosses ***nc*** depending on the within-family selection intensity *i*.

Cross-selection index (CSI)	Gain term	Diversity term
PM	$V^{(i=0)}(\boldsymbol{nc})$	–
OCS-He* (3 different He*)	$V^{(i=0)}(\boldsymbol{nc})$	$D^{(i=0)}(\boldsymbol{nc})$
UC	$V^{(i=2.06)}(\boldsymbol{nc})$	–
UCPC-He* (3 different He*)	$V^{(i=2.06)}(\boldsymbol{nc})$	$D^{(i=2.06)}(\boldsymbol{nc})$

*$He^* = \{0.15; 0.10; 0.01\}$ to be reached linearly (s = 1) at the end of simulation ($t^* = 60$ years). $V^{(i=0)}(\boldsymbol{nc})$ is the averaged parental mean (PM) of crosses in **nc** and $V^{(i=2.06)}(\boldsymbol{nc})$ is the averaged usefulness criterion (UC) of crosses in **nc** considering a within-family selection intensity of 2.06. $D^{(i=0)}(\boldsymbol{nc})$ and $D^{(i=2.06)}(\boldsymbol{nc})$ are the expected genetic diversity in the progeny before and after within-family selection, respectively.*

UC, OCS-He* and UCPC-He* with He* = 0.01 considering estimated marker effect at the 2,000 SNPs (GS scenario) and PM based only on phenotypic evaluation (PS scenario). We followed several variables on the 80 DH progeny/family × 20 crosses realized every year. At each cohort $T \in [0,60]$ with $T = 0$ co rresponding to the last burn-in cohort, we computed the additive genetic variance as the variance of the 1,600 DH progeny true breeding values (TBVs): $\sigma_A^2(T) = var(TBV(T))$. We followed the mean genetic merit of all progeny $\mu(T)$ = mean($TBV(T)$) and of the 10 most performant progeny $\mu_{10}(T)$= mean(max($TBV(T)$)) as a proxy of realized performance that could be achieved at a commercial level by releasing these lines as varieties. Then, we centered and scaled the two genetic merits to obtain realized cumulative genetic gains in units of genetic standard deviation at the end of the burn-in ($T = 0$), at the whole progeny level $G(T) = (\mu(T) - \mu(0)) / \sqrt{\sigma_A^2(0)}$ and at the commercial level $G_{10}(T) = (\mu_{10}(T) - \mu(0)) / \sqrt{\sigma_A^2(0)}$.

The interest of long-term genetic gain relies on the ability to breed at long term, which depends on the short-term economic success of breeding. Following this rationale, we penalized strategies that compromised the short-term commercial genetic gain using the discounted cumulative gain following Dekkers et al. (1995) and Chakraborty et al. (2002). In practice, we computed the weighted sum of the commercial gain value in each generation $\sum_{T=1}^{60} w_T \, G_{10}(T)$, where the discounted weights $w_T = 1/(1+\rho)^T, \forall T \in [1,60]$ were scaled to have $\sum_{T=1}^{60} w_T = 1$ and ρ is the interest rate per generation. The discounted weights measure how much breeders will care about future genetic gain compared to today's genetic gain, also referred as the "net present value" of long-term gain in finance. For ρ = 0, the weights were $w_{T \in [1,60]}$ = 1/60; i.e., the same importance was given to all cohorts. We compared different values of ρ and reported results for ρ = 0, ρ = 0.04 giving approximatively seven times more weight to short-term gain (after 10 years) compared to long-term gain (after 60 years) and ρ = 0.2 giving nearly no weight to gain after 30 years of breeding.

We also measured the additive genic variance at QTLs $\sigma_a^2(T) = \sum_{j=1}^{m} 4 \, p_j(T)(1 - p_j(T))\beta_j^2$, the mean expected heterozygosity at QTLs (He, Nei, 1973) $He(T) = m^{-1} \sum_{j=1}^{m} 2 \, p_j(T)(1 - p_j(T))$, and the number of QTLs where the favorable allele was fixed or lost in the progeny, with $p_j(T)$ the allele frequency at QTL $j \in [1,m]$ in the 1,600 DH progeny and β_j the additive effect of the QTL j. In addition, we considered the ratio of additive genetic over genic variance σ_A^2 / σ_a^2. hich provides an estimate of the amount of additive genic variance captured by negative covariances between QTLs, known as the Bulmer effect under directional selection (Bulmer, 1971, Bulmer, 1980; Lynch and Walsh, 1999). All these variables were further averaged on the 10 simulation replicates, and the standard error divided by the square root of the number of replicates is reported.

RESULTS

Simulation 1

Compared to the usual approach that ignores the effect of selection on parental contributions, accounting for the effect of within-family selection increased the squared correlation (R^2) between predicted genetic diversity and genetic diversity in the selected fraction of progeny (**Figures 3A**, **B**) for all three types of crosses. The squared correlation between predicted genetic diversity and post-selection genetic diversity for intrageneration crosses was only slightly increased (E1 × E1: from 0.811 to 0.822 and E2 × E2: from 0.880 to 0.888), while the squared correlation for sets of crosses involving also intergeneration crosses showed a larger increase (from 0.937 to 0.987) (**Figures 3A**, **B**). Using post-selection parental contributions instead of ante-selection parental contributions also reduced the mean prediction error of He (predicted – empirical He) (**Figures 4A**, **B**) for all three types of crosses. The mean prediction error for intrageneration crosses

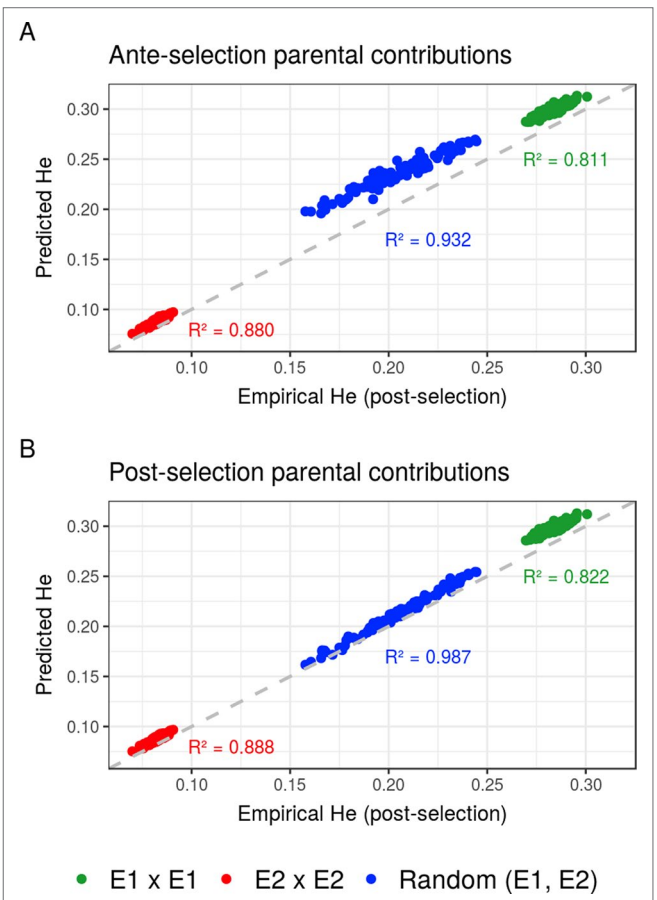


FIGURE 3 | Squared correlations (R^2) between predicted genetic diversity (He) and empirical He in the selected fraction of progeny of a set of 20 biparental crosses in the TRUE scenario considering **(A)** ante-selection parental contributions or **(B)** post-selection parental contributions to predict He. In total, 100 sets of each three types of crosses (intrageneration: E1xE1 and E2xE2 or randomly intragenerations and intergenerations): random (E1, E2) are shown, and the squared correlations between predicted and empirical post-selection He are given in the corresponding color.

was only slightly reduced (E1 × E1: from 0.006 to 0.005 and E2 × E2: from 0.016 to 0.015), while the mean prediction error for sets involving intergeneration crosses was more reduced (from 0.032 to 0.008) (**Figures 4A**, **B**). The mean prediction error of He was reduced but still positive when considering post-selection parental contributions, which means that the genetic diversity in the selected fraction of progeny remains overestimated. Note that the ante-selection contributions predicted well the empirical genetic diversity before selection for all three types of crosses (mean prediction error = 0.000 and $R^2 > 0.992$, results not shown).

Simulation 2

Interest of UC Over PM

Considering known QTL effects (TRUE scenario), we observed that UC yielded significantly higher short- and long-term genetic gain at commercial level (G_{10}) than PM (on average, G_{10} = 9.316 [±0.208] compared to 8.338 [±0.195] 10 years post burn-in and G_{10} = 18.293 [±0.516] compared to 15.744 [±0.449] 60 years post burn-in; **Figures 5B**, **C**; **Supplementary Material** [**Table S1 File S4**]). When considering the whole progeny mean performance (G), PM nonsignificantly outperformed UC for the first 5 years (on average, G = 4.647 [±0.174] compared to 4.633 [±0.138] 5 years post burn-in), and after 5 years, UC significantly outperformed PM (on average, G = 7.620 [±0.158] compared to 7.197 [±0.199] 10 years post burn-in) [**Figure 5A**, **Supplementary Material (Table S1 File S4)**]. UC showed higher genic (σ_a^2) and genetic (σ_A^2) additive variances than PM (**Figures 6A**, **B**), but both yielded a genic and genetic variance near zero after 60 years of breeding. The genetic over genic variance ratio (σ_A^2 / σ_a^2) was also higher for UC compared to PM (**Figure 6C**). The evolution of genetic diversity (He) along years followed the same tendency as the genic variance (**Figure 7A**, **Figure 6A**). UC fixed more favorable alleles at QTLs after 60 years (**Figure 7B**) and lost less favorable alleles at QTLs than PM in all 10 simulation replicates, with an average of 243.1 (±4.547) QTLs where the favorable allele was lost compared to 274.9 (±4.283) QTLs for PM [**Figure 7C**; **Supplementary Material (Table S1 File S4)**].

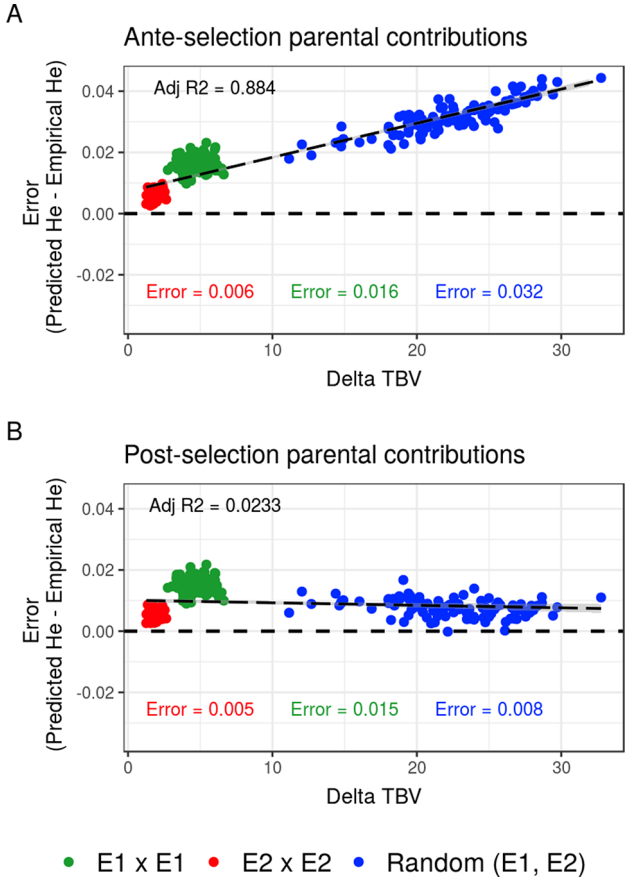


FIGURE 4 | Mean prediction error (predicted – empirical) of predicting the genetic diversity (He) in the selected fraction of progeny of a set of 20 biparental crosses in the TRUE scenario depending on the mean difference of performance between parents (Delta true breeding value TBV). Mean prediction error is measured as the predicted He – empirical post-selection He, considering **(A)** ante-selection parental contributions or **(B)** post-selection parental contributions to predict He. In total, 100 sets of each three types of crosses (intrageneration: E1 × E1 and E2 × E2 or randomly intra and inter-generations): random (E1, E2) are shown, and the averaged errors are given in the corresponding color.

Targeted Diversity Trajectory

Considering known QTL effects (TRUE scenario), the tested optimal cross-selection methods OCS-He* and UCPC-He* showed lower short-term genetic gain at the whole progeny level (G; **Figure 5A**) and at the commercial level (G_{10}; **Figures 5B**, **C**) but significantly higher long-term genetic gains than UC at 60 years **Supplementary Material (Table S1 File S4)**. The lower the targeted diversity He*, the higher the short-term and midterm genetic gain at both whole progeny (G; **Figure 5A**) and commercial (G_{10}; **Figures 5B**, **C**) levels. The higher the targeted diversity He*, the higher the long-term genetic gain except for OCS-He* = 0.10 and OCS-He* = 0.01 that performed similarly after 60 years (on average, G_{10} = 21.925 [±0.532] and 21.892 [±0.525]; **Figure 5B**, **Supplementary Material** [**Table S1 File S4**]). The highest targeted diversity (He* = 0.15) showed a strong penalty at the short term and midterm, while the intermediate targeted diversity (He* = 0.10) showed a lower penalty at the short term and midterm compared to the lowest targeted diversity (He* = 0.01) (**Figures 5A–C**).

For all targeted diversities and all simulation replicates, accounting for within-family selection (UCPC-He*) yielded a significantly higher short-term commercial genetic gain (G_{10}) after 5 and 10 years compared to OCS-He* [**Figures 5B**, **C**; **Supplementary Material (Table S1 File S4)**]. Long-term commercial genetic gain (G_{10}) after 60 years was also higher for UCPC-He* than for OCS-He* with He* = 0.01 in the 10 simulation replicates (on average, G_{10} = 22.869 [±0.641] compared to 21.892 [±0.525]) and less importantly with He* = 0.10 in nine out of 10 replicates (on average, G_{10} = 22.474 [±0.645] compared to 21.925 [±0.532]). However, for He* = 0.15, UCPC-He* outperformed OCS-He* at the long term in only three out of 10 replicates (on average, G_{10} = 20.665 [±0.573] compared to 20.938 [±0.553]) [**Figures 5B**, **C**; **Supplementary Material (Table S1 File S4)**]. The discounted cumulative gain giving more weight to short-term than to long-term gain (ρ = 0.04) was higher for UCPC-He* than

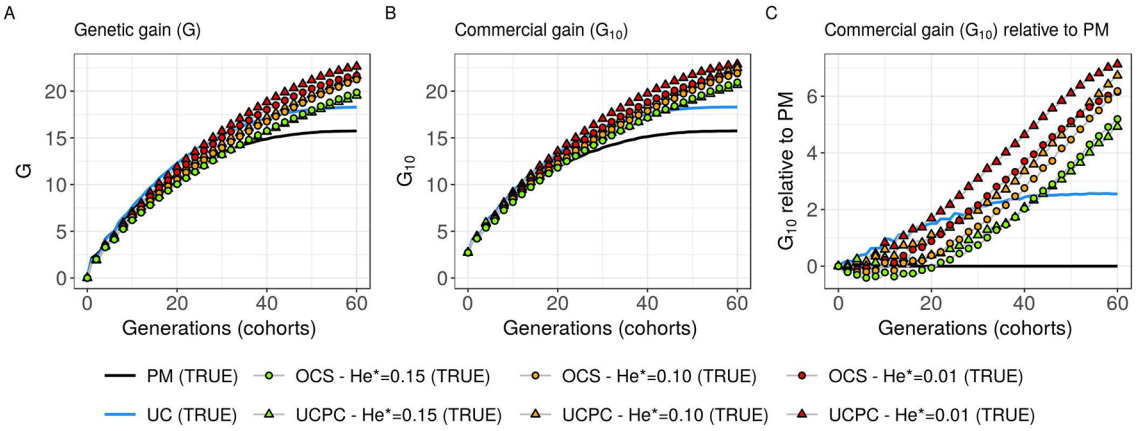


FIGURE 5 | Genetic gains for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Genetic gain (G) measured as the mean of the whole progeny, **(B)** commercial genetic gain (G_{10}) measured as the mean of the 10 best progeny, and **(C)** G_{10} relative to selection based on parental mean (PM).

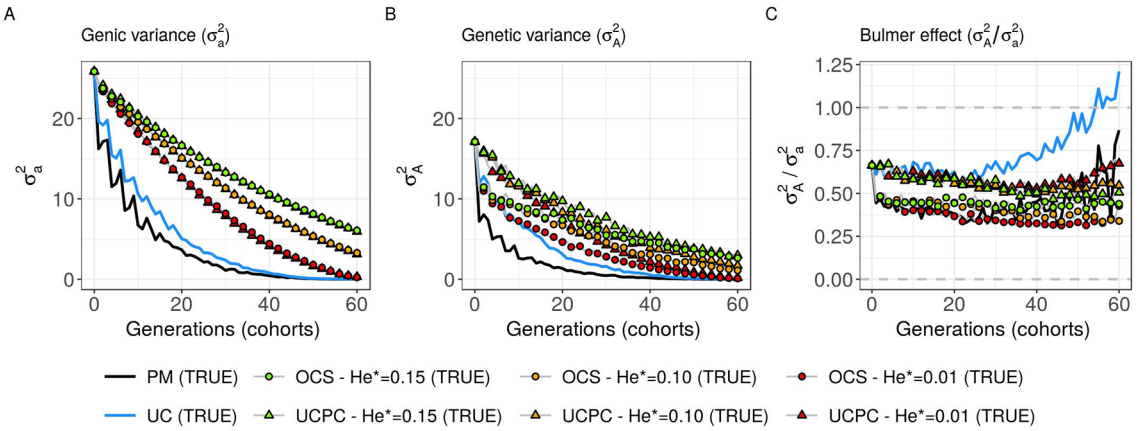


FIGURE 6 | Genetic and genic additive variances for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection, and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Additive genic variance (σ^2_a) measured on the whole progeny, **(B)** additive genetic variance (σ^2_A) measured on the whole progeny, and **(C)** ratio of genetic over genic variance (σ^2_A / σ^2_a) reflecting the Bulmer effect.

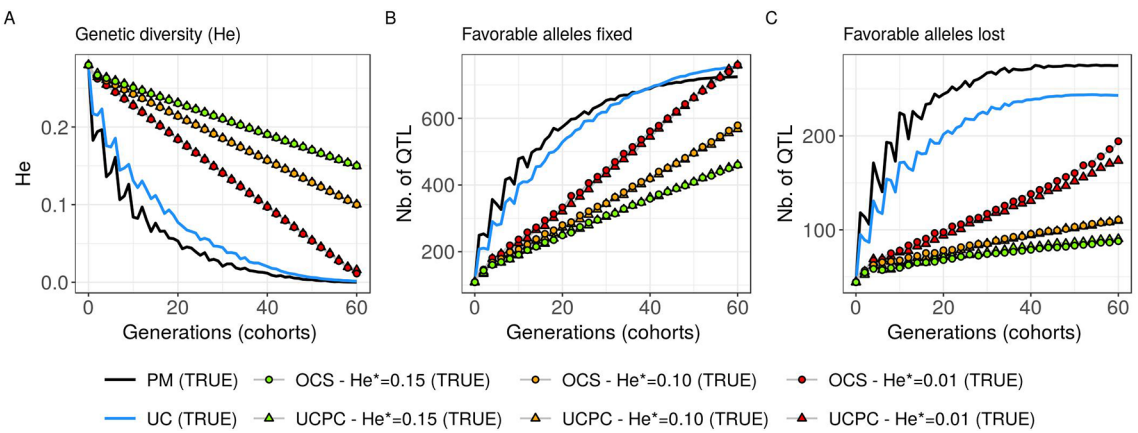


FIGURE 7 | Genetic diversity at QTLs for different cross-selection indices in the TRUE scenario (PM: parental mean, UC: usefulness criterion, OCS-He*: optimal cross-selection, and UCPC-He*: UCPC-based optimal cross-selection) according to the generations. **(A)** Genetic diversity at QTLs in the whole progeny (*He*), **(B)** number of QTLs where the favorable allele is fixed in the whole progeny, and **(C)** number of QTLs where the favorable allele is lost in the whole progeny.

OCS-He* in all simulation replicates for He* = 0.01 (on average, 12.321 [±0.284] compared to 11.675 [±0.262]), in all simulation replicates for He* = 0.10 (on average, 11.788 [±0.280] compared to 11.278 [±0.264]) and in nine out of 10 simulation replicates for He* = 0.15 (on average, 11.176 [±0.250] compared to 10.884 [±0.250]) (**Table 2**). Discounted cumulative gain giving the same weight to short- and long-term gain (ρ = 0) was also higher for UCPC-He* compared to OCS-He* (**Table 2**). When giving almost no weight to long-term gain after 30 years (ρ = 0.2), the best CSI appeared to be UC [on average, 6.822 (±0.145)] followed by the UCPC-He* with the lowest constraint on diversity (i.e., He* = 0.01) [on average, 6.682 (±0.143)].

For a given He*, the additive genic variance (σ_a^2; **Figure 6A**) and genetic diversity at QTLs (He; **Figure 7A**) were constrained by the targeted diversity trajectory for both UCPC-He* or OCS-He*. However, UCPC-He* and OCS-He* behaved differently for genetic variance (σ_A^2 ; **Figure 6A**) resulting in differences for the ratio genetic over genic variances (σ_A^2 / σ_a^2; **Figure 6C**). UCPC-He* yielded a higher ratio than OCS-He* (**Figure 6C**) independently of the targeted diversity He* at short term and midterm. For low targeted diversity (He* = 0.01), UCPC-He* showed in all 10 replicates a lower number of QTLs where the favorable allele was lost compared to OCS-He* (**Figure 7C**; **Supplementary Material [Table S1 File S4]**, on average 173.6 [±4.031] QTLs-194.3 [±2.633] QTLs).

GS Scenario With Estimated Marker Effects

Considering estimated marker effects (GS scenario) yielded lower genetic gain than when considering known marker effects [**Figures 5–8** and **Supplementary Material (Tables S1 and S2 File S4)**]. However, the short- and long-term superiority of the UC over the CSI ignoring within cross variance (PM) was consistent with estimated effects (on average, G_{10} = 8.338 [±0.237] compared to 7.713 [±0.256] 10 years post burn-in and G_{10} = 15.367 [±0.358] compared to 13.287 [±0.436] 60 years post burn-in; **Figure 8**, **Supplementary Material [Table S2 File S4]**). Similarly, the long-term superiority of UCPC-He* = 0.01 over UC was conserved in all 10 replicates (on average, G_{10} = 16.398 [±0.426] compared to 14.438 [±0.320] 40 years post burn-in and G_{10} = 18.161 [±0.470] compared to 15.367 [±0.358] 60 years post burn-in; **Figure 8**, **Supplementary Material [Table S2 File S4]**). Before the 40th year, UC and UCPC-He* = 0.01 performed similarly **Supplementary Material (Table S2 File S4)**. In GS scenario, UCPC-He* = 0.01 outperformed OCS-He* = 0.01 during the first 20 years in all 10 replicates (on average, G_{10} = 8.162 [±0.208] compared to 7.734 [±0.237] 10 years post burn-in and G_{10} = 11.881 [±0.272] compared to 11.313 [±0.323] 20 years post burn-in; **Figure 8**, **Supplementary Material [Table S2 File S4]**). After 20 years, UCPC-He* = 0.01 outperformed OCS-He* = 0.01 in eight out of 10 replicates (on average, G_{10} = 16.398 [±0.426] compared to 15.850 [±0.384] 40 years post burn-in and G_{10} = 18.161 [±0.470] compared to 17.528 [±0.438] 60 years post burn-in; **Figure 8**, **Supplementary Material [Table S2 File S4]**). Observations on the genic variance (σ_a^2) and genetic variance (σ_A^2) were consistent as well. We also observed that UCPC-He* = 0.01 yielded a lower number of QTLs where the favorable allele was lost (on average, 218.8 [±3.852]) compared to OCS-He* = 0.01 (on average, 234.5 [±3.908]) (**Figure 8**). PM not considering the marker information, i.e., phenotypic selection (PS scenario), yielded lower short- and long-term genetic gains than PM considering marker information (GS scenario) (on average, G_{10} = 6.402 [±0.166] compared to 7.713 [±0.256] 10 years post burn-in and G_{10} = 10.810 [±0.329] compared to 13.287 [±0.436) 60 years post burn-in; **Figure 8**, **Supplementary Material [Table S2 File S4]**).

DISCUSSION

Predicting the Next-Generation Diversity

Accounting for within-family selection increased the squared correlation and reduced the mean error of post-selection genetic diversity prediction (**Figures 3**, **4**). The gain in squared correlation (**Figure 3**) and the reduction in mean error (**Figure 4**), were more important for parents showing differences in performance. This result is consistent with observations in Allier et al. (2019b), where crosses between two phenotypically distant parents yielded post-selection parental contributions that differ from their expectation before selection (i.e., 0.5). The mean prediction error was always positive, which can be explained by the use in Eq. 9 of genome-wide parental contributions to progeny in lieu of parental contributions at individual QTLs to predict allelic frequency changes due to selection **Supplementary Material (File S2)**. As a result, the predicted extreme frequencies at QTLs in the progeny are shrunk toward the mean frequency, leading to an overestimation of the

TABLE 2 | Discounted cumulative gain in TRUE scenario for three different parameters ρ giving more weight to short-term gain in different levels and assuming known QTL effects (TRUE scenario).

Cross-selection index (CSI)	Discounted cumulative gain		
	ρ = 0	ρ = 0.04	ρ = 0.2
UCPC - He* = 0.01	15.949 (±0.398)	12.321 (±0.284)	6.682 (±0.143)
UCPC - He* = 0.10	15.174 (±0.386)	11.788 (±0.280)	6.593 (±0.158)
UC	14.408 (±0.355)	11.689 (±0.266)	6.822 (±0.145)
OCS - He* = 0.01	15.148 (±0.346)	11.675 (±0.262)	6.360 (±0.149)
OCS - He* = 0.10	14.630 (±0.349)	11.278 (±0.264)	6.230 (±0.149)
UCPC - He* = 0.15	14.205 (±0.334)	11.176 (±0.250)	6.454 (±0.149)
OCS - He* = 0.15	14.056 (±0.337)	10.884 (±0.250)	6.103 (±0.155)
PM	12.609 (±0.280)	10.392 (±0.217)	6.345 (±0.155)

Mean discounted cumulative gain with ρ = 0 (constant weight along years), ρ = 0.04 (decreasing weight along years) and ρ = 0.2 (nearly null weights after 30 years) on the ten independent replicates. CSI are ordered in decreasing discounted cumulative gain with ρ = 0.04.

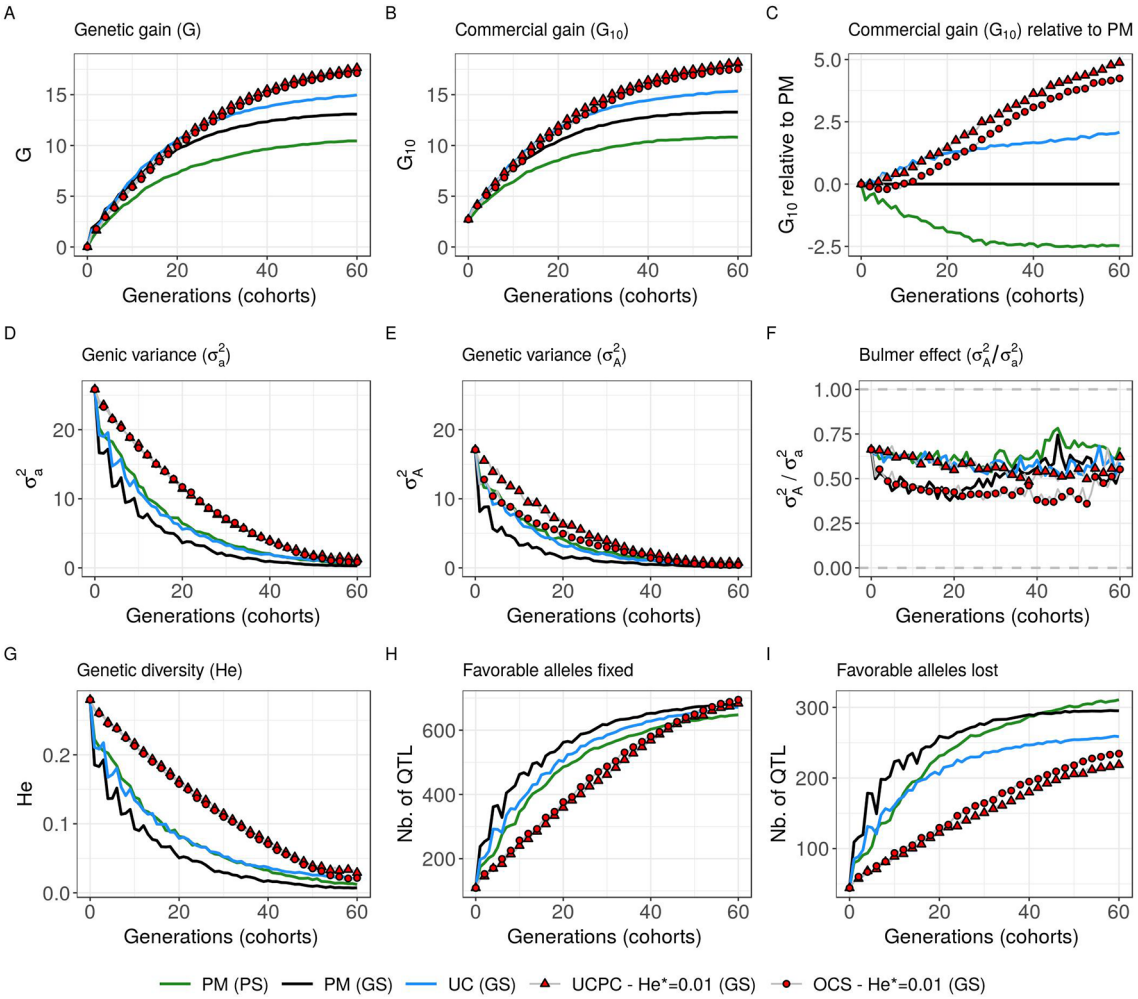


FIGURE 8 | Evolution of different variables for different cross-selection indices according to the generations in the GS scenario (PM, parental mean; UC, usefulness criterion; OCS-He*, optimal cross-selection; and UCPC-He*, UCPC-based optimal cross-selection for He* = 0.01) and in the PS scenario (PM, parental mean). **(A)** Genetic gain at whole progeny level (G), **(B)** genetic gain at commercial level (G_{10}), and **(C)** G_{10} relatively to PM (GS), genetic gain is measured on true breeding values. **(D)** Genic variance at QTLs (σ^2_a). **(E)** genetic variance of true breeding values (σ^2_A) and **(F)** ratio of genic over genetic variance (σ^2_A / σ^2_a). **(G)** genetic diversity at QTLs and number of QTLs where the favorable allele was fixed **(H)** and lost **(I)**.

expected heterozygosity (He) (results not shown). Local changes in allele frequency under artificial selection could be predicted following Falconer and Mackay (1996) and Gallais et al. (2007), but this approach would assume linkage equilibrium between QTLs, which is a strong assumption that does not correspond to the highly polygenic trait that we simulated.

Effect of UC on Short- and Long-Term Recurrent Selection

In a first approach, we considered no constraint on diversity during cross-selection and compared cross-selection maximizing the UC or maximizing the PM in the TRUE scenario, assuming known QTL effects and positions. The UC yielded higher short-term genetic gain at commercial level (G_{10}; **Figures 5B**, **C**). This was expected because UC predicts the mean performance of the best fraction of progeny. When considering the genetic gain at the mean progeny level (G; **Figure 5A**), UC needed 5 years to outperform PM. These results underline that UC maximizes the mean performance of the next generation issued from the intercross of selected progeny, sometimes at the expense of the current generation progeny mean performance. This observation is consistent with the fact that candidate parents of the sixth cohort came all from the three first cohorts generated considering UC and thus the sixth cohort took full advantage of the use of UC (**Figure 1A**). This tendency was also observed in simulations by Müller et al. (2018) considering the EMBV approach, akin to the UC for normally distributed additive traits. The UC also showed a higher long-term genetic gain at both commercial (G_{10}) and whole progeny level (G) compared to intercrossing the best candidate parents (PM). This long-term gain was driven by a higher additive genic variance at QTLs (σ^2_a ; **Figure 6A**) and a lower genomic covariance between QTLs (σ^2_A / σ^2_a ; **Figure 6C**) resulting in a higher additive genetic

variance in UC compared to PM (σ_A^2 ; **Figure 6B**). Note that with lower σ_a^2 the ratio σ_A^2 / σ_a^2 becomes less interpretable in the long-term (**Figure 6C**). UC also better managed the fixation (**Figure 7B**) or the maintenance (**Figure 7C**) of the favorable allele at QTLs compared to PM. These results highlight the interest of considering within cross variance in cross-selection for improving long-term genetic gain as observed in Müller et al. (2018).

Accounting for Within-Family Variance in Optimal Cross-Selection

Assuming known marker effects, we observed that considering a constraint on diversity, i.e., optimal cross-selection, always maximized the long-term genetic gain, at the cost of a variable penalty for short-term gain, compared to no constraint on diversity (e.g., UC). We further compared the OCS (Gorjanc et al., 2018) with the UCPC-based optimal cross-selection that accounts for the fact that only a selected fraction of each family contributes to the next generation. In the optimization framework considered, we compared the ability of UCPC (referred to as UCPC-He*) and OCS (referred to as OCS-He*) to convert a determined loss of diversity into genetic gain. For a given diversity trajectory, UCPC-He* yielded higher short-term commercial gain than OCS-He*. Both, OCS-He* and UCPC-He* yielded similar additive genic variance (σ_a^2), but we observed differences in terms of the ratio σ_A^2 / σ_a^2. As expected under directional selection, the ratio σ_A^2 / σ_a^2 was positive and inferior to one, revealing a negative genomic covariance between QTLs (Bulmer, 1971). UCPC-He* yielded a higher ratio, i.e., lower repulsion, and thus a higher additive genetic variance (σ_A^2) than OCS-He* for a similar He*. This explains the higher long-term genetic gain at commercial and whole progeny levels observed for UCPC-He*. This result supports the idea, suggested in Allier et al. (2019a), that accounting for complementarity between parents when defining crossing plans is an efficient way to favor recombination events to reveal part of the additive genic variance hidden by repulsion between QTLs. For low targeted diversity (He* = 0.01), UCPC-He* also appeared to better manage the rare favorable alleles at QTLs than OCS-He*. These results highlighted the interest of UCPC-based optimal cross-selection to convert the genetic diversity into genetic gain by maintaining more rare favorable alleles and limiting repulsion between QTLs. In case of higher targeted diversity (He* = 0.15), the loss of diversity was likely not sufficient to fully express the additional interest of UCPC compared to OCS to convert diversity into genetic gain. In this case, UCPC-He* and OCS-He* performed similarly. Accounting for within cross variance to measure the expected gain of a cross in optimal cross-selection was already suggested in Shepherd and Kinghorn (1998). More recently, Akdemir and Isidro-Sánchez (2016) and Akdemir et al. (2018) accounted for within cross variance considering linkage equilibrium between QTLs. Akdemir and Isidro-Sánchez (2016) also observed that accounting for within cross variance during cross-selection yielded higher long-term mean performance with a penalty at short-term mean progeny performance.

Short-term economic returns of a breeding program condition the resources invested to maintain/increase response to selection and therefore long-term competitive capacity. Hence, to fully take advantage of their benefit at long term, it is necessary to make sure that tested breeding strategies do not compromise too much the short-term commercial genetic gain. For this reason, we considered the discounted cumulative commercial gain following Dekkers et al. (1995) and Chakraborty et al. (2002) as a summary variable to evaluate CSI while giving more weight to short-term gain in different levels. UCPC-He* outperformed OCS-He* for a given He* either considering uniform weights ($\rho = 0$) or giving approximately seven times more weight to short-term gain compared to long-term gain ($\rho = 0.04$). This was also true when focusing only on short-term gain ($\rho = 0.2$), but in this case the best model was UC without accounting for diversity (**Table 2**).

Practical Implementations in Breeding

UCPC With Estimated Marker Effects

In simulations, we first considered 1,000 QTLs with known additive effects sampled from a centered normal distribution. For a representative subset of CSIs (PM, UC, UCPC-He*, and OCS-He* with He* = 0.01; **Figure 8**), we considered estimated effects at 2,000 SNPs. The main conclusions obtained with known and estimated marker effects were consistent, supporting the practical interest of UCPC-based optimal cross-selection (**Figure 8**). The difference was that the superiority of UCPC-based optimal cross-selection over optimal cross-selection not accounting for within-family selection in GS scenario was not significant after 60 years **Supplementary Material (Table S2 File S4)**. With estimated marker effects instead of known QTL effects, the predicted progeny variance (σ^2) corresponded to the variance of the predicted breeding values, which are shrunk compared to TBVs, depending on the model accuracy (referred to as variance of posterior mean [VPM] in Lehermeier et al.). An alternative would be to consider the marker effects estimated at each sample of a Monte Carlo Markov Chain process, e.g., using a Bayesian ridge regression, to obtain an improved estimate of the additive genetic variance (referred to as posterior mean variance [PMV] in Lehermeier et al., 2017a; Lehermeier et al., 2017b).

In practice, QTL effects are unknown, so the selection of progeny cannot be based on TBVs, and thus the selection accuracy (h) is smaller than one. In our simulation study assuming unknown QTLs (GS scenario), progeny were selected based on estimated breeding values taking into account genotypic information as well as replicated phenotypic information, which led to a high selection accuracy, as it can be encountered in breeding. Thus, the assumption $h = 1$ used in Eq. 6 for GS scenario is reasonable. In order to shorten the cycle length of the breeding scheme, selection of progeny can be based on predicted GEBVs of genotyped but not phenotyped progeny. In such a case, the selection accuracy (h) will be considerably reduced. In such a situation, one can advocate to use PMV instead of VPM in the computation of UCPC and to take into account the proper selection accuracy (h) within crosses adapted to the selection scheme. When selection is based on predicted values, i.e.,

genotyped but not phenotyped progeny, the shrunk predictor VPM should be a good approximation of $(h\sigma)^2$.

UCPC-Based Optimal Cross-Selection

In this study, we assumed fully homozygous parents and two-way crosses. However, neither the optimal cross-selection nor UCPC-based optimal cross-selection is restricted to homozygote parents. Considering heterozygote parents in optimal cross-selection is straightforward. Following the extension of UCPC to four-way crosses (Allier et al., 2019b), UCPC optimal cross-selection can be used for phased heterozygous individuals, as it is commonly the case in perennial plants or animal breeding. Animal breeders are interested in Mendelian sampling variance for individual and cross-selection (Segelke et al., 2014; Bonk et al., 2016; Bijma et al., 2018) and might be interested to incorporate it into OCS strategies. We considered an inbred line breeding program, but the extension to hybrid breeding is of interest for species such as maize. The use of testcross effects, i.e., estimated on hybrids obtained by crossing candidate lines with lines from the opposite heterotic pool, in UCPC-based optimal cross-selection is straightforward, and so the UCPC-based optimal cross-selection can be used to improve each heterotic pool individually. In order to jointly improve two pools, further investigations are required to include dominance effects in UCPC-based optimal cross-selection. In addition, this would imply that crossing plans in both pools are jointly optimized to manage genetic diversity within pools and complementarity between pools.

We considered a within-family selection intensity corresponding to the selection of the 5% most performant progeny as candidates for the next generation. Equal selection intensities were assumed for all families, but in practice due to experimental constraints or optimized resource allocation (e.g., generate more progeny for crosses showing high progeny variance but low progeny mean), within-family selection intensity can be variable. Different within-family selection intensities (see Eqs. 8 and 9) can be considered in UCPC-based optimal cross-selection, but an optimization regarding resource allocation of the number of crosses and the selection intensities within crosses calls for further investigations. However, in marker-assisted selection schemes based on QTL detection results (Bernardo et al., 2006), an optimization of selection intensities per family was observed to be only of moderate interest.

Proposed UCPC-based optimal cross-selection was compared to OCS in a targeted diversity trajectory context. We considered a linear trajectory, but any genetic diversity trajectory can be considered (e.g., **Figure 2**). The optimal diversity trajectory cannot be easily determined and depends on breeding objectives and data considered. Optimal contribution selection in animal breeding considers a similar ϵ-constraint optimization with a targeted inbreeding trajectory determined by a fixed annual rate of inbreeding (e.g., 1% advocated by the Food and Agriculture Organization (FAO), Woolliams et al., 1998). Woolliams et al. (2015) argued that the optimal inbreeding rate is also not straightforward to define. An alternative formulation of the optimization problem to avoid the use of a fixed constraint is to consider a weighted index $(1-\alpha)V(\boldsymbol{nc})+\alpha D(\boldsymbol{nc})$, where α is the weight balancing the expected gain $V(\boldsymbol{nc})$ and constraint $D(\boldsymbol{nc})$ (De Beukelaer et al., 2017). However, the appropriate choice of α is difficult and is not explicit either in terms of expected diversity or expected gain.

Introgression of Diversity and Anticipation of a Changing Breeding Context

We considered candidate parents coming from the three last overlapping cohorts (**Figure 1**) in order to reduce the number of candidate crosses during the progeny covariances prediction (UCPC) and the optimization process. This yielded elite candidate parents that were not directly related (no parent–progeny) and that did not show strong differences in performances, which is standard in a commercial plant breeding program focusing on yield improvement. However, when the genetic diversity in a program is so low that long-term genetic gain is compromised, external genetic resources need to be introgressed by crosses with internal elite parents. As suggested by results of simulation 1, we conjecture that the advantage of UCPC-based optimal cross-selection over OCS increases in such a context where heterogeneous, i.e., phenotypically distant, genetic materials are crossed. This requires investigations that we hope to address in subsequent research.

Our simulations also assumed fixed environments and a single targeted trait over 60 years. However, in a climate change context and with rapidly evolving societal demands for sustainable agricultural practices, environments and breeders objectives will likely change over time. In a multitrait context, the multiobjective optimization framework proposed in Akdemir et al. (2018) can be adapted to UCPC-based optimal cross-selection. The upcoming but yet unknown breeding objectives make the necessity to manage genetic diversity even more important than highlighted in this study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.25387/g3.7405892.

AUTHOR CONTRIBUTIONS

ST, CL, AC, and LM supervised the study. AA performed the simulations and wrote the manuscript. ST worked on the implementation in the simulator. All authors reviewed and approved the manuscript.

FUNDING

This research was funded by RAGT2n and the ANRT CIFRE grant no. 2016/1281 for AA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01006/full#supplementary-material

REFERENCES

Akdemir, D., and Isidro-Sánchez, J. (2016). Efficient breeding by genomic mating. *Front. Genet.* 7, 210. doi: 10.3389/fgene.2016.00210

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2018). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672. doi: 10.1101/209080

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Moreau, L., Charcosset, A. (2019a). Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. *Theor. Appl. Genet*. 132, 1321–1334. doi: 10.1007/s00122-019-03280-w

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

Bernardo, R., Moreau, L., and Charcosset, A. (2006). Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. *Crop Sci.* 46, 1972–1980. doi: 10.2135/cropsci2006.01-0057

Bijma, P., Wientjes, Y. C. J., and Calus, M. P. L. (2018). Increasing genetic gain by selecting for higher Mendelian sampling variance. *Proc. World Congr. Genet. Appl. Livest. Prod. Genet. Gain-Breed. Strategies* 2, 47.

Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.* 48, 36. doi: 10.1186/s12711-016-0214-0

Bulmer, M. (1971). The stability of equilibria under selection. *Heredity* 27, 157–162. doi: 10.1038/hdy.1971.81

Bulmer, M. (1980). *The mathematical theory of quantitative genetics*. Newyork Oxford University Press.

Chakraborty, R., Moreau, L., and Dekkers, J. C. (2002). A method to optimize selection on multiple identified quantitative trait loci. *Genet. Sel. Evol.* 34, 145. doi: 10.1186/1297-9686-34-2-145

Clark, S. A., Hickey, J. M., and van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol. GSE* 43, 18. doi: 10.1186/1297-9686-43-18

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J.M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

De Beukelaer, H. D., Badke, Y., Fack, V., and Meyer, G. D. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genet.* 206: 1127–1138. doi: 10.1534/genetics.116.194449

Dekkers, J. C. M., Birke, P. V., and Gibson, J. P. (1995). Optimum linear selection indexes for multiple generation objectives with non-linear profit functions. *Anim. Sci.* 61, 165–175. doi: 10.1017/S1357729800013667

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics. 4th ed*. Harlow, England: Pearson.

Fradgley, N., Gardner, K. A., Cockram, J., Elderfield, J., Hickey, J. M., Howell, P., et al. (2019). A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. *PLoS Biol.* 17, e3000071. doi: 10.1371/journal.pbio.3000071

Gallais, A., Moreau, L., and Charcosset, A. (2007). Detection of marker–QTL associations by studying change in marker frequencies with selection. *Theor. Appl. Genet.* 114, 669–681. doi: 10.1007/s00122-006-0467-z

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6, e28334. doi: 10.1371/journal.pone.0028334

Gerke, J. P., Edwards, J. W., Guill, K. E., Ross-Ibarra, J., and McMullen, M. D. (2015). The genomic impacts of drift and selection for hybrid performance in maize. *Genetics* 201, 1201–1211. doi: 10.1534/genetics.115.182410

Giraud, H., Lehermeier, C., Bauer, E., Falque, M., Segura, V., Bauland, C., et al. (2014). Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic qtl for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* 198, 1717–1734. doi: 10.1534/genetics.114.169367

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

Gorjanc G., and Hickey J. M. (2018) AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics* 34, 3408–3411. doi: 10.1093/bioinformatics/bty375

Haimes, Y., Lasdon, L. S., and Wimer, D. (1971). On a bicriterion formation of the problems of integrated system identification and system optimization. *IEEE Trans. Syst. Man Cybern.* SMC-1, 296–297. doi: 10.1109/TSMC.1971.4308298

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLoS Genet.* 6, e1001139. doi: 10.1371/journal.pgen.1001139

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Guelph: University of Guelph.

Heslot, N., Jannink, J.-L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Sci.* 55, 1–12. doi: 10.2135/cropsci2014.03.0249

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42, 35. doi: 10.1186/1297-9686-42-35

Kinghorn, B. P., Banks, R., Gondro, C., Kremer, V. D., Meszaros, S. A., Newman, S., et al. (2009). "Strategies to exploit genetic variation while maintaining diversity," in *adaptation and fitness in animal populations* (Dordrecht: Springer), 191–200. doi: 10.1007/978-1-4020-9005-9_13

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genet. Sel. Evol.* 43, 4. doi: 10.1186/1297-9686-43-4

Lehermeier, C., de los Campos, G., Wimmer, V., and Schön, C.-C. (2017a). Genomic variance estimates: with or without disequilibrium covariances? *J. Anim. Breed. Genet.* 134, 232–241. doi: 10.1111/jbg.12268

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

Lin, Z., Cogan, N. O. I., Pembleton, L. W., Spangenberg, G. C., Forster, J. W., Hayes, B. J. et al., (2016). Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. *Plant Genome* 9. doi: 10.3835/plantgenome2015.06.0046

Lin, Z., Shi, F., Hayes, B. J., and Daetwyler, H. D. (2017). Mitigation of inbreeding while preserving genetic gain in genomic breeding programs for outbred plants. *Theor. Appl. Genet.* 130, 969–980. doi: 10.1007/s00122-017-2863-y

Lynch, M., and Walsh, B. (1999). *Evolution and selection of quantitative traits*. Sunderland, MA., Sinauer Associates.

Meuwissen, T. H. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* 75, 934–940. doi: 10.2527/1997.754934x

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

Misztal, I. (2008). Reliable computing in estimation of variance components. *J. Anim. Breed. Genet.* 125, 363–370. doi: 10.1111/j.1439-0388.2008.00774.x

Mohammadi, M., Tiede, T., and Smith, K. (2015). PopVar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55, 2068–2077. doi: 10.2135/cropsci2015.01.0030

Müller, D., Schopp, P., and Melchinger, A. E. (2018). Selection on expected maximum haploid breeding values can increase genetic gain in recurrent genomic selection. *G3 Genes Genomes Genet*. 8, 1173–1181. doi: 10.1534/g3.118.200091

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *J. Dairy Sci.* 95, 377–388. doi: 10.3168/jds.2011-4254

Pszczola, M., Strabel, T., Mulder, H. A., and Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within

and to the reference population. *J. Dairy Sci.* 95, 389–400. doi: 10.3168/jds.2011-4338

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rauf, S., Teixeira da Silva, J. A., Khan, A. A., and Naveed, A. (2010). Consequences of plant breeding on genetic diversity. *Int. J. Plant Breed.* 4, 1–21.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theor. Appl. Genet.* 132, 81–96. doi: 10.1007/s00122-018-3196-1

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L. et al., (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* 8. doi: 10.3835/plantgenome2014.10.0074

Schnell, F., and Utz, H. (1975). "F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern," in *Bericht* über *die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter* (Austria: BAL Gumpenstein), 243–248.

Segelke, D., Reinhardt, F., Liu, Z., and Thaller, G. (2014). Prediction of expected genetic variation within groups of offspring for innovative mating schemes. *Genet. Sel. Evol.* 46, 42. doi: 10.1186/1297-9686-46-42

Shepherd, R. K., and Kinghorn, B. P. (1998). A tactical approach to the design of crossbreeding programs, in *Proceedings of the sixth world congress on genetics applied to livestock production: 11-16 january*, (Armidale) 431–438.

Storn, R., and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* 11, 341–359. doi: 10.1023/A:1008202821328

Van Inghelandt, D., Reif, J. C., Dhillon, B. S., Flament, P., and Melchinger, A. E. (2011). Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theor. Appl. Genet.* 123, 11–20. doi: 10.1007/s00122-011-1562-3

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* 94, 73–83. doi: 10.1017/S0016672312000274

Woolliams, J. A., Gwaze, D. P., Meuwissen, T. H., Planchenault, D., Renard, J. P., Thibier, M., et al. (1998). Secondary guidelines for the development of national farm animal genetic resources management plans. *Manage. Small Popul. Risk.*

Woolliams, J. A., Berg, P., Dagnachew, B. S., and Meuwissen, T. H. E. (2015). Genetic contributions and their optimization. *J. Anim. Breed. Genet.* 132, 89–99. doi: 10.1111/jbg.12148

Wray, N., and Goddard, M. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26, 431. doi: 10.1186/1297-9686-26-5-431

Zhong, S., and Jannink, J.-L. (2007). Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics* 177, 567–576. doi: 10.1534/genetics.107.075358

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Allier, Lehermeier, Charcosset, Moreau and Teyssèdre. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fgene.2019.00954

Last-Generation Genome–Environment Associations Reveal the Genetic Basis of Heat Tolerance in Common Bean (*Phaseolus vulgaris* L.)

Felipe López-Hernández[1,2]* *and Andrés J. Cortés*[1,3]

[1] Corporación Colombiana de Investigación Agropecuaria (Agrosavia) - Cl La Selva, Rionegro, Colombia, [2] Facultad de Ciencias – Grupo de Investigación en Sistemática Molecular, Universidad Nacional de Colombia - Sede Medellín, Medellín, Colombia, [3] Facultad de Ciencias Agrarias - Departamento de Ciencias Forestales, Universidad Nacional de Colombia - Sede Medellín, Medellín, Colombia

Genome–environment associations (GEAs) are a powerful strategy for the study of adaptive traits in wild plant populations, yet they still lack behind in the use of modern statistical methods as the ones suggested for genome-wide association studies (GWASs). In order to bridge this gap, we couple GEA with last-generation GWAS algorithms in common bean to identify novel sources of heat tolerance across naturally heterogeneous ecosystems. Common bean (*Phaseolus vulgaris* L.) is the most important legume for human consumption, and breeding it for resistance to heat stress is key because annual increases in atmospheric temperature are causing decreases in yield of up to 9% for every 1°C. A total of 78 geo-referenced wild accessions, spanning the two gene pools of common bean, were genotyped by sequencing (GBS), leading to the discovery of 23,373 single-nucleotide polymorphism (SNP) markers. Three indices of heat stress were developed for each accession and inputted in last-generation algorithms (*i.e.* SUPER, FarmCPU, and BLINK) to identify putative associated loci with the environmental heterogeneity in heat stress. Best-fit models revealed 120 significantly associated alleles distributed in all 11 common bean chromosomes. Flanking candidate genes were identified using 1-kb genomic windows centered in each associated SNP marker. Some of these genes were directly linked to heat-responsive pathways, such as the activation of heat shock proteins (*MED23*, *MED25*, *HSFB1*, *HSP40*, and *HSP20*). We also found protein domains related to thermostability in plants such as *S1* and *Zinc finger A20* and *AN1*. Other genes were related to biological processes that may correlate with plant tolerance to high temperature, such as time to flowering (*MED25*, *MBD9*, and *PAP*), germination and seedling development (*Pkinase_Tyr*, *Ankyrin-B*, and *Family Glicosil-hydrolase*), cell wall stability (*GAE6*), and signaling pathway of abiotic stress via abscisic acid (histone-like transcription factors *NFYB* and *phospholipase C*) and auxin (*Auxin response factor* and *AUX_IAA*). This work offers putative associated loci for marker-assisted and genomic selection for heat tolerance in common bean. It also demonstrates that it is feasible to identify genome-wide environmental associations with modest sample sizes by using a combination of various carefully chosen environmental indices and last-generation GWAS algorithms.

Keywords: heat stress, local adaptation, genome-wide association studies (GWAS), environmental indices, SUPER, FarmCPU, BLINK

OPEN ACCESS

Edited by:
Nunzio D'Agostino, University of Naples Federico II, Italy

Reviewed by:
Ali Soltani, Michigan State University, United States
Atena Oladzadabbasabadi, North Dakota State University, United States

***Correspondence:**
Felipe López-Hernández
llopez@agrosavia.co

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *25 June 2019*
Accepted: *06 September 2019*
Published: *22 November 2019*

Citation:
López-Hernández F and Cortés AJ (2019) Last-Generation Genome–Environment Associations Reveal the Genetic Basis of Heat Tolerance in Common Bean (Phaseolus vulgaris L.). Front. Genet. 10:954. doi: 10.3389/fgene.2019.00954

INTRODUCTION

Exploring the genetic basis of adaptive traits in wild plant populations has been accelerated by modern genomic strategies such as genome–phenotype [genome-wide association study (GWAS)] and genome–environment association (GEA) studies (Frank et al., 2016). GEA commonly associates single-nucleotide polymorphisms (SNPs) and environmental variables based on the accessions' sampling site in order to infer adaptation to abiotic stress. Genotyping by sequencing (GBS) has in turn been revealed as one of the best methods for GEA due to its potential to discover a considerable amount of SNP markers throughout the genome. For instance, coupling GEA and GBS recently allowed identifying adaptive variation for drought tolerance (Cortés and Blair, 2018). However, despite the fact that the GEA framework uses the latest genomic tools available, it has not yet taken full advantage of newer and more promising statistical approaches to detect genomic signatures of environmental adaptation while controlling for confounding effects.

GEA studies often rely on GWAS models, which typically couple mixed linear models (MLMs) (Zhang et al., 2010) with kinship and population structure analyses in order to correct for false positives. Yet new GWAS algorithms have recently been developed to gain statistical power to detect associated markers, increase efficiency, and decrease computational complexity (Wang et al., 2014b). The strategy to reconstruct the kinship matrix is the most relevant difference between recent methods of individual marker tests such as Factored Spectrally Transformed Linear Mixed Model (FaST-LMM-Select), Compressed MLM (CMLM) (Li et al., 2014), and Settlement of MLM Under Progressively Exclusive Relationship (SUPER), the latter being the most statistically powerful (Wang et al., 2014b; Liu et al., 2016). SUPER drastically reduces the amount of genetic markers used to infer kinship relationships by dividing the SNP dataset into *bins* (Wang et al., 2014b). Most influential *bins*, known as pseudo-nucleotides of quantitative rank underlying the phenotype (PseudoQTNs), are then optimized in size and number using maximum likelihood and linkage disequilibrium (LD). On the contrary, FaST-LMM-Select chooses SNPs to infer kinship relationships based only on a physical distance criterion, while CMLM uses kinship estimates between pairs of groups clustered based on their kinship value in order to reduce the size of the fixed effect and increase the computational power. Tests of multiple loci such as the multi-locus mixed model (MLMM) (Segura et al., 2012) have been developed, too. Both strategies, individual markers (CMLM, FaST-LMM-Select, and SUPER) and multiple loci (MLMM) tests, effectively control the false-positive rate. Yet these algorithms have a higher rate of false negatives after the partition imposed on the SNP dataset to recreate the kinship matrix.

Alternative methods such as Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016) and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Huang et al., 2019) have been developed to control both the false-positive rate and the confounding variable that disfavors the real associations. FarmCPU and BLINK divide a typical MLMM into two parts that are used iteratively, a fixed effect model (FEM) and a random effect model (REM). BLINK replaces restricted maximum likelihood (REML) in FarmCPU's REM with Bayesian information criteria (BIC) in a FEM. Additionally, BLINK uses LD information to replace the bin method. SUPER, FarmCPU, and BLINK can therefore be considered last-generation GWAS models. These powerful algorithms, already tested for conventional GWAS, are promising to identify adaptive loci under a GEA framework.

In turn, the potential of GEA studies to identify new sources of tolerance to abiotic stresses is undeniable (Cortés and Blair, 2018) and could aid the study of the genetic adaptation to adverse conditions that have not previously been approached from a GEA perspective, which is the case of heat stress (HS). Annual increases in atmospheric average temperature have been responsible for yield losses of 9% for every 1°C across the vast majority of agricultural species. This situation is likely to worsen as by 2100 global average temperature is estimated to be 3°C above the present value (Abrol and Ingram, 1996), jeopardizing worldwide yields.

Common bean (*Phaseolus vulgaris* L.), a not perennial (Gentry, 1969), is one of the most produced legumes with ~27 million tons worldwide, China and America being the main producers (FAO, 2018), yet tolerance to HS is generally low in this species. Beans are nutritionally rich due to their high content of proteins, folic acid, iron, dietary fiber, and complex carbohydrates and constitute a main alimentary supply for communities in Latin America, Africa, and Asia (Sgarbieri and Whitaker, 1982; Pachico, 1993). Since these regions are also highly vulnerable to HS, increased atmospheric average temperature would impact not only yields in small-scale farms but also human nutrient intake *via* common bean (Jones, 1999). Most common bean varieties used by farmers are better adapted to regions of medium to high elevations or to sowing times during the colder seasons in tropical areas (Porch and Jahn, 2001). Some authors have reported optimal temperatures between 18°C and 20°C (Wantanbe, 1953; Qi et al., 1998; Porch, 2006; Rosas et al., 2000) for the cultivation of this legume. The reproductive phase is the most sensitive phenological stage to HS, with temperatures above 28°C to 32°C (Gonçalves et al., 1997; Caramori et al., 2001; Silva et al., 2007; Rainey and Griffiths, 2019) decreasing the number of pods and seeds and therefore reducing yield (Weaver and Timm, 1988; Monterroso and Wien, 2019). In order to compensate for yield losses due to low tolerance of cultivated common bean to high temperatures, a prompt characterization of the genetic sources of HS tolerance in wild populations is needed.

Nowadays, there is a lack of knowledge on how the most recent GWAS models work under a GEA paradigm. Additionally, there is an urgent need to identify loci linked to HS tolerance in wild common bean germplasm collections, which would aid the development of common bean varieties resistant to high temperatures. Therefore, for this study, we set the following objectives: (1) synthetize environmental variables in order to estimate HS tolerance in wild common bean germplasm collections, which would allow identifying tolerant accessions; (2) explore the utility of the most promising modern GWAS models (CMLM, SUPER, FarmCPU, and BLINK) for GEA

studies; and (3) implement GEA models with last-generation GWAS algorithms in order to capture adaptive genetic variation to HS, candidate to be integrated into common bean breeding programs. This first exploration of the environmental adaptation of wild common bean to HS will ultimately offer putative associated loci for marker-assisted and genomic selection strategies by using a combination of various well-chosen environmental indices and last-generation GWAS algorithms, while testing the utility of the latter under a GEA paradigm.

MATERIALS AND METHODS

Plant Material and GBS

The present work was developed with a total of 78 accessions of wild common bean. All genotypes were transferred by the Genetic Resources Unit of the International Center for Tropical Agriculture (CIAT) and are conserved under the genetic resources treaty of the Food and Agriculture Organization of the United Nations (FAO collection). The accessions are a representative sample of groups of genes and races, the selection being based on core collections for wild bean samples according to Tohme et al. (1996). Despite adaptation to environmental stress conditions evolved differently in the two gene pools of common bean (Soltani et al., 2017; Soltani et al., 2018; Oladzad et al., 2019), we carried out the GEA models including both gene pools in order to maximize the statistical power to detect significantly associated markers by increasing (1) the number of wild accessions and (2) the environmental contrast (Mesoamerican environments of wild common bean typically experience more heat events than Andean environments, **Figure S8**). Georeferencing was provided by the Genetic Resources Unit at CIAT (**Table S1**).[1]

Processing of plant material, genomic DNA extraction, GBS library preparation using the *Apek1* enzyme (Cortés and Blair, 2018), and sequencing and bioinformatic processing for the 78 accessions were carried out as described by Cortés and Blair (2018), following Elshire et al. (2011) and Bradbury et al. (2007) and using as reference genome the common bean assembly (Schmutz et al., 2014). SNP markers with missing data that exceeded 20% or frequency of the minor allele (MAF) that did not exceed 5% were excluded from the GEA dataset in the 78 genotyped accessions in order to finally obtain a matrix of 23,373 SNP markers with an average depth of 13.6 X.

Compilation of Bioclimatic Data and HS Indices

In order to estimate heat tolerance for wild common bean, we extracted from the WorldClim[2] database, at a 2.5-min resolution, environmental variables using the georeferencing of each accession. A total of six bioclimatic variables, putatively related with HS, were considered, as follows: BIO1 = annual mean temperature, BIO5 = maximum temperature of warmer month, BIO8 = mean temperature of the wettest quarter, BIO9 = mean temperature of the driest quarter, BIO10 = mean temperature of the warmest 4-month period, and T_j = average of absolute maximum temperature during the reproductive phase. Extraction was carried out using the *dismo* package of R v.3.4.4 (R Core Team). Historical temperature values were obtained as monthly averages from 1970 to 2000. Values of each bioclimatic variable were adjusted for the year 2000 according to the average annual increase in temperature for each hemisphere, using the following expressions (Trenberth et al., 2007):

$$\mathrm{T}_{2000} = \mathrm{Ti} + \left(2000 - i\right) \times 0.031675\,[^{\circ}\mathrm{C}]\ \text{for the Northern Hemisphere} \quad (1)$$

$$\mathrm{T}_{2000} = Ti + \left(2000 - \mathrm{i}\right) \times 0.01325\,[^{\circ}\mathrm{C}]\ \text{for the Southern Hemisphere} \quad (2)$$

where i is the year of collection of each accession, Ti is the bioclimatic variable for the year when the accession was collected, and T_{2000} is the value of each bioclimatic variable for the year 2000.

We generated three indices based on environmental data from wild common bean accessions in order to understand natural adaptation to high temperatures and identify associated genetic markers. The first index was built using the evapotranspiration model from (Thornthwaite, 1948), which contained an expression for monthly heat index, heat index Thornthwaite (HIT), as follows (equation 3):

$$HIT_{original} = \sum_{i=1}^{k} \frac{T_m^{1.514}}{5} \quad (3)$$

For all $T_{\mathrm{m}} > 0$, T_{m} is the average mean monthly temperature in any phenological stage of the plant, and k the number of months.

This index (HIT_{original}) uses average temperature (T_{m}) and not maximum temperature (T_j), despite the latter being more informative for HS events. Thus, we used two adjustments to refine HIT_{original}. First, we used the absolute maximum temperature instead of the average temperature. Second, we narrowed the window of temperatures only across the reproductive phase (T_j), during which plants are most sensitive to HS events (Rainey and Griffiths, 2019). Since seeds were collected for each accession as part of the original sampling, the reproductive phase has an approximate duration of 2 months prior to the month when sampling took place. The modified HIT_{original} index was expressed through the following equation:

$$HIT = \sum_{i=1}^{2} \frac{T_j^{1.514}}{5} \quad (4)$$

For all $T_j > 0$, T_j is the average of absolute maximum temperature during the reproductive phase and i is the month within that phase.

On the other hand, we built a second index of HS, heat stress index (HSI), as detailed in equation 5. This index is based on the temperature threshold during the reproductive phase (T_{max} = 28–32°C) above which common bean exhibits low grain yields (Gonçalves et al., 1997; Caramori et al., 2001; Silva et al., 2007; Rainey and Griffiths, 2019). Therefore, this suggested HS index

[1] http://genebank.ciat.cgiar.org/genebank/main.do
[2] http://www.worldclim.org

compares T_{max} = 30°C and the maximum temperature during the reproductive phase T_j adjusted for the year 2000.

$$HSI = \left(\frac{T_j - 30}{30} \right) \times 100; -100 \leq HSI \leq 100 \quad (5)$$

Finally, the first main principal component of all six bioclimatic variables explained 94.37% of the overall variance and was chosen as a third index of HS (hereinafter referred to as PCA1). Using all three indices aims characterizing different components of the adaptation to HS. Two important assumptions of these HS indices should be noted. First, poorly adapted genotypes are inexistent because the distribution of accessions in the study areas is assumed to be in equilibrium with the niche requirements (Forester et al., 2016). Second, it is assumed that HS indices are stable over the years, since they are based on climatic data averaged over three decades. Ecological balance and stability of these HS indices are a prerequisite for GEA analysis (Cortés et al., 2013; Cortés and Blair, 2018). Since normality is also required for GWAS-type models, normality of each bioclimatic variable was verified using the skewness, kurtosis, and Shapiro–Wilk statistics ($P \geq 0.05$) using the *agricolae* package (De Mendiburu, 2014) in R v.3.4.4 (R Core Team). Dispersion diagrams, as well as Pearson (r) and Spearman (ρ) correlations, were made among all bioclimatic variables and HS indices in R v.3.4.4 (R Core Team).

Analysis of Kinship and Population Structure

Using the panel of 23,373 SNP markers, we estimated random and fixed effects in order to reduce the rate of false positives of each GEA model (*i.e.* MLM, CMLM, SUPER, FarmCPU, and BLINK). Random effects accounted for kinship relationships, while fixed effects accounted for population structure. Kinship was built in different ways according to the peculiarities of each algorithm. The MLM used a kinship matrix computed across all markers using the Loiselle, VanRaden, and EMMA methods available in the *GAPIT* package (Tang et al., 2016) of R v.3.4.4 (R Core Team). As an exploratory phase, we tested the power of these three different methods in capturing random effects in a GEA with MLM models. MLM models were selected for this purpose because they consider all 23,373 SNP markers. MLM models were designed using the combination of all three HS indices as response variable "I" (HIT, his, and PCA1), two population stratification methods as fixed effects "Q" (PC and TESS3), and three kinship methods as random effects "K" (Loiselle, VanRaden, and EMMA) for a total of 18 MLM models (3I × 2Q × 3K). Among all 18 MLM models, those that used the EMMA algorithm to reconstruct the kinship matrix were the most powerful. Thus, the following GEA models only considered the EMMA algorithm.

Based on this exploratory phase, only the EMMA algorithm was implemented for the reconstruction of the kinship relationships in the improved MLM algorithms (*i.e.* CMLM) and the last-generation GWAS models (*i.e.* SUPER, FarmCPU, and BLINK), each of which had different criteria for sub-setting the SNP dataset (PseudoQTNs) according to their specifications (Wang et al., 2014b; Liu et al., 2016; Huang et al., 2019).

Population stratification was explored using two strategies. First, a traditional molecular principal component analysis (hereinafter referred to as PC) was carried out in TASSEL v.5 (Bradbury et al., 2007). Second, spatial population structure was reconstructed using *TESS3* (Caye et al., 2016) as implemented in R v.3.4.4 (R Core Team). *TESS3* is a novel package that infers population structure from genotypic and geographical information. The optimum number of ancestral populations (K) was determined using a cross-entropy method implemented with the *snmf* function in the *LEA* package (Frichot and François, 2015) of R v.3.4.4 (R Core Team). The *snmf* algorithm was executed with 1,000 repetitions and a fluctuating K value from 2 to 10. The cross-entropy inference was further improved by exploring the percentage of masked genotypes at thresholds of 5% and 20%, as suggested by Frichot and François (2015) and Ariani et al. (2018), respectively. Results of population stratification were compared explicitly with previous studies carried out in wild common bean by Ariani et al. (2018). We selected a clustering coefficient (Q) cutoff of ≥0.7, following Ariani et al. (2018) and Bitocchi et al. (2012), for assigning genotypes to subpopulations.

Identification of Loci Associated With HS Indices

After the exploratory phase with 18 MLM models, we built 30 GEA models using improved MLM (CMLM) and last-generation GWAS (*i.e.* SUPER, FarmCPU, and BLINK) algorithms to explore single-marker associations. The improved MLM and last-generation GWAS algorithms increase the statistical power while better controlling the false-positive rate. FarmCPU and BLINK are particularly powerful at further controlling the false-negative rate (Huang et al., 2019). GEA models were obtained from the combination of all three HS indices as response variable "I" (HIT, HSI, and PCA1), two population stratification methods as fixed effects "Q" (PC and TESS3), and a unique kinship method as random effect "K" (EMMA with PseudoQTNs) for a total of 30 GEA models constructed by means of one improved MLM algorithm (CMLM) and three last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK). GEA models considered a total of six CMLM models (3I × 2Q × 1K), six FarmCPU models (3I × 2Q × 1K), six BLINK models (3I × 2Q × 1K), and 12 SUPER models. SUPER models were initially implemented as suggested by Wang et al. (2014b) in order to be computationally efficient, yet expecting the same statistical power as any MLM and CMLM models. To overcome this issue, these first-stage SUPER models were coupled with the MLM and CMLM algorithms for a total of 12 second-stage SUPER models (3I × 2Q × 1K × 1 first-stage GWAS algorithm × 2 second-stage GWAS algorithms).

Models were abbreviated as follows: $I_{M\text{-}Fc\text{-}Rc}$, where "I" refers to the HS index, "M" is the GWAS model family, and "Fc" and "Rc" are the algorithms used to reconstruct the fixed and random covariates, respectively. For example, the model $HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$ used HIT as the HS index, FarmCPU as the GWAS method, *TESS3*'s inference as a fixed covariate, and EMMA's kinship as a random covariate. This nomenclature was modified to account for the SUPER algorithm since it employed two different GWAS models in the first and last steps. The first step always used a GLM

model, but the last step used a MLM or CMLM model. Therefore, SUPER models were marked as $I_{SUPER(M)\text{-}Fc\text{-}Rc}$, where "M" is the model used in the last step (MLM or CMLM) (**Table S2**).

In order to choose the optimal GEA models, we drew Q–Q and Manhattan diagrams of the *P*-values with customized R scripts and used these diagrams to evaluate the rate of false positives. Highly significant associations were determined using a Bonferroni correction of *P*-values at an $\alpha = 0.05$, which led to a significance threshold of 2.14×10^{-6} or $-\log_{10}2.14 \times 10^{-6} = 5.67$ for CMLM models (2,373 effective SNP markers), 2.13×10^{-6} or $-\log_{10}2.13 \times 10^{-6} = 5.67$ for SUPER models (23,421 effective SNP markers), and 5.89×10^{-6} or $-\log_{10}5.89 \times 10^{-6} = 5.23$ for FarmCPU and BLINK models (8,494 effective SNP markers). Therefore, we used the Bonferroni threshold in order to evaluate the rate of false positives by visual interpretation of the Q–Q plots. In addition, a relax threshold of $-\log_{10}$ *P*-value = 4, as previously suggested (Pasam et al., 2012; Soltani et al., 2017; Soltani et al., 2018; Oladzad et al., 2019), was used only in the exploratory phase with 18 MLM models in order to identify weaker associations, since it is documented that the Bonferroni threshold is very restrictive or conservative in MLMs (Joo et al., 2016).

Identification of Candidate Genes

Putative candidate genes were identified by inspecting conservative flanking sections of 1 kb around each associated locus from all GEA models. Flanking sections were captured using the common bean reference genome v2.1 (Schmutz et al., 2014) and the *PhytoMine* and *BioMart* tools from the Phytozome v.12.3 platform.[3] Identified genes were further annotated using the GO,[4] PFAM,[5] PANTHER,[6] KEGG,[7] and UniProt[8] databases by means of Phytozome (see note C). Authors such as Oladzad et al. (2019) and Soltani et al. (2017; 2018) have suggested a genomic window to look for flanking genes of ~100 kb in common bean. On the other hand, LD in wild common bean, measured as marker correlation R^2, was reported to decay to 0.8 per every 81 kb (Rossi et al., 2009). Thus, we further explored a genomic window of 81 kb (40.5 kb upstream to 40.5 kb downstream of the significantly associated SNP markers) using the common bean reference genome v2.1 and the annotation tools as described above.

RESULTS

Among the entire set of 78 wild common bean accessions, we identified five accessions (G2648, G23511A, G13094, G12869, and G11071) putatively tolerant to HS based on three different bioclimatic indices (HIT, HSI, and PCA1). Incorporating these indices as response variables in GEA models led to 18 traditional MLM models that used three contrasting kinship reconstruction methods and 30 improved traditional mixed (*i.e.* ECLMLM) and last-generation GWAS models (*i.e.* SUPER, FarmCPU, and BLINK) that only used the EMMA algorithm for kinship reconstruction. None of the improved traditional mixed models yielded significant results. On the other hand, 15 last-generation GWAS models increased the statistical power to detect 120 significant associations in a GEA framework. A joint inference across these models and the three indices allowed having a more comprehensive understanding of the adaptive landscape and genetic architecture of heat tolerance. We recovered 22 genes, flanking 24 SNP markers, previously reported as candidates for heat tolerance (Wang et al., 2004; Ikeda et al., 2011; Lopes-Caitar et al., 2013; Oladzad et al., 2019; Soltani et al., 2019) and involved in the activation of heat shock proteins (HSPs), protein domains related to thermostability in plants and signaling pathways of abiotic stress via abscisic acid and auxin. These allelic variants require further validation and are ideal to be incorporated into common bean breeding programs for resistance to high temperatures.

Each Bioclimatic Index Captured a Different Component of HS

The three HS indices captured different facets of HS. All six bioclimatic variables (annual average temperature, maximum temperature of the warmest month, average temperature of the wettest quarter, average temperature of the driest quarter, average temperature of the warmest quarter, and average of the absolute maximum temperature of the reproductive phase) and three HS indices (HIT, HSI, and PCA1) exhibited a normal behavior (Shapiro–Wilk $P \geq 0.05$, **Figure S1**). HIT and PCA1 presented a positive bias with a skewness statistics of 0.160 and 0.271, respectively. On the other hand, HSI had a negative skewness with a skewness value of −0.166. All three HS indices allowed us to approximate the same HS event by different strategies. If different indices had distinct skewness values, contrasting extreme values described different facets of the HS event (**Figure S1**). Correlation coefficients estimated by Pearson (r) and Spearman (ρ) methods respectively ranged from 0.82 to 1 and from 0.78 to 1 among all bioclimatic variables and the HIT and HSI indices. The index built with the PCA1 had a negative correlation with all six bioclimatic variables and the other two HS indices (**Figure S2**) with Pearson (r) and Spearman correlation coefficients (ρ) ranging from −0.92 to −0.99 and −0.94 to −0.99, respectively. Therefore, despite differences in the extreme values, there is correspondence among all six bioclimatic variables and the three indices. Normality, together with the assumptions of stability over time and genotype–ecological niche equilibrium, makes these three HS indices suitable as response variables in GWAS models within a GEA framework aiming to capture various components of HS.

All 23,373 SNP Markers Recovered Six Subpopulations

Population structure as revealed by a PC (molecular PCA) analysis with 23,373 SNP markers suggested a total of six subpopulations (**Figure 1**). Also, cross-entropy validation implemented in *TESS3* with the same markers suggested an optimum of six subpopulations from Mesoamerica to northern Argentina (**Figure 1B**). Both methods, *TESS3* and PC, suggested six subpopulations: MW1 (Mesoamerican Wild 1), MW2 (Mesoamerican Wild 2), MW3

[3] https://phytozome.jgi.doe.gov/pz/portal.html
[4] http://geneontology.org/
[5] https://pfam.xfam.org/
[6] http://www.pantherdb.org/
[7] https://www.genome.jp/kegg/
[8] https://www.uniprot.org/

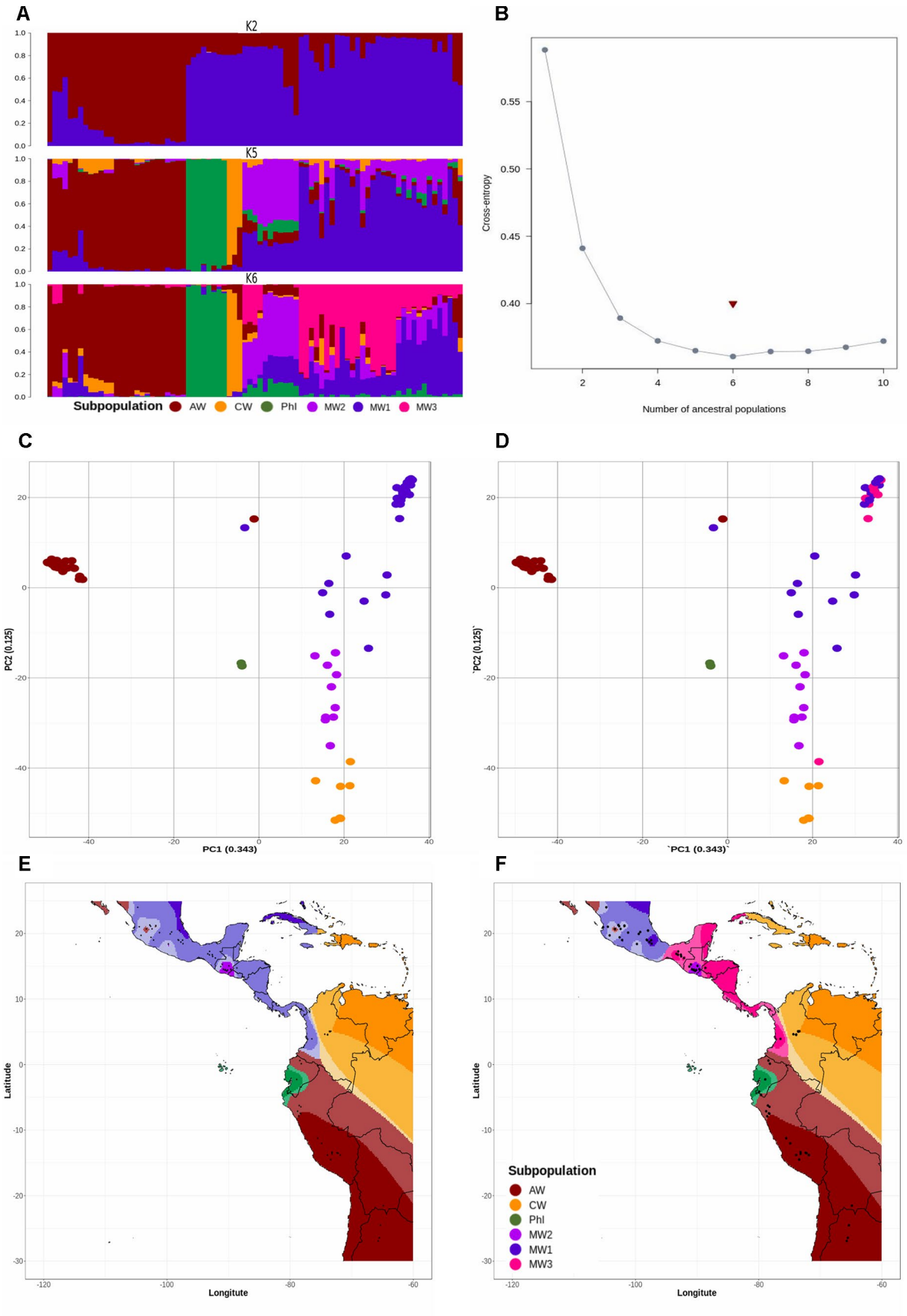

A
K2
K5
K6
Subpopulation AW CW Phl MW2 MW1 MW3
B
Cross-entropy
Number of ancestral populations
C
PC2 (0.125)
PC1 (0.343)
D
`PC2 (0.125)`
`PC1 (0.343)`
E
Latitude
Longitude
F
Latitude
Longitude
Subpopulation
AW
CW
Phl
MW2
MW1
MW3

FIGURE 1 | (A) Spatial population clustering and ancestry coefficients estimated with TESS3 using the number of gene pools (K = 2), the number of subpopulations suggested by other studies (K = 5), and the best number of subpopulations suggested by cross-entropy validation test (K = 6). The genotypes are sorted by latitude from northern Argentina to Mesoamerica. The subpopulations are MW1 (Mesoamerican Wild 1), MW2 (Mesoamerican Wild 2), MW3 (Mesoamerican Wild 3), PhI (Northern Peru-Ecuador Wild), AW (Andean Wild), and CW (Colombian Wild), colored in blue, purple, pink, green, red, and yellow, respectively. **(B)** Cross-entropy plot when the number of cluster (K) ranges between 1 and 10. The snmf algorithm was executed with 1,000 repetitions and a fluctuating K value from 2 to 10. The cross-entropy inference was further improved by exploring the percentage of masked genotypes at thresholds of 5% and 20%. **(C**, **D)** Population structure revealed by a molecular principal component analysis based on 23,373 SNP markers, using number of subpopulations K = 5 **(C)** and K = 6 **(D)**. Subpopulations are colored as in **(A)**. The percentage of explained variation by each axis is shown within parenthesis in the label of the corresponding axis. **(E**, **F)** Spatial interpolation of population ancestry coefficients across the geographic distribution of the genotypes analyzed. Subpopulations are colored as in **(A)**.

(Mesoamerican Wild 3), PhI (Northern Peru–Ecuador Wild), AW (Andean Wild), and CW (Colombian Wild) (**Figures 1D–F**). When we looked at the five subpopulations partition suggested by Ariani et al. (2018) based on following previous works 19,126 SNP markers flanking the *Cvi*AII restriction site, we did not recover Ariani's MW3 (**Figures 1C–E**), but instead the new subpopulation CW reappeared in both analyses (*TESS3* and PC).

EMMA Algorithm Was More Powerful at Reconstructing Kinship Relationships

As an exploratory phase, we built 18 traditional MLM models incorporating as random effects kinship matrices estimated with the Loiselle, VanRaden, and EMMA algorithms and as fixed effects estimates from TESS3 and PC (molecular PCA) algorithms across all 23,373 SNP markers. The three kinship algorithms were congruent among them and with the inferred population structure, revealing the typical Mesoamerican–Andean gene pool split (**Figure S3**). None of these 18 traditional MLMs recovered associated markers at a Bonferroni threshold of $-\log_{10}$ P-value = 5.67 (**Figures 2–4A**, **B**, **S4**, **S5A–L**, and **S6A–F**). Three loci systematically crossed the lax threshold of $-\log_{10}$ P-value = 4. They were on chromosomes Pv01 (S1_42870591) and Pv11 (S1_466464831 and S1_471851336) in all 18 traditional MLM models (**Figures 2A**, **B**, **3A**, **B**, **4A**, **B**, **S4A–L**, **S5A–L**, and **S6A–F**). Three of the models built with the EMMA algorithm ($HIT_{MLM\text{-}PC\text{-}EMMA}$, $HSI_{MLM\text{-}PC\text{-}EMMA}$, and $PCA1_{MLM\text{-}PC\text{-}EMMA}$) further identified three other alleles that crossed the lax threshold with greater significance (**Figures 2–4A**, **B**). Thus, the EMMA-based kinship matrix was defined as the random effect for the 30 improved traditional mixed and last-generation GWAS models.

A Total of 120 Loci in 15 Models Were Associated With the Three HS Bioclimatic Indices

We generated a total of 30 GEA models by implementing the algorithms CMLM (six models), SUPER (12 models), FarmCPU (six models), and BLINK (six models) using three HS indices as response variables, two methods of population stratification (PC and TESS3) as a fixed effect, and kinship reconstruction using the EMMA algorithm as a random effect. None of the six CMLM (**Figures 2–4C, D** and **S6G–L**) models yielded associated markers at a Bonferroni threshold of $-\log_{10}$ P-value = 5.67. However, at a lax threshold of $-\log_{10}$ P-value = 4, these CMLM models captured the same three associated loci identified by the 18 traditional MLMs. Three CMLM models that used the PCA1 as a covariable ($HIT_{CMLM\text{-}PC\text{-}EMMA}$, $HSI_{CMLM\text{-}PC\text{-}EMMA}$, and $PCA1_{CMLM\text{-}PC\text{-}EMMA}$) captured, at a lax threshold, one additional associated locus each (**Figures 2–4C**, **D**).

We implemented a GLM model in the first step of the SUPER algorithm as suggested by Wang et al. (2014b) due to its computational efficiency, same as MLM and CMLM models. MLM and CMLM models were implemented for the last step of the SUPER algorithm with each of the three HS indices. From all these 12 SUPER models, the only ones that reported associated markers at a Bonferroni threshold of $-\log_{10}$ P-value = 5.67 were $HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$ (**Figures 2E**, **F**), $HIT_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$ (**Figures 3E**, **F**), and $PCA1_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$ (**Figures 4E**, **F**), from now on named as $HIT_{SUPER\text{-}PC\text{-}EMMA}$, $HSI_{SUPER\text{-}PC\text{-}EMMA}$, and $PCA1_{SUPER\text{-}PC\text{-}EMMA}$, respectively, for better reading. The remaining nine SUPER models [$HIT_{SUPER(CMLM)\text{-}TESS3\text{-}EMMA}$, $HSI_{SUPER(CMLM)\text{-}TESS3\text{-}EMMA}$, $PCA1_{SUPER(CMLM)\text{-}TESS3\text{-}EMMA}$, $HIT_{SUPER(MLM)\text{-}PC\text{-}EMMA}$, $HSI_{SUPER(MLM)\text{-}PC\text{-}EMMA}$, $PCA1_{SUPER(MLM)\text{-}PC\text{-}EMMA}$, $HIT_{SUPER(MLM)\text{-}TESS3\text{-}EMMA}$, $HSI_{SUPER(MLM)\text{-}TESS3\text{-}EMMA}$, and $PCA1_{SUPER(MLM)\text{-}TESS3\text{-}EMMA}$], abbreviated as "failed" SUPER models, only identified between 17 and 12 SNP markers that crossed the lax threshold of $-\log_{10}$ P-value = 4 (**Figures S7A–P**). On the other hand, all 12 FarmCPU (**Figures 2–4G**, **H** and **5A–F**) and BLINK (**Figures 2–4I**, **J** and **5G–L**) models reported associated markers at a Bonferroni threshold of $-\log_{10}$ P-value = 5.23. Regardless what population stratification method was used (PC or TESS3) as a fixed effect, the 15 last-generation models SUPER (three), FarmCPU (six), and BLINK (six) identified a total of 120 associated loci at a Bonferroni threshold (**Table 1**). A total of 61 from the 120 SNP markers were captured by a single GEA model, and the remaining 59 SNP markers were associated with more than one of these GEA models; thus, we obtained a total of 270 GWAS redundant outputs (**Table S3**). The 120 significantly associated SNP markers were distributed in 105 regions across the common bean genome (**Figure 6**). Chromosomes Pv03, Pv01, Pv11, and Pv07 harbored the highest number of markers with 18, 15, 14, and 14 SNPs in 16, 12, 11, and 12 regions, respectively. Chromosomes Pv06, Pv08, Pv04, Pv02, and Pv10 had 10, 10, 10, 9, and 9 associated markers grouped in 10, 9, 10, 9, and 6 regions, respectively. Pv09 and Pv05 were the chromosomes with the fewest associated markers with seven and four SNPs, grouped in six and four regions (**Table S4**). On the other hand, PCA1 was the HS index with the highest number of markers with 96 SNPs in 83 regions through the entire genome. The HS indices HSI and HIT had 57 and 37 associated markers grouped in 54 and 33 regions, respectively, across all chromosomes (**Table S4**). Also, the last-generation GWAS algorithm with the highest number of associated markers was BLINK with 91 SNPs in 80 regions through the entire genome.

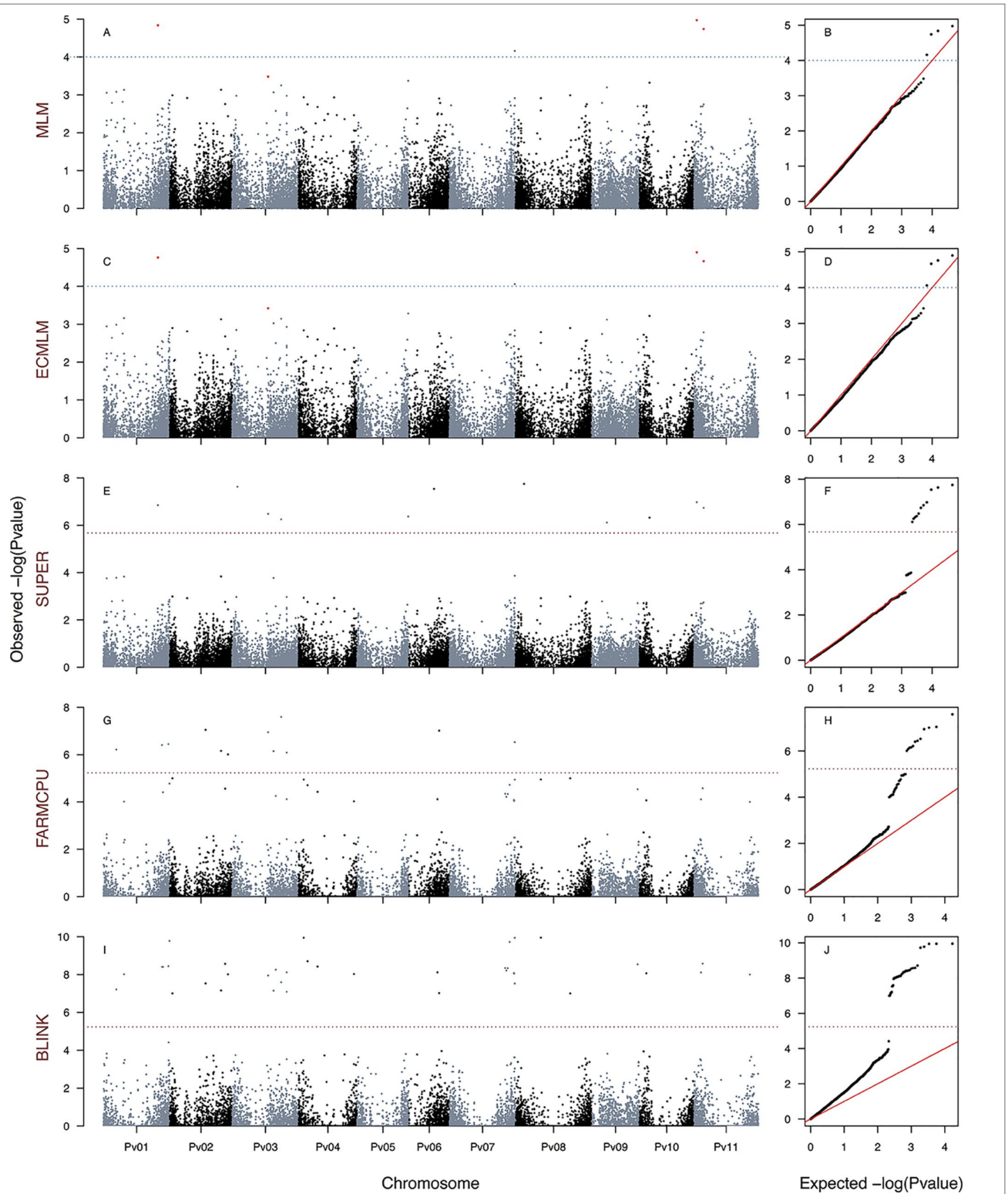


FIGURE 2 | Manhattan and Q–Q plots of the optimum genome–environment association (GEA) analysis for heat tolerance in 78 common bean accessions based on 23,373 SNP markers, using the HSI index (equation 4). The Manhattan and Q–Q plots are generated according to traditional MLM algorithms, compressed MLM algorithms, and last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK) using kinship matrix as a random effect by the EMMA algorithm and the first six principal components (**Figure 1D**) as a fixed effect. These models are $HSI_{MLM\text{-}PC\text{-}EMMA}$ **(A, B)**, $HSI_{CMLM\text{-}PC\text{-}EMMA}$ **(C, D)**, $HSI_{SUPER\text{-}PC\text{-}EMMA}$ **(E, F)**, $HSI_{FARMCPU\text{-}PC\text{-}EMMA}$ **(G, H)**, and $HSI_{BLINK\text{-}PC\text{-}EMMA}$ **(I, J)**. The red dashed horizontal line marks the *P*-value threshold after Bonferroni correction for multiple comparisons. The blue dashed horizontal line marks the lax *P*-value threshold. Black and blue colors highlight different common bean (Pv) chromosomes.

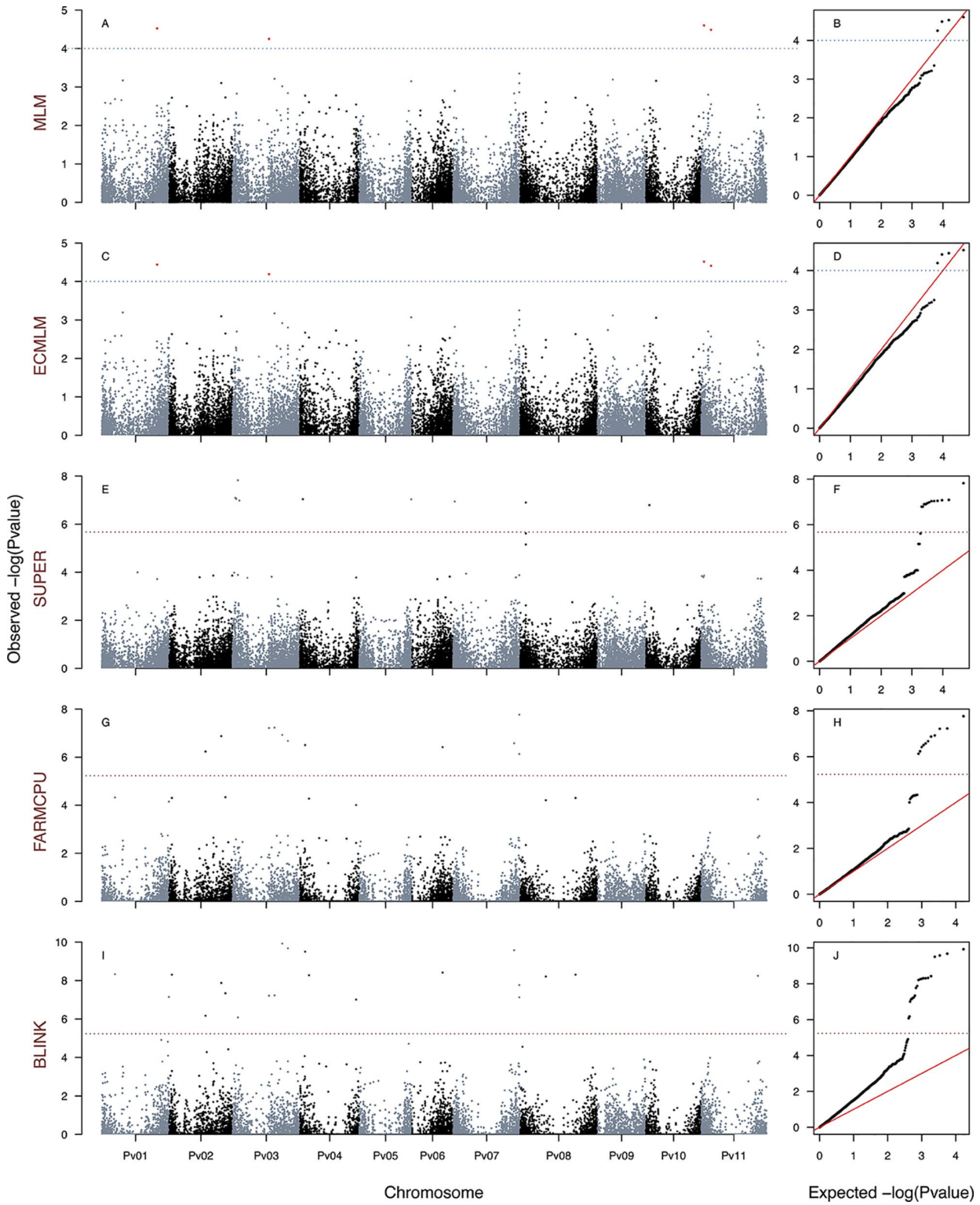


FIGURE 3 | Manhattan and Q–Q plots of the optimum genome–environment association (GEA) analysis for heat tolerance in 78 common bean accessions based on 23,373 SNP markers, using the HIT index (equation 5). The Manhattan and Q–Q plots are generated according to traditional MLM algorithms, compressed MLM algorithms, and last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK) using kinship matrix as a random effect by the EMMA algorithm and the first six principal components (**Figure 1D**) as a fixed effect. These models are $HIT_{MLM\text{-}PC\text{-}EMMA}$ **(A, B)**, $HIT_{CMLM\text{-}PC\text{-}EMMA}$ **(C, D)**, $HIT_{SUPER\text{-}PC\text{-}EMMA}$ **(E, F)**, $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$ **(G, H)**, and $HIT_{BLINK\text{-}PC\text{-}EMMA}$ **(I, J)**. The red dashed horizontal line marks the *P*-value threshold after Bonferroni correction for multiple comparisons. The blue dashed horizontal line marks the lax *P*-value threshold. Black and blue colors highlight different common bean (Pv) chromosomes.

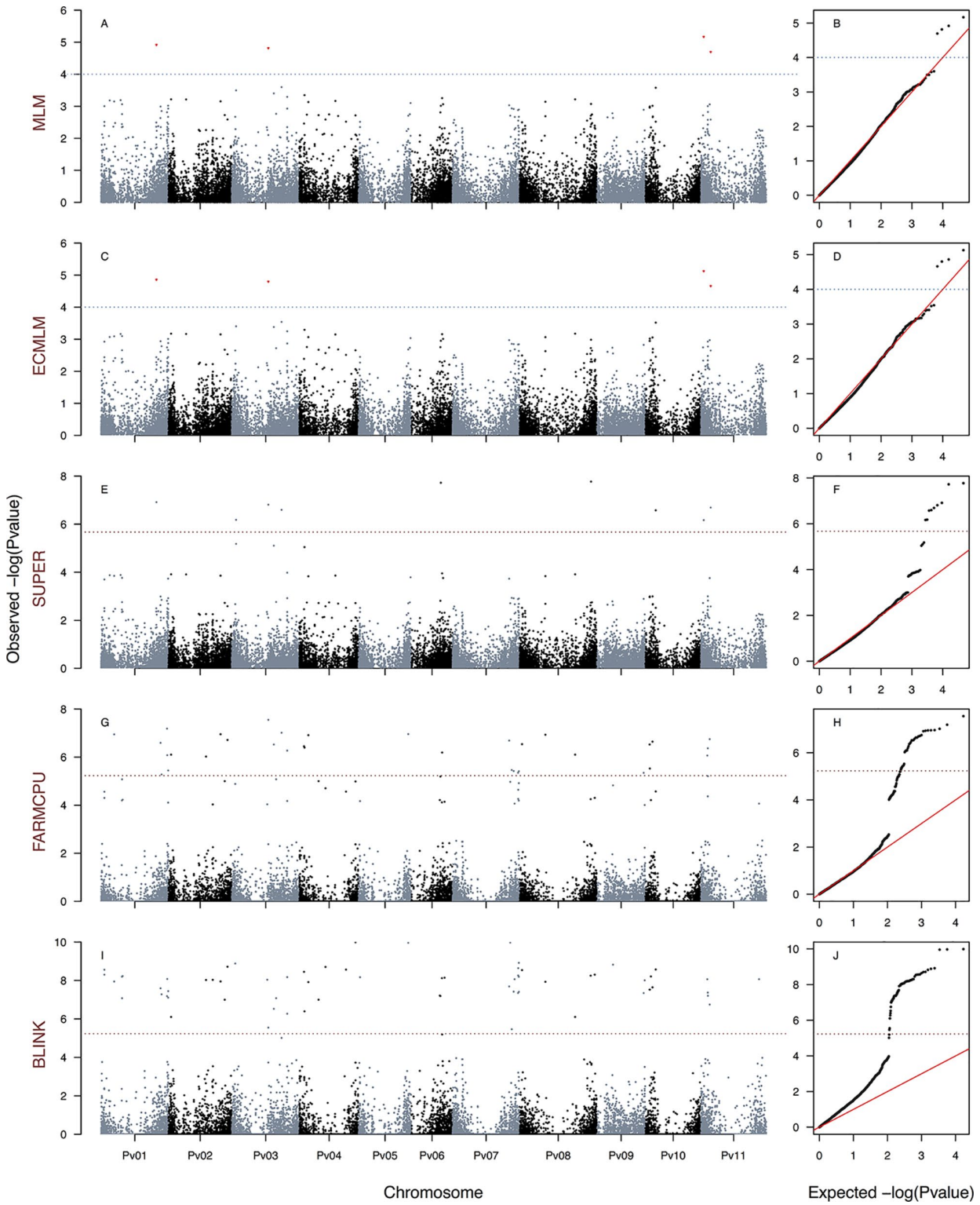


FIGURE 4 | Manhattan and Q–Q plots of the optimum genome–environment association (GEA) analysis for heat tolerance in 78 common bean accessions based on 23,373 SNP markers, using the PCA1 index. The Manhattan and Q–Q plots are generated according to traditional MLM algorithms, compressed MLM algorithms, and last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK) using kinship matrix as a random effect by the EMMA algorithm and the first six principal components (**Figure 1D**) as a fixed effect. These models are $PCA1_{MLM\text{-}PC\text{-}EMMA}$ **(A, B)**, $PCA1_{CMLM\text{-}PC\text{-}EMMA}$ **(C, D)**, $PCA1_{SUPER\text{-}PC\text{-}EMMA}$ **(E, F)**, $PCA1_{FARMCPU\text{-}PC\text{-}EMMA}$ **(G, H)**, and $PCA1_{BLINK\text{-}PC\text{-}EMMA}$ **(I, J)**. The red dashed horizontal line marks the *P*-value threshold after Bonferroni correction for multiple comparisons. The blue dashed horizontal line marks the lax *P*-value threshold. Black and blue colors highlight different common bean (Pv) chromosomes.

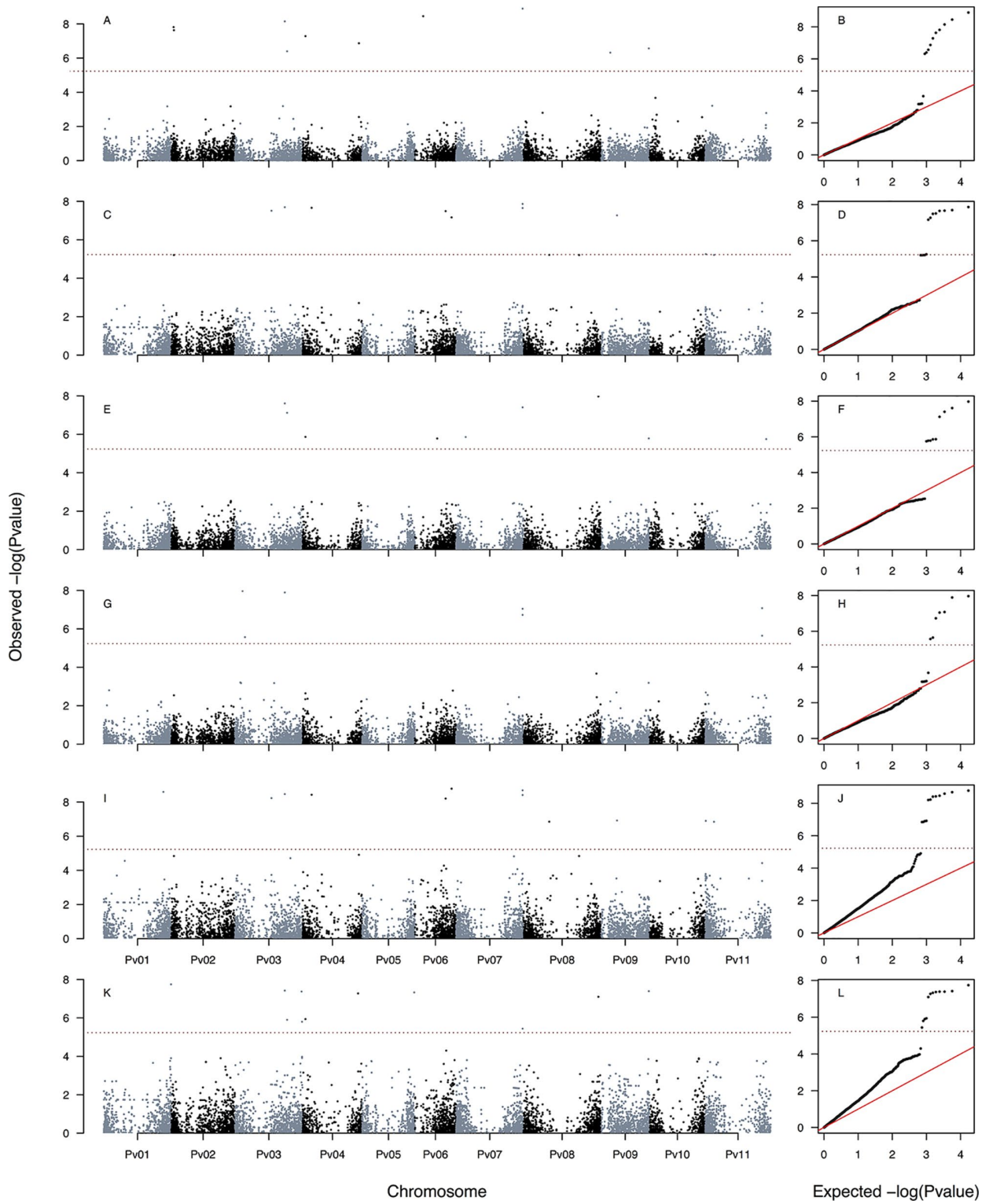


FIGURE 5 | Manhattan and Q–Q plots of the optimum genome–environment association (GEA) analysis for heat tolerance in 78 common bean accessions based on 23,373 SNP markers according to last-generation GWAS algorithms FarmCPU and BLINK. The covariates used in these six models provided are kinship matrix as a random effect using EMMA algorithm and the population structure as fixed effect using TESS3 (**Figure 1F**). These last-generation GWAS models are $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ **(A**, **B)**, $HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$ **(C**, **D)**, $PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$ (**E**, **F**), $HSI_{BLINK\text{-}TESS3\text{-}EMMA}$ **(G**, **H)**, $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$ **(I**, **J)**, and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$ **(K**, **L)**. The red dashed horizontal line marks the *P*-value threshold after Bonferroni correction for multiple comparisons. Black and blue colors highlight different common bean (Pv) chromosomes.

TABLE 1 | Summary statistics of the 15 gene–environment association (GEA) models for the 120 single-nucleotide polymorphism (SNP) markers associated with the three heat stress (HS) indices (HSI, HIT, and PCA1) in 78 common bean accessions based on the optimum association analysis (**Figures 2–5**).

Model	Pv	Number of associated SNPs	Average $-\log_{10}$ (*P*-value)	Average R^2 (%)	Number of exclusive associated SNPs	Number of associated regions	Number of regions containing genes	Number of genes	Number of genes related to HS
$HIT_{BLINK\text{-}PC\text{-}EMMA}$	1, 2, 3, 4, 6, 7, 8, 11	21	8.0 ± 1.1	64.62 ± 0.10	0	20	6	6	3
$HIT_{BLINK\text{-}TESS3\text{-}EMMA}$	1, 3, 4, 6, 7, 8, 9, 11	12	7.9 ± 0.8	60.86 ± 0.07	1	11	3	3	3
$HIT_{FARMCPU\text{-}PC\text{-}EMMA}$	2, 3, 4, 6, 7	11	6.8 ± 0.5	68.71 ± 0.10	0	10	4	4	2
$HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$	3, 4, 6, 7, 9	8	7.5 ± 0.2	57.09 ± 0.09	0	7	2	2	2
$HIT_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$	3, 4, 5, 7, 8, 10	12	7.0 ± 0.3	26.71 ± 0.02	9	9	5	5	5
$HSI_{BLINK\text{-}PC\text{-}EMMA}$	1, 2, 3, 4, 6, 7, 8, 9, 10, 11	39	8.2 ± 0.8	59.91 ± 0.10	1	36	9	9	4
$HSI_{BLINK\text{-}TESS3\text{-}EMMA}$	3, 7, 11	5	7.3 ± 0.6	63.36 ± 0.06	1	4	0	0	0
$HSI_{FARMCPU\text{-}PC\text{-}EMMA}$	1, 2, 3, 6, 7	12	6.5 ± 0.5	67.83 ± 0.10	0	12	6	6	4
$HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$	2, 3, 4, 6, 7, 9	10	7.4 ± 0.9	52.58 ± 0.18	5	10	5	5	5
$HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$	1, 3, 5, 6, 8, 9, 10, 11	11	6.8 ± 0.6	26.10 ± 0.05	3	11	4	4	3
$PCA1_{BLINK\text{-}PC\text{-}EMMA}$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	73	7.8 ± 0.8	56.44 ± 0.08	30	62	16	18	8
$PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$	1, 3, 4, 5, 8, 9	11	7.0 ± 0.7	48.59 ± 0.16	5	10	5	5	2
$PCA1_{FARMCPU\text{-}PC\text{-}EMMA}$	1, 2, 3, 4, 5, 6, 7, 8, 10, 11	27	6.6 ± 0.4	61.19 ± 0.09	0	26	14	14	9
$PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$	3, 4, 6, 7, 8, 9, 11	9	6.6 ± 0.9	47.82 ± 0.13	3	9	2	2	2
$PCA1_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$	1, 3, 6, 8, 10, 11	9	6.8 ± 0.6	42.86 ± 0.02	3	9	2	4	3

The FarmCPU and SUPER algorithms had 46 and 24 associated markers grouped in 44 and 21 regions, respectively, across all chromosomes (**Table S4**).

From 15 significant-GEA models, $PCA1_{BLINK\text{-}PC\text{-}EMMA}$, $HSI_{BLINK\text{-}PC\text{-}EMMA}$, $PCA1_{FARMCPU\text{-}PC\text{-}EMMA}$, and $HIT_{BLINK\text{-}PC\text{-}EMMA}$ were the models with the highest number of markers with 73, 39, 27, and 21 SNPs in 62, 36, 26, and 20 regions, respectively, through the entire genome. The models $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$, $HIT_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}PC\text{-}EMMA}$, $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$, $HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$, and $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ had 12, 12, 12, 11, 11, 11, and 10 SNPs in 11, 9, 12, 10, 11, 10, and 10 regions, respectively, across all chromosomes. $PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $PCA1_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$, and $HSI_{BLINK\text{-}TESS3\text{-}EMMA}$ were the models with the fewest associated markers with nine, nine, eight, and five SNPs, grouped in nine, nine, seven, and four regions, respectively (**Table 1**).

Also, the models $PCA1_{BLINK\text{-}PC\text{-}EMMA}$, $HIT_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$, and $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ had the highest number of exclusive markers that no other model captured, with 30, 9, 5, and 5 SNPs, respectively. The models $HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $PCA1_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $HSI_{BLINK\text{-}PC\text{-}EMMA}$, $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$, and $HSI_{BLINK\text{-}TESS3\text{-}EMMA}$ had the fewest exclusive markers with three, three, three, one, one, and one SNPs, respectively. The remaining models from the 15 significant-GEA models did not have exclusive SNP markers (**Table 1**). On the other hand, the 120 significantly associated SNP markers explained 54.28%, 52.73%, and 51.73% of the variation (effects) for PCA1, his, and HIT, respectively (**Table S4**). Furthermore, we averaged the R^2 of all associated SNPs by each of the significant 15 models throughout the genome of common bean, getting a range of average effects between 68.71% ($HIT_{FARMCPU\text{-}PC\text{-}EMMA}$) and 26.10% ($HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$) (**Table 1**). In summary, from the entire set of 30 GEA models implemented with improved traditional MLMs and last-generation GWAS algorithms, only 15 reported associations at a Bonferroni threshold, for a total of 120 associated markers.

Associated Markers Flanked 22 Genes Related With the HS Response, Such as Activation of HSPs and Abiotic Stress Signaling

We identified 120 associated loci across 15 of the 30 run GEA models at a Bonferroni-corrected significance threshold of –log10 P-value = 5.23 for 12 FarmCPU and BLINK models and at a Bonferroni-corrected threshold of –log10 P-value = 5.67 for three SUPER models. Among the 15 GEA models that captured significantly associated markers, only one (HITBLINK-TESS3-EMMA) did not identify any flanking gene. The other 14 models captured 36 flanking genes (**Table S3**). An ontology analysis revealed that 22 of these genes, flanking 24 associated markers, related with biological processes of the response to heat tolerance in plants (**Figure 6**, **Table 2**).

The chromosomes with the highest number of genes related to heat tolerance were Pv02, Pv06, Pv03, Pv01, Pv08, and Pv11, with five, four, three, three, three, and two genes, respectively

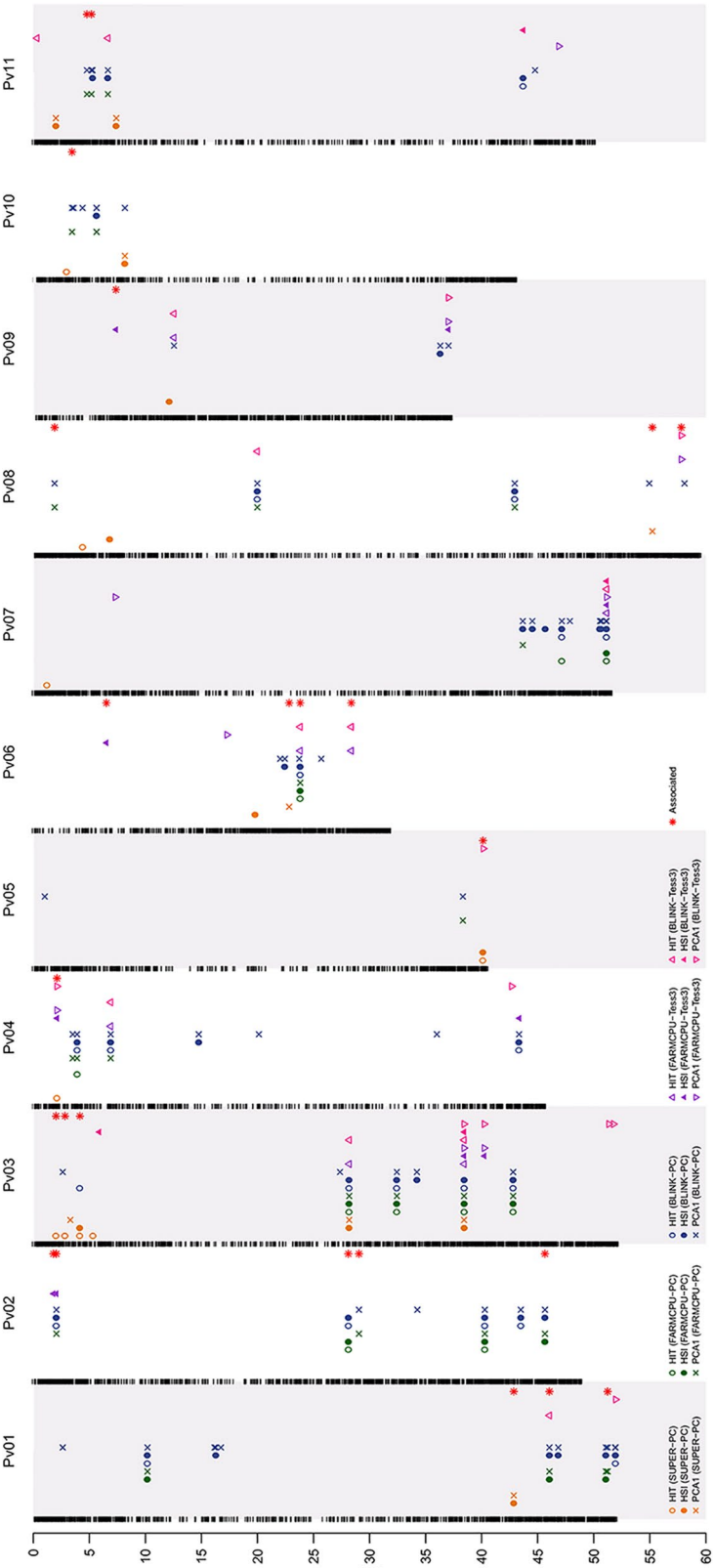


FIGURE 6 | Physical map of the 23,373 SNP markers for all 11 common bean (Pv) chromosomes. Physical position is shown in millions of base pairs (Mb). Each black hyphen corresponds to a SNP marker. The columns of yellow plotting symbols indicate markers associated by the SUPER algorithm with HIT (circle), HSI (filled circle), and PCA1 (cross) indices, using EMMA as a random effect and PC as a fixed effect (**Figures 2E**, **F**, **3E**, **F**, and **4E**, **F**). The columns of green plotting symbols indicate markers associated by the FarmCPU algorithm with HIT (circle), HSI (filled circle), and PCA1 (cross) indices, using EMMA as a random effect and PC as a fixed effect (**Figures 2G**, **H**, **3G**, **H**, and **4G**, **H**). The columns of blue plotting symbols indicate markers associated by the BLINK algorithm with HIT (circle), HSI (filled circle), and PCA1 (cross) indices, using EMMA as a random effect and PC as a fixed effect (**Figures 2I, J, 3I, J**, and **4I, J**). The columns of purple plotting symbols indicate markers associated by the FarmCPU algorithm with HIT (triangle point up), HSI (filled triangle point-up blue), and PCA1 (triangle point down) indices, using EMMA as a random effect and TESS3 as a fixed effect (**Figure 5A, F**). The columns of pink plotting symbols indicate markers associated by BLINK algorithm with HIT (triangle point up), HSI (filled triangle point-up blue), and PCA1 (triangle point down) indices, using EMMA as a random effect and TESS3 as a fixed effect (**Figure 5G, L**). The columns of red stars indicate markers associated with genes related with the HS response, such as activation of heat shock proteins (HSPs) and abiotic stress signaling. Regions are defined here as overlapping 1,000-bp sections that flanked associated markers (**Table 2**).

TABLE 2 | List of 24 single-nucleotide polymorphism (SNP) markers associated and flanked (genomic window of 1 kb) to 22 genes related with the heat stress (HS) response such as activation of heat shock proteins (22.73%), abiotic stress signaling (18.18%), germination and seedling development (18.18%), flowering time (9.09%), protein domain thermostability (9.09%), molecular chaperones (9.09%), and stability of the cell wall (4.55%) using PhytoMine [B] and reference genome of common bean v2.1.

Gene Name	GEA Model	Associated SNPs	Gen
	Activation of heat shock proteins—five genes (22.73%)		
Phvul.003G021100	$HIT_{SUPER(CMLM)-PC-EMMA}$	S1_103273611	***MED23***
Phvul.003G028300	$HIT_{SUPER(CMLM)-PC-EMMA}$	S1_104075622	***MED25***
Phvul.003G038600	$HIT_{SUPER(CMLM)-PC-EMMA}$, $HSI_{SUPER(CMLM)-PC-EMMA}$, $HIT_{BLINK-PC-EMMA}$	S1_105404421	***Hsp40***—Pv03
Phvul.002G136100	$HSI_{FARMCPU-PC-EMMA}$, $HIT_{FARMCPU-PC-EMMA}$, $HSI_{BLINK-PC-EMMA}$, $HIT_{BLINK-PC-EMMA}$	S1_80309359	***Hsp40***—Pv02
Phvul.006G182100	$HIT_{FARMCPU-TESS3-EMMA}$, $HIT_{BLINK-TESS3-EMMA}$	S1_268677251	***Hsp40***—Pv06
Phvul.002G019100	$HSI_{FARMCPU-TESS3-EMMA}$	S1_54254560	***HSFB1*** *(HSF4)*
Phvul.008G227900	$PCA1_{FARMCPU-TESS3-EMMA}$, $PCA1_{BLINK-TESS3-EMMA}$	S1_381855152	***HSP20***
	Abiotic stress signaling—four genes (18.18%)		
Phvul.008G204500	$PCA1_{SUPER(CMLM)-PC-EMMA}$	S1_379270378	***NFY***
Phvul.002G142500	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_81263655	***AUX_IAA***
Phvul.006G014100	$HSI_{FARMCPU-TESS3-EMMA}$	S1_246823134	Phospholipase C ***PLC***
Phvul.001G202000	$HSI_{FARMCPU-PC-EMMA}$, $PCA1_{FARMCPU-PC-EMMA}$, $HSI_{BLINK-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$, $HIT_{BLINK-TESS3-EMMA}$	S1_46052073	Auxin response factor
	Germination and seedling development–four genes (18.18%)		
Phvul.005G175800	$HSI_{SUPER(CMLM)-PC-EMMA}$, $HIT_{SUPER(CMLM)-PC-EMMA}$	S1_239620265	Glycoside hydrolases family 28
Phvul.002G016700	$HSI_{FARMCPU-TESS3-EMMA}$	S1_53999562	Family ***AP2/ERF***
Phvul.001G171600	$HSI_{SUPER(CMLM)-PC-EMMA}$, $PCA1_{SUPER(CMLM)-PC-EMMA}$	S1_42870591	***Ankyrin-B (Ankyrin-2)***
Phvul.011G054400	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_469234219	***Pkinase_Tyr***
	Flowering time—two genes (9.09%)		
Phvul.006G119900	$PCA1_{SUPER(CMLM)-PC-EMMA}$	S1_263134744	Poly(A) polymerase ***PAP***
Phvul.004G017600	$HIT_{SUPER(CMLM)-PC-EMMA}$, $PCA1_{FARMCPU-TESS3-EMMA}$, $HSI_{FARMCPU-TESS3-EMMA}$, $PCA1_{BLINK-TESS3-EMMA}$	S1_155643598	***MBD9***
	Protein domain thermostability—two genes (9.09%)		
Phvul.011G058100	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_469639214	Zinc finger A20 and AN1
Phvul.002G287600	$PCA1_{FARMCPU-PC-EMMA}$, $HSI_{FARMCPU-PC-EMMA}$, $HSI_{BLINK-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_97861798	S1
	Molecular chaperones—two genes (9. 09%)		
Phvul.009G032500	$HSI_{FARMCPU-TESS3-EMMA}$	S1_391075082	***14-3-3*** proteins family
Phvul.010G024400	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_424636676	***FKBP***
	DNA transcription—two genes (9. 09%)		
Phvul.008G022950	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_325947871	***BRIX***
Phvul.006G130200	$HSI_{FARMCPU-PC-EMMA}$, $HIT_{FARMCPU-PC-EMMA}$, $PCA1_{FARMCPU-PC-EMMA}$, $HSI_{BLINK-PC-EMMA}$, $HIT_{BLINK-PC-EMMA}$, $HIT_{FARMCPU-TESS3-EMMA}$, $HIT_{BLINK-TESS3-EMMA}$	S1_264118438	***YTH*** protein domain
	Stability of the cell wall—one gene (4.55%)		
Phvul.001G267000	$PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$	S1_51236796	***GAE6***

[B] *https://phytozome.jgi.doe.gov/pz/portal.html*

(**Table S4**). The chromosomes with only one gene related to heat tolerance were Pv04, Pv05, Pv09, and Pv10. Pv07 was the only chromosome that did not report gene related to heat tolerance. On the other hand, the PCA1 was the HS index with the highest number of genes related to heat tolerance with 14 genes. Furthermore, the HS indices HSI and HIT had 12 and 9 genes related to heat tolerance, respectively (**Table S4**). Also, the last-generation GWAS algorithm with the highest number of genes related to heat tolerance was FarmCPU with 18 genes. The BLINK and SUPER algorithms had 14 and 8 genes related to heat tolerance, respectively (**Table S4**).

From 15 significant-GEA models, $PCA1_{FARMCPU-PC-EMMA}$, $PCA1_{BLINK-PC-EMMA}$, $HIT_{SUPER(CMLM)-PC-EMMA}$, $HSI_{FARMCPU-TESS3-EMMA}$, $HSI_{BLINK-PC-EMMA}$, and $HSI_{FARMCPU-PC-EMMA}$ were the models with the highest number of genes related to heat tolerance with nine, eight, five, five, four, and four genes, respectively. On the other hand, $HSI_{SUPER(CMLM)-PC-EMMA}$, $PCA1_{SUPER(CMLM)-PC-EMMA}$, $HIT_{BLINK-TESS3-EMMA}$, $HIT_{BLINK-PC-EMMA}$, $PCA1_{FARMCPU-TESS3-EMMA}$, $PCA1_{BLINK-TESS3-EMMA}$, $HIT_{FARMCPU-TESS3-EMMA}$, and $HIT_{FARMCPU-PC-EMMA}$ were the models with the fewest number of genes related to heat tolerance with three, three, three, three, two, two, two, and two genes, respectively. $HSI_{BLINK-TESS3-EMMA}$ was the only GEA model that had no associated genes (**Table 1**).

A total of 22 genes flanked 24 loci because three different copies of the *HSP40* (Wang et al., 2004) gene were reported on three different chromosomes (Pv02, Pv03, and Pv06) using eight GEA models that incorporated HIT and HSI as response environmental variables. Four other genes from the set of 22 were also related to pathways of response to HS, such as activation of HSPs [*MED23* (Kim et al., 2004), *MED25* (Mathur et al., 2011), and *HSFB1* (Ikeda et al., 2011) in Pv02; and *HSP20* (Lopes-Caitar et al., 2013) in Pv08]. This set of five genes (*HSP20*, *HSP40*, *MED23*, *MED25*, and *HSFB1*) was recovered by 11 redundant GEA models [$HIT_{BLINK-PC-EMMA}$,

$HIT_{BLINK\text{-}TESS3\text{-}EMMA}$, $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$, $HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $HIT_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $HSI_{SUPER(CMLM)\text{-}PC\text{-}EMMA}$, $HSI_{BLINK\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$, and $PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$] (**Table 2**). These precursor genes of HSPs can play a crucial role in protecting plants against stress by reestablishing normal protein conformation and thus cellular homeostasis (Wang et al., 2004). Four significant SNP markers were found within the coding sequencing of the duplicated *HSP40* genes (Pv02 and Pv06), *MED23* and *MED25*. We also found two genes associated with protein domains related to thermostability in plants such as *S1* in Pv02 and *Zinc finger A20 and AN1* (Dixit and Dhankher, 2011) in Pv11 (**Table 2**).

We also recovered nine genes associated with biological processes likely correlated with plant tolerance to high temperatures, such as flowering time (*MBD9* in Pv04 and *PAP* in Pv06) (Peng et al., 2006; Trost et al., 2014), regulation of molecular chaperones (*FKBP* in PV10 and *14-3-3* proteins in Pv09) (Wang et al., 2004; Gollan et al., 2012), germination and seedling development [*Pkinase_Tyr* family in Pv11, *Ankyrin-B* in Pv01 (Hanks and Quinn, 1991; Bae et al., 2008), glycoside hydrolase *GH* family (González-Carranza et al., 2002) in Pv05, and transcription factors family *AP2/ERF* (Jofuku et al., 1994; Büttner and Singh, 1997) in Pv02], and cell wall stability (*GAE6* in Pv01) (Usadel et al., 2004). Additionally, four genes were involved in the signaling pathways of abiotic stress via abscisic acid [histone-like transcription factors *NFYB* (Warpeha et al., 2007) in Pv08 and phospholipase C *PLC* (Peters et al., 2010) in Pv06] and auxin (auxin response factor in Pv01 and *AUX_IAA* in Pv02) (Hagen and Guilfoyle, 2002; Ellis et al., 2005) (**Table 2**). On the other hand, since HS compromises molecular processes inherent to DNA transcription, it is not unexpected that we found two transcription factors [*BRIX* (Weis et al., 2015) in Pv08 and protein domains *YTH* (Wang et al., 2014a) in Pv06] involved in plant development and response to abiotic stress such as drought and heat. Overall, the biological processes related to HS over-represented among the associated genes were thermal shock protein activation (22.73%), abiotic stress signaling (18.18%), and germination and seedling development (18.18%) (**Table 2**).

Additionally, we explored a genomic window of 81 kb (40.5 kb upstream to 40.5 kb downstream of the associated SNP using the common bean v2.1, **Table S5**) based on LD criterion, finding 541 new genes for a total of 578 genes. Among the 578 genes, we found eight genes related to HSPs (three *HSP40*, two *HSP20*, one *HSFA5*, and two *HSP17.6*) in addition to the five genes found in the window of 1 kb (three *HSP40*, one *HSP20*, one *HSFB1*, one *MED23*, and one *MED25*) for a total of thirteen genes. The eight new genes related to HSP were distributed like this: three *HSP40* in chromosomes Pv01, Pv06, and Pv07; two *HSP20* in chromosomes Pv05 and Pv08; one *HSFA5* in chromosome in Pv01; and two *HSP17.6* in chromosome Pv08.

Last-Generation GWAS Models Complemented Each Other Despite Some Redundancy

Based on the previous gene recovery and classification, 11 GEA models were the best at explaining the activation of HSPs as the genetic basis of heat tolerance, by reporting seven loci across five chromosomes (**Tables 2** and **S3**) as potential candidates to be integrated into breeding programs. These seven loci were related to genes belonging to the HSPs' activation signaling pathway. From these 11 GEA models, the ones that best explained the HS indices were $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$ (68.71%), $HSI_{FARMCPU\text{-}PC\text{-}EMMA}$ (67.83%), and $PCA1_{FARMCPU\text{-}PC\text{-}EMMA}$ (61.19%). In other words, the last-generation GWAS model families that best explained the HS indices were FarmCPU and BLINK. Meanwhile, SUPER models reported the weakest effects (42.86%) (**Table 1**).

Among the 11 most-explanatory GEA models, 10 models, distributed in four main clusters, were redundant. $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ was the unique non-redundant model that captured a gene related to heat tolerance (*HSFB1*) (**Table 2**). The clustering criterion was that models within the same cluster captured the same gene. The first cluster had three models ($HIT_{SUPER\text{-}PC\text{-}EMMA}$, $HSI_{SUPER\text{-}PC\text{-}EMMA}$, and $HIT_{BLINK\text{-}PC\text{-}EMMA}$) that reported a paralogous copy of the *HSP40* gene in chromosome Pv03. The second cluster had four models ($HIT_{BLINK\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}PC\text{-}EMMA}$, $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$, and $HSI_{BLINK\text{-}PC\text{-}EMMA}$) that reported a paralogous of the same gene in chromosome Pv02. The third cluster had two models ($HIT_{FARMCPU\text{-}TESS3\text{-}EMMA}$ and $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$) that identified the other paralogues of *HSP40* in chromosome Pv06. The fourth cluster was made of two models ($PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$ and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$) that captured the same *HSP20* gene in chromosome Pv08. On the other hand, the genes that were captured by non-redundant models were *MED23* and *MED25* (both with $HIT_{SUPER\text{-}PC\text{-}EMMA}$) and *HSFB1* (with $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$). The $HIT_{SUPER\text{-}PC\text{-}EMMA}$ model was not redundant with other models when capturing these two genes, but this model was redundant with the first cluster when capturing the Pv02 *HSP40* paralogues.

The $HIT_{BLINK\text{-}PC\text{-}EMMA}$ model simultaneously reported the paralogous *HSP40* gene in chromosomes Pv02 (SNP marker S1_80309359, effect = 56.22%) and Pv03 (SNP marker S1_105404421, effect = 56.74%), from the first and second clusters, respectively. The LD between both SNP markers reported by $HIT_{BLINK\text{-}PC\text{-}EMMA}$ had an R^2 of 6.2% (P-value = 0.045). In other words, both SNP markers were recovered by the same model ($HIT_{BLINK\text{-}PC\text{-}EMMA}$) and accounted for different effects of paralogous copies of the *HSP40* gene in different chromosomes. So we selected the $HIT_{BLINK\text{-}PC\text{-}EMMA}$ model as the representative model of the first and second clusters. On the other hand, we selected the most explanatory models (highest effects) as representative models for the third and fourth clusters. Thus, we chose the $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$ model (effect = 60.86%) as the representative model of the third cluster (**Table 1**). This model identified the *HSP40* gene in chromosome Pv06. Finally, we selected the $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$ model (effect = 48.59%) as the representative model of the fourth cluster (**Table 1**). This model captured the *HSP20* gene in chromosome Pv08. Therefore, the 11 models that best explained the activation of HSPs can be condensed into five non-redundant models, which are $HIT_{BLINK\text{-}PC\text{-}EMMA}$, $HIT_{SUPER\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$, and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$. Each of these five non-redundant GEA models captured a unique gene of the HSPs' activation signaling pathway, including regulators

of mediators, activators, and expression genes ($HIT_{BLINK\text{-}PC\text{-}EMMA}$ captured *HSP40* in Pv03 and Pv02, $HIT_{SUPER\text{-}PC\text{-}EMMA}$ captured *MED23* and *MED25*, $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ captured *HSFB1*, $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$ captured *HSP40* in Pv06, and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$ captured *HSP20*) (**Table 2**).

DISCUSSION

The discriminatory power provided by kinship covariates used as random effects has been of great interest in the development of promising GWAS algorithms (CMLM, SUPER, FarmCPU, and BLINK). However, last-generation GWAS algorithms have given greater importance to the selection of SNP markers for the kinship reconstruction than to the reconstruction method itself. The three HS indices (HIT, HSI, and PCA1) and 15 last-generation GWAS models that generated significant results captured complementary components of the genetic architecture of heat tolerance. We found a total of 24 loci associated to 22 genes related to biological processes of the HS response in plants. Also, among the 24 loci, we captured seven loci as potential candidates to be integrated into breeding programs, since they were flanking five genes belonging to the signaling pathway that activates HSPs.

Bioclimatic Indices Capture Complementary Genetic Effects Conferring Heat Tolerance

Each HS index captures a different facet of the HS event. The HIT index uses accumulated information of maximum temperatures during the reproductive phase of common bean and therefore is more informative over time in capturing extreme values related to HS. Because of this dynamic nature of HIT, models that integrated HIT were more successful at capturing genes related to the activation of HSPs. In addition, the $HIT_{SUPER\text{-}PC\text{-}EMMA}$ model, which integrates the HIT index as a response variable, captured unique results that no other model recovered, by reporting key genes in the activation of HSPs such as *MED23* and *MED25* activators, which are key genes in the reconstruction of the genetic basis of heat tolerance.

On the other hand, the HIS index is built on thresholds of maximum temperature during the reproductive phase reported by some authors for plants in the tropics (Gonçalves et al., 1997; Caramori et al., 2001; Silva et al., 2007; Rainey and Griffiths, 2019). Thus, this index could be more informative phenologically in capturing extreme values related to HS events. This is based on the fact that models constructed with HIS tended to capture more unique genes than any other index. Furthermore, $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$ was the only model that captured the HS gene heat shock factor *HSFB1* (*HSF4*). Among the set of genes captured by 11 GEA models, *HSF4*, a regulatory gene in the expression of HSPs in *Arabidopsis thaliana* (Ikeda et al., 2011), is the gene that has greater regulatory importance both in the activation of HSPs and other molecular mechanisms of response to abiotic stress. Then, although the HIS index fails to capture the amount of genes that the HIT index did, perhaps because of its stationary nature, it manages to identify unique results that are essential to reconstruct the complexity of the genetic effects that confer heat tolerance.

Finally, the index based on PCA1 exhibits variability that the first two indices did not offer. PCA1 integrates other bioclimatic variables besides T_f, yet still related to abiotic stress events. The wide variability offered by PCA1 is evident in the large coverage of the GEA models that relied on this index. These models capture more candidate genes than the previous ones (14 from 22 genes). They also capture more biological processes related to HS (*e.g.* abiotic stress signaling, germination and development of seedlings and flowering time). However, they recover few genes related with the activation of HSPs proteins. The models $PCA1_{FARMCPU\text{-}TESS3\text{-}EMMA}$ and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$ capture unique genes such as *HSP20*, reported in soybean as activator of HSPs (Lopes-Caitar et al., 2013), and reported in common bean as one of the three most over-expressed genes under HS using RNA-sequencing (Soltani et al., 2019).

Each index captures unique genes associated with the activation of HSPs, but each of them also identifies different paralogous copies of the same gene. The models that used the HSI and HIT indices, recover genes *upstream* to HSPs genes in the pathway of activation of HSPs (*i.e. HSFB1*, *MED23*, and *MED25*). On the other hand, genes of the family of low molecular weight sHSPs (small HSPs), such as *HSP40*, are found in Pv02 and Pv03 chromosomes using the HIT and HSI indices, and in Pv06 chromosome using the HIT. Also, other low molecular weight HSPs such as *HSP20* are captured by the PCA1 index. Thus, models that integrate different indices manage to identify mediating, activating and expression genes (sHSPs), providing a more comprehensive understanding of the genetic architecture of heat tolerance. Although the three indices fail to capture all the conserved families of HSPs such as *Hsp70*, *Hsp60*, *Hsp90* and *Hsp100*, they detect associations flanking several genes of the family sHSPs, such as *HSP20* and *HSP40*, this is possibly because sHSP family is the most prevalent in plants (Vierling, 1991). In addition, gene diversification and subspecialization may reflect molecular adaptation to stress conditions that are unique to specific populations (Wang et al., 2004). On the other hand, high abundance of sHSPs in multiple cellular compartments suggests that they may have an important role in acquisition of stress tolerance in plants (Wang et al., 2004). In this sense, the expression of sHSPs genes, as those detected in this study by means of the three different indices, despite not being the proteins that have higher folding potential (as *Hsp70* and *Hsp90)*, can be key regulatory steps of the molecular response to HS (by modulating genes such as *HSFB1*, *MED23*, and *MED25*, that we also managed to detect).

An Assortment of Various Last-Generation GWAS Models Offer Better Alternatives for GEA Studies

Each last-generation GWAS algorithm implemented in this study differs in the internal strategy that uses to reconstruct the random Kinship covariable. While the kinship method is consistent across algorithms, the implementation of Pseudo QTNs differs. On the other hand, a prerequisite for GWAS models is the use of fixed covariables for population structure, being the principal components analysis (PC) the most traditional method. However, the generation of alternative strategies such as the one implemented

in TESS3, which is more powerful to reconstruct the stratification of the population as evidenced by the works of Caye et al. (2016), Ariani et al. (2018) and Varshney et al. (2017), led us to consider TESS3 as a promising method to be integrated into the GEA models. The results obtained by models that use TESS3 as a fixed covariate, demonstrate its importance to capture candidate genes not recovered by any other GEA model, such as an activator of HSP proteins (*HSFB1*) and two HSPs of low molecular weight (*HSP20*, and *HSP40* in Pv06). On the other hand, the implementation of the PC method as a fixed covariate in GEA models is also useful, because the models that integrate this method capture unique genes such as HSPs of low molecular weight (*HSP40* genes in Pv02 and Pv03) and activators of HSP proteins (*MED23* and *MED25*).

In summary, 30 GEA models were built with TESS3 and PC as fixed covariables, from an improved traditional MLM algorithm (CMLM) and three last-generation GWAS models (SUPER, FarmCPU, and BLINK). Of the 30 GEA models, 14 used last GWAS algorithms and reported genes linked to biological processes related to HS. A total of 11 of these 14 models captured genes related to molecular mechanism of activation of HSPs proteins. This molecular process was given greater focus due to its importance for heat tolerance and its relationship with other stresses. The 11 GEA models that identified HSP activation genes can be condensed into five non-redundant GEA models, conserving the same number of associated genes.

We did not find the 'holy grail' for GWAS models, which is a unique model that would summarize all 14 GEA models. The majority of the 14 GEA models that used FarmCPU and BLINK algorithms are redundant in the results related to activation of HSPs, regardless whether these considered TESS3 or PC as fixed covariables. The coincidence between the results obtained by FarmCPU and BLINK had already been reported by the authors of the BLINK algorithm for flowering time in corn (Huang et al., 2019), and was attributed to the way both strategies are conceived. They operate by separating the mixed model into a fixed sub-model and a random sub-model, differing only in the parameter-estimation method (Huang et al., 2019). This is why both methodologies converge to the same results for heat tolerance in common beans and flowering time in corn. However, in our study an exception to the redundancy between $HIT_{BLINK\text{-}PC\text{-}EMMA}$ and $HIT_{FARMCPU\text{-}PC\text{-}EMMA}$ algorithms was that the exact identity of the associated markers within the candidate genes differed. Besides, despite that the authors of BLINK reported that this method captures more associated genes to flowering time than FarmCPU, we found the opposite pattern when it comes to heat tolerance in common bean. This suggests that the algorithms could be sensible to the use of different response variables (*e.g.* environmental vs. phenotypic).

The Q–Q plot can provide information on two main aspects of GWAS data: whether the statistical testing is well controlled for challenges such as population stratification and whether there is any association. In the last aspect, we could see some associations at the end of the Q–Q plot crossing the Bonferroni threshold. The population structure control is still a challenge in GWAS and our Q–Q plots show signs of inflation. This inflation could partially be produced by causal SNPs (or SNPs in LD with causal variants), that at the same time are strongly differentiated among gene pools. This scenario is possible because both gene pools come from contrasting environments in terms of exposure to HS events. Mesoamerican genotypes generally experience more heat events than Andean genotypes (**Figure S8**).

In conclusion, the five non-redundant GEA models that best explain the activation of HSPs as the genetic basis of heat tolerance are $HIT_{BLINK\text{-}PC\text{-}EMMA}$, $HIT_{SUPER\text{-}PC\text{-}EMMA}$, $HSI_{FARMCPU\text{-}TESS3\text{-}EMMA}$, $HIT_{BLINK\text{-}TESS3\text{-}EMMA}$ and $PCA1_{BLINK\text{-}TESS3\text{-}EMMA}$. Each of these models captures a key gene in the pathway of activation of sHSPs, including genes involved in the regulation, activation and expression of the signal (Vierling, 1991). Therefore, using an assortment of last-generation GWAS methods, various environmental indices and different methods to account for fixed covariates, is much more informative than trying to select a single optimum GWAS model. Our work presents for the first time a powerful strategy to explore GEAs throughout a wide range of different last-generation GWAS models. This opens the door for new ways to couple environmental information in the study of complex characters, such as heat tolerance.

Modern GEA Is Capable of Revealing the Genetic Basis of a Complex Adaptive Trait Despite Limited Sampling

HS affects several physiological, cellular and molecular processes in plant cells, affecting fluidity of the cell membrane (Savchenko et al., 2002), protein (Ahmad et al., 2009) and cytoskeletal stability (Bita and Gerats, 2013), chromatin structure (Khraiwesh et al., 2012), the production of reactive oxygen species (ROS) (Camejo et al., 2006) as well as metabolic coupling (Bita and Gerats, 2013). Consequently, HS generates responses in plant cells at molecular and cellular levels, such as activation of HSPs (Wang et al., 2004), calcium signaling (Larkindale and Huang, 2004), phosphorylation, changes in the transcription (Bita and Gerats, 2013) and hormonal responses via Abscisic Acid (Larkindale and Knight, 2002), Ethylene or Auxin (Evrard et al., 2013; Larkindale and Huang, 2004). Yet, HS also affects processes such as flowering time, germination and abscission of floral organs (Bita and Gerats, 2013). The genes reported in this work may be causal or in LD with causal genes, involved in the majority of these processes. Although, we captured at least one gene in each of these biological processes, the highest number of associated genes were involved in the activation of HSPs. This could be attributed to the ability of the sHSPs family (*e.g. HSP20* and *HSP40*) and HSF genes (*HSFB1*) to activate HSPs as well as other physiological, cellular, and molecular mechanisms of heat tolerance in plants, such as hormonal signaling routes (Wang et al., 2004), photosystem II protection (Kotak et al., 2007, Soltani et al., 2019), DNA translation control (Malik et al., 1999) and elimination of reactive oxygen species (ROS) (Bita and Gerats, 2013). In addition, if we focused in genes related to HSPs, the resolution to detect these proteins decreases with a wider window of 81 kb. Among the 578 genes, we found eight genes related to HSPs (three HSP40, two HSP20, one HSFA5, and two HSP17.6) in addition to the five genes found in the windows of 1 kb (three HSP40, one HSP20, one HSFB1, one MED23, and one MED25) for a total of thirteen genes. However, these thirteen genes are the 2,25 % of the 578 genes found in a genomic window of 81 kb, while in a narrower genomic window of 1 kb, the five genes related to HSP are the 13.5% of the 37 genes.

Although we were unable to reconstruct the entire pathways of HSP protein activation, hormonal responses, time to flowering and seedling development, we found key genes in these biological processes, by only using environmental information from the accession's sampling sites. This strategy is valuable in optimizing time and costs for association studies using wild material.

We have demonstrated that combining diverse and contrasting samples with cautiously synthesized environmental variables, through a range of diverse last-generation models, offers an unprecedented power for GEA studies in the absence of phenotyping and with moderate sample sizes. By doing this, we identified a broad genetic basis for heat tolerance in common bean, and captured adaptive loci related to the activation of HSPs (*HSFB1*, *MED23*, and *MED25*) as well as HSPs of low molecular weight (*HSP20* and *HSP40*). Small HSP family genes were actually identified as relevant in the recent work by Soltani et al. (2019), where authors detected *HSP21* as one of the three most over-expressed genes in common bean under HS using RNA-sequencing.

On the other hand, the use of traditional GWAS models and raw environmental information should be avoided since they lack statistical power to detect associated markers. Several authors had already pointed this limitation (Cortés and Blair, 2018; Frank et al., 2016; Lasky et al., 2015). Therefore, we suggest coupling synthesized environmental variables with diverse last-generation models, in order to reveal more accurately the adaptive genetic variation to different types of stress in collections of wild germplasm.

PERSPECTIVES

This study demonstrates that the implementation of last-generation GWAS models under a GEA framework with carefully chosen environmental indices improves the reconstruction of the genetic basis of adaptation to HS. New studies across a variety of species and populations subjected to different stresses will benefit by using last-generation GWAS models within a well thought GEA design in order to capture better sources of genetic adaptation. We are looking forward to seeing more studies that follow these lines within the oncoming years.

On the other hand, the genes identified in this study as candidates for heat tolerance have the potential to be used in plant breeding programs after validation by means of strategies such as gene expression studies and Whole Genome re-Sequencing (WGS) (Barbulescu et al., 2018). The latter will make available all the genetic variability present in each accession. Additionally, it would be ideal that the indices explored in this work were contrasted with measurements of heat tolerance in greenhouse and at field conditions under controlled treatments (Zuiderveen et al., 2016). It would also be appropriate to consider for these experiments the same group of accessions used in the present work as well as accessions of related species that are well-known for their heat tolerance (*i.e. Phaseolus acutifolius*). Ultimately, validated candidate genes could be integrated into molecular editing strategies (Lang-Mladek et al., 2010; Pecinka et al., 2010; LeBlanc et al., 2018).

As part of a larger project, promissory accessions identified in this work will be evaluated together with advanced lines and related species under HS conditions at Coastal Colombia. These materials are currently undergoing seed multiplication at the greenhouses so that field establishment can take place in 2020.

Finally, by exploring the genetic basis of heat tolerance using indices constructed from phenotypic information, it will be possible to couple GBS and WGS data with last-generation GWAS models and genomic selection approaches (Crossa et al., 2011). In parallel, there have been recent GWAS developments relying on Artificial Intelligence (AI) (*i.e. deep learning*) and Machine Learning (ML) strategies that deserve further exploration under a GEA framework.

DATA AVAILABILITY STATEMENT

The filtered dataset and data analysis pipeline are archived at the Dryad Digital Repository (https://doi.org/10.5061/dryad.9k862c8).

AUTHOR CONTRIBUTIONS

AC conceived this study. AC carried out DNA extractions to produce GBS data. LL recovered historical environmental data. LL analyzed and interpreted environmental and GBS data with guidance from AC. LL wrote the manuscript with contributions from AC.

FUNDING

The genotyping and early analyses done as part of this work were funded by the Lundell and Tullberg grants, with support from the grants 4.J1-2016-00418 and BS2017-0036 from Vetenskapsrådet (VR) and Kungliga Vetenskapsakademien (KVA) to AC as PI. The Geneco Mobility Fund to AC and the Fulbright Specialist Award to M.W. Blair are acknowledged for encouraging synergistic discussions around common bean genetics in Nashville (TN, USA) and Rionegro (Antioquia, Colombia) during 2015 and 2019, respectively. The Network for Vegetable Research and the Training & Development Department from the Colombian Corporation for Agricultural Research are thanked for sponsoring LL's internship. The Editorial Fund from the same institute is recognized for subsidizing the publication fee of this work.

ACKNOWLEDGMENTS

We thank the Genetic Resources Unit at the International Center for Tropical Agriculture and D.G. Debouck for providing the seeds of the accessions that were considered in this study. We are also grateful with the Genomic Diversity Facility of the Institute of Biotechnology at Cornell University for support in SNP genotyping and calling. Some of the ideas presented in this manuscript were shared by LL and discussed with N. Palacios, M. Soto-Suárez and E. Torres-Rojas as part of the C2B2 conference, held in Bogotá (Colombia) on November 2018. Insights from M.W. Blair, E. Burbano, I. Cerón, C.H. Galeano, C.I. Medina, D. Peláez, P. Reyes, J.M. Rojas and A. Tofiño are also very much appreciated.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00954/full#supplementary-material

TABLE S1 | Identity of the 78 common bean accessions used in this study. The G identification number (from the Genetic Resources Unit at the International Center for Tropical Agriculture), country of origin and georeferences. The raw bioclimatic variables used are BIO1 = annual mean temperature, BIO5 = maximum temperature of warmer months, BIO8 = mean temperature of the wettest quarter, BIO9 = mean temperature of the driest quarter, BIO10 = mean temperature of the warmest 4-month period, and T_j = average of absolute maximum temperature during the reproductive phase. The bioclimatic-based heat indices are the HSI, HIT, and PCA1. The membership value of 78 common bean accessions using TESS3 algorithm (K = 6).

TABLE S2 | List of notations of GEA models generated by last-generation GWAS (SUPER, FarmCPU, and BLINK) and improvement traditional MLM algorithms (CMLM). These notations show, for each model, the family of GWAS models used and algorithms implemented for random and fixed effects. The models were abbreviated as follows: $I_{M\text{-}Fc\text{-}Rc}$, where "I" refers to the HS index, "M" is the GWAS model family, and "Fc" and "Rc" are the algorithms used to reconstruct the fixed and random covariates, respectively.

TABLE S3 | Genome–environment association (GEA) analyses for heat tolerance according to last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK) for each 270 SNP markers associated with HIT, HSI, and PCA1 indices in 78 common bean accessions based on the optimum association GEA analysis (**Figures 2**–**5**).

TABLE S4 | Summary statistics of last-generation GWAS algorithms (SUPER, FarmCPU, and BLINK). The HS indices and chromosomes (Pv) are presented for each SNP marker associated in 78 common bean accessions based on the optimum association analysis (**Figures 2**–**5**).

TABLE S5 | List of 578 genes flanking 120 associated SNP markers in an expanded genomic window of 81 kb using PhytoMine (see note B) and the reference genome of common bean v2.1

FIGURE S1 | Boxplot, histogram, skewness, kurtosis, and Shapiro–Wilk statistics of the six bioclimatic variables [BIO1 = annual mean temperature **(A)**, BIO5 = maximum temperature of warmer months **(B)**, BIO8 = mean temperature of the wettest quarter **(C)**, BIO9 = mean temperature of the driest quarter **(D)**, BIO10 = mean temperature of the warmest 4-month period **(E)**, and T_j = average of absolute maximum temperature during the reproductive phase **(F)**] and the three HS indices [HSI **(G)**, HIT **(H)**, and PCA1 **(I)**] for the 86 common bean accessions used in this study.

FIGURE S2 | Dispersion diagrams generate by means of Pearson **(A)** and Spearman **(B)** correlations for all bioclimatic variables and between each HS index.

FIGURE S3 | Heat maps of kinship matrices estimated with the VanRaden **(A)**, Loiselle **(B)**, and EMMA **(C)** algorithms across all 23,373 SNP markers.

FIGURE S4 | Manhattan and Q–Q plots of the exploratory phase of genome–environment association (GEA) analysis, for heat tolerance in 78 common bean accessions based on 23,373 SNP markers according to traditional MLM algorithm with the population structure as a fixed effect using the first six principal components (**Figure 1D**). Also, these MLM models use kinship matrix as a random effect by means of Loiselle and VanRaden algorithms. These MLM models are $HSI_{MLM\text{-}PC\text{-}LOISELLE}$ **(A**, **B)**, $HIT_{MLM\text{-}PC\text{-}LOISELLE}$ **(C**, **D)**, $PCA1_{MLM\text{-}PC\text{-}LOISELLE}$ **(E**, **F)**, $HSI_{MLM\text{-}PC\text{-}VANRADEN}$ **(G**, **H)**, $HIT_{MLM\text{-}PC\text{-}VANRADEN}$ **(I**, **J)**, and $PCA1_{MLM\text{-}PC\text{-}VANRADEN}$ (**K**, **L**). The blue dashed horizontal line marks the lax *P*-value threshold. The red dots are SNP markers that systematically crossed the lax threshold in the exploratory phase from all 18 MLM models (S1_42870591 in Pv01 and S1_466464831 and S1_471851336 in Pv11). Black and green colors highlight different common bean (Pv) chromosomes.

FIGURE S5 | Manhattan and Q–Q plots of the exploratory phase of genome–environment association (GEA) analysis, for heat tolerance in 78 common bean accessions based on 23,373 SNP markers according to a traditional MLM algorithm with the population structure using TESS3 (**Figure 1F**) as a fixed effect. Also, these MLM models use kinship matrix as a random effect by means of EMMA and Loiselle algorithms. These MLM models are $HSI_{MLM\text{-}TESS3\text{-}EMMA}$ **(A**, **B)**, $HIT_{MLM\text{-}TESS3\text{-}EMMA}$ **(C**, **D)**, $PCA1_{MLM\text{-}TESS3\text{-}EMMA}$ **(E**, **F)**, $HSI_{MLM\text{-}TESS3\text{-}LOISELLE}$ **(G**, **H)**, $HIT_{MLM\text{-}TESS3\text{-}LOISELLE}$ **(I**, **J)**, and $PCA1_{MLM\text{-}TESS3\text{-}LOISELLE}$ **(K**, **L)**. The blue dashed horizontal line marks the lax *P*-value threshold. The red dots are SNP markers that systematically crossed the lax threshold in the exploratory phase from all 18 MLM models (S1_42870591 in Pv01 and S1_466464831 and S1_471851336 in Pv11). Black and green colors highlight different common bean (Pv) chromosomes.

FIGURE S6 | The Manhattan and Q–Q plots of the exploratory phase of genome–environment association (GEA) analysis, for heat tolerance in 78 common bean accessions based on 23,373 SNP markers according to traditional MLM algorithm with population structure as a fixed effect using TESS3 (**Figure 1F**) and kinship matrix as a random effect using the VanRaden algorithm. These MLM models are $HSI_{MLM\text{-}TESS3\text{-}VANRADEN}$ **(A**, **B)**, $HIT_{MLM\text{-}TESS3\text{-}VANRADEN}$ **(C**, **D)**, $PCA1_{MLM\text{-}TESS3\text{-}VANRADEN}$ **(E**, **F)**. The red dots are SNP markers that systematically crossed the lax threshold in the exploratory phase from all 18 MLM models (S1_42870591 in Pv01 and S1_466464831 and S1_471851336 in Pv11). The Manhattan and Q–Q plots of genome–environment association (GEA) analysis, for heat tolerance in 78 common bean accessions based on 23,373 SNP markers according to compressed MLM algorithms with the population structure using TESS3 (**Figure 1F**) as fixed effect and kinship matrix as a random effect using EMMA algorithm. These CMLM models are $HSI_{CMLM\text{-}TESS3\text{-}EMMA}$ **(G**, **H)**, $HIT_{CMLM\text{-}TESS3\text{-}EMMA}$ **(I**, **J)**, and $PCA1_{CMLM\text{-}TESS3\text{-}EMMA}$ **(K**, **L)**. The blue dashed horizontal line marks the lax *P*-value threshold. Black and green colors highlight different common bean (Pv) chromosomes.

FIGURE S7 | Manhattan and Q–Q plots of genome–environment association (GEA) analysis by means of SUPER algorithm, for heat tolerance in 78 common bean accessions based on 23,373 SNP. GLM model is used in the first step of these nine "failed" SUPER models, and the last step used CMLM (A–F) and MLM (G–P) algorithms. The nine "failed" SUPER are HSISUPER(CMLM)- TESS3-EMMA (A, B), HITSUPER(CMLM)-TESS3-EMMA (C, D), PCA1SUPER(CMLM)-TESS3-EMMA (E, F), HSISUPER(MLM)-PC-EMMA (G, H), HITSUPER(MLM)-PC-EMMA (I, J), PCA1SUPER(MLM)-PC-EMMA (K, L), HSISUPER(MLM)-TESS3-EMMA (M, N), HITSUPER(MLM)-TESS3-EMMA (O, P), and PCA1SUPER(MLM)-TESS3-EMMA (Q, R). Black and green colors highlight different common bean (Pv) chromosomes.

FIGURE S8 | Historical maximum temperature values obtained from the monthly averages from years 1970 to 2000. Extraction of information from WorldClim (see note B). Map construction was done through a customized R-Script using the *raster* package of R v. 3.6.1 (R Core Team).

REFERENCES

Abrol, Y. P., and Ingram, K. T. (1996). "Effects of higher day and night temperatures on growth and yields of some crop plants," in *Global climate change and agricultural production: direct and indirect effects of changing hydrological, pedological, and plant physiological processes*. Eds. F. A. Bazzaz, and W. G. Sombroek (West Sussex: Food and Agriculture Organization of the United Nations), 345.

Ahmad, A., Diwan, H., and Abrol, Y. P. (2009). "Global climate change, stress and plant productivity," in *Abiotic stress adaptation in plants.* (Dordrecht: Springer Netherlands), 503–521. doi: 10.1007/978-90-481-3112-9_23

Ariani, A., Berny Mier y Teran, J. C., and Gepts, P. (2018). Spatial and Temporal Scales of Range Expansion in Wild *Phaseolus vulgaris*. *Mol. Biol. Evol.* 35, 119–131. doi: 10.1093/molbev/msx273

Bae, W., Lee, Y. J., Kim, D. H., Lee, J., Kim, S., Sohn, E. J., et al. (2008). AKR2A-mediated import of chloroplast outer membrane proteins is essential for chloroplast biogenesis. *Nat. Cell Biol.* 10, 220–227. doi: 10.1038/ncb1683

Barbulescu, D. M., Fikere, M., Malmberg, M. M., Spangenberg, G. C., Cogan, N. O., and Daetwyler, H. D. (2018). Imputation to whole-genome sequence increases the power of genome wide association studies for blackleg resistance in canola. *AusCanola 2018 Co-hosts* 29.

Bita, C. E., and Gerats, T. (2013). Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. *Front. Plant Sci.* 4, 273. doi: 10.3389/fpls.2013.00273

Bitocchi, E., Nanni, L., Bellucci, E., Rossi, M., Giardini, A., Zeuli, P. S., et al. (2012). Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 109, E788–E796. doi: 10.1073/pnas.1108973109

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Büttner, M., and Singh, K. B. (1997). *Arabidopsis thaliana* ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5961–5966. doi: 10.1073/pnas.94.11.5961

Camejo, D., Jiménez, A., Alarcón, J. J., Torres, W., Gómez, J. M., and Sevilla, F. (2006). Changes in photosynthetic parameters and antioxidant activities following heat-shock treatment in tomato plants. *Funct. Plant Biol.* 33, 177. doi: 10.1071/FP05067

Caramori, P. H., Gonçalves, S. L., Wrege, M. S., Henrique, J., Oliveira, D., De, R. T., et al. (2001). Zoneamento de riscos climáticos e definição de datas de semeadura para o feijão no Paraná Climatic risk zoning and definition of best sowing dates for common beans in Paraná state, Brazil. *Rev. Bras. Agrometeorol.* 9, 477–485.

Caye, K., Deist, T. M., Martins, H., Michel, O., and François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* 16, 540–548. doi: 10.1111/1755-0998.12471

Cortés, A. J., and Blair, M. W. (2018). Genotyping by sequencing and genome–environment associations in wild common bean predict widespread divergent adaptation to drought. *Front. Plant Sci.* 9, 128. doi: 10.3389/fpls.2018.00128

Cortés, A. J., Monserrate, F. A., Ramírez-Villegas, J., Madriñán, S., and Blair, M. W. (2013). Drought tolerance in wild plant populations: the case of common beans (Phaseolus vulgaris L.). *PLoS One* 8 (5), e62898. doi: 10.1371/journal.pone.0062898

Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *J. Crop Improv.* 25, 239–261. doi: 10.1080/15427528.2011.558767

De Mendiburu, F. (2014). Agricolae: statistical procedures for agricultural research. R package version, 1(1). Available at: http://tarwi.lamolina.edu.pe/~fmendiburu%0D

Dixit, A. R., and Dhankher, O. P. (2011). A novel stress-associated protein 'AtSAP10' from *Arabidopsis thaliana* confers tolerance to nickel, manganese, zinc, and high temperature stress. *PLoS One* 6, e20921. doi: 10.1371/journal.pone.0020921

Ellis, C. M., Nagpal, P., Young, J. C., Hagen, G., Guilfoyle, T. J., and Reed, J. W. (2005). AUXIN RESPONSE FACTOR1 and AUXIN RESPONSE FACTOR2 regulate senescence and floral organ abscission in *Arabidopsis thaliana*. *Development* 132, 4563–4574. doi: 10.1242/DEV.02012

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Evrard, A., Kumar, M., Lecourieux, D., Lucks, J., von Koskull-Döring, P., and Hirt, H. (2013). Regulation of the heat stress response in *Arabidopsis* by MPK6-targeted phosphorylation of the heat stress factor HsfA2. *PeerJ* 1, e59. doi: 10.7717/peerj.59

FAO (2018). FAOSTAT. NAME OF DATA COLLECTION (production indices). http://www.fao.org/faostat/. Available at: http://www.fao.org/faostat/en/#data/QI/metadata (Accessed July 27, 2018).

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120. doi: 10.1111/mec.13476

Frank, A., Oddou-Muratorio, S., Lalagüe, H., Pluess, A. R., Heiri, C., and Vendramin, G. G. (2016). Genome–environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. *New Phytol.* 210, 589–601. doi: 10.1111/nph.13809

Frichot, E., and François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi: 10.1111/2041-210X.12382

Gentry, H. S. (1969). Origin of the common bean, *Phaseolus vulgaris*. *Econ. Bot.* 23, 55–69. doi: 10.1007/BF02862972

Gollan, P. J., Bhave, M., and Aro, E.-M. (2012). The FKBP families of higher plants: exploring the structures and functions of protein interaction specialists. doi: 10.1016/j.febslet.2012.09.002

Gonçalves, S. L., Wrege, M. S., Caramori, P. H., Mariot, E. J., and Abucarub Neto, M. (1997). Probabilidade de ocorrência de temperaturas superiores a 30 C no florescimento do feijoeiro (*Phaseolus vulgaris* L.), cultivado na safra das águas no estado do Paraná. *Rev. Bras. Agrometeorol.* 5, 99–107.

González-Carranza, Z. H., Whitelaw, C. A., Swarup, R., and Roberts, J. A. (2002). Temporal and spatial expression of a polygalacturonase during leaf and flower abscission in oilseed rape and *Arabidopsis*. *Plant Physiol.* 128, 534–543. doi: 10.1104/pp.010610

Hagen, G., and Guilfoyle, T. (2002). Auxin-responsive gene expression: genes, promoters and regulatory factors. *Plant Mol. Biol.* 49, 373–385. doi: 10.1023/A:1015207114117

Hanks, S. K., and Quinn, A. M. (1991). [2] Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* 200, 38–62. doi: 10.1016/0076-6879(91)00126-H

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* 8. doi: 10.1093/gigascience/giy154

Ikeda, M., Mitsuda, N., and Ohme-Takagi, M. (2011). *Arabidopsis* HsfB1 and HsfB2b act as repressors of the expression of heat-inducible Hsfs but positively regulate the acquired thermotolerance. *Plant Physiol.* 157, 1243–1254. doi: 10.1104/PP.111.179036

Jofuku, K. D., den Boer, B. G., Van Montagu, M., and Okamuro, J. K. (1994). Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2. *Plant Cell* 6, 1211–1225. doi: 10.1105/tpc.6.9.1211

Jones, A. L. (1999). *PHASEOLUS BEAN: Post-harvest Operations*. Mejia, B, Danilo, and Lewis, Eds. Centro internacional de agricultura tropical (CIAT). www.cgiar.org/ciat Available at: http://www.fao.org/3/a-av015e.pdf.

Joo, J. W. J., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biol.* 17 (1), 62. doi: 10.1186/s13059-016-0903-6

Khraiwesh, B., Zhu, J.-K., and Zhu, J. (2012). Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim. Biophys. Acta—Gene Regul. Mech.* 1819, 137–148. doi: 10.1016/J.BBAGRM.2011.05.001

Kim, T. W., Kwon, Y.-J., Kim, J. M., Song, Y.-H., Kim, S. N., and Kim, Y.-J. (2004). MED16 and MED23 of mediator are coactivators of lipopolysaccharide- and heat-shock-induced transcriptional activators. *Proc. Natl. Acad. Sci. U. S. A.* 101, 12153–12158. doi: 10.1073/pnas.0401985101

Kotak, S., Larkindale, J., Lee, U., von Koskull-Döring, P., Vierling, E., and Scharf, K.-D. (2007). Complexity of the heat stress response in plants. *Curr. Opin. Plant Biol.* 10, 310–316. doi: 10.1016/J.PBI.2007.04.011

Lang-Mladek, C., Popova, O., Kiok, K., Berlinger, M., Rakic, B., Aufsatz, W., et al. (2010). Transgenerational inheritance and resetting of stress-induced loss of epigenetic gene silencing in *Arabidopsis*. *Mol. Plant* 3, 594–602. doi: 10.1093/MP/SSQ014

Larkindale, J., and Huang, B. (2004). Thermotolerance and antioxidant systems in *Agrostis stolonifera*: involvement of salicylic acid, abscisic acid, calcium, hydrogen peroxide, and ethylene. *J. Plant Physiol.* 161, 405–413. doi: 10.1078/0176-1617-01239

Larkindale, J., and Knight, M. R. (2002). Protection against heat stress-induced oxidative damage in *Arabidopsis* involves calcium, abscisic acid, ethylene, and salicylic acid. *Plant Physiol.* 128, 682–695. doi: 10.1104/pp.010320

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome–environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1, e1400218. doi: 10.1126/sciadv.1400218

LeBlanc, C., Zhang, F., Mendez, J., Lozano, Y., Chatpar, K., Irish, V. F., et al. (2018). Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *Plant J.* 93, 377–386. doi: 10.1111/tpj.13782

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12, 73. doi: 10.1186/s12915-014-0073-5

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12, e1005767. doi: 10.1371/journal.pgen.1005767

Lopes-Caitar, V. S., de Carvalho, M. C., Darben, L. M., Kuwahara, M. K., Nepomuceno, A. L., Dias, W. P., et al. (2013). Genome-wide analysis of the *Hsp20* gene family in soybean: comprehensive sequence, genomic organization and expression profile analysis under abiotic and biotic stresses. *BMC Genomics* 14, 577. doi: 10.1186/1471-2164-14-577

Malik, M. K., Slovin, J. P., Hwang, C. H., and Zimmerman, J. L. (1999). Modified expression of a carrot small heat shock protein gene, *Hsp17.7*, results in increased or decreased thermotolerance. *Plant J.* 20, 89–99. doi: 10.1046/j.1365-313X.1999.00581.x

Mathur, S., Vyas, S., Kapoor, S., and Tyagi, A. K. (2011). The mediator complex in plants: structure, phylogeny, and expression profiling of representative genes in a dicot (*Arabidopsis*) and a monocot (rice) during reproduction and abiotic stress. *Plant Physiol.* 157, 1609–1627. doi: 10.1104/pp.111.188300

Monterroso, V. A., and Wien, H. C. (2019). Flower and pod abscission due to heat stress in beans. *J. Am. Soc. Hortic. Sci.* 115, 631–634. doi: 10.21273/jashs.115.4.631

Oladzad, A., Porch, T., Rosas, J. C., Moghaddam, S. M., Beaver, J., Beebe, S. E., et al. (2019). Single and multi-trait GWAS identify genetic factors associated with production traits in common bean under abiotic stress environments. *Genes Genomes Genet.* 9 (6), 1881–1892. doi: 10.1534/g3.119.400072

Pachico, D. (1993). *The demand for bean technology*. Cali, Colombia: CIAT.

Pasam, R. K., Sharma, R., Malosetti, M., van Eeuwijk, F. A., Haseneyer, G., Kilian, B., et al. (2012). Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol.* 12, 16. doi: 10.1186/1471-2229-12-16

Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., and Mittelsten Scheid, O. (2010). Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in *Arabidopsis*. *Plant Cell* 22, 3118–3129. doi: 10.1105/tpc.110.078493

Peng, M., Cui, Y., Bi, Y.-M., and Rothstein, S. J. (2006). AtMBD9: a protein with a methyl-CpG-binding domain regulates flowering time and shoot branching in *Arabidopsis*. *Plant J.* 46, 282–296. doi: 10.1111/j.1365-313X.2006.02691.x

Peters, C., Li, M., Narasimhan, R., Roth, M., Welti, R., and Wang, X. (2010). Nonspecific phospholipase C NPC4 promotes responses to abscisic acid and tolerance to hyperosmotic stress in *Arabidopsis*. *Plant Cell* 22, 2642–2659. doi: 10.1105/TPC.109.071720

Porch, T. G. (2006). Application of stress indices for heat tolerance screening of common bean. *J. Agron. Crop Sci.* 192, 390–394. doi: 10.1111/j.1439-037X.2006.00229.x

Porch, T. G., and Jahn, M. (2001). Effects of high-temperature stress on microsporogenesis in heat-sensitive and heat-tolerant genotypes of *Phaseolus vulgaris*. *Plant Cell Environ.* 24, 723–731. doi: 10.1046/j.1365-3040.2001.00716.x

Qi, A., Smithson, J. B., and Summerfield, R. J. (1998). Adaptation to climate in common bean (*Phaseolus vulgaris* L.): photothermal flowering responses in the Eastern, Southern and Great Lakes regions of Africa. *Exp. Agric.* 34, 153–170. doi: 10.1017/S0014479798002026

Rainey, K. M., and Griffiths, P. D. (2019). Inheritance of heat tolerance during reproductive development in snap bean (*Phaseolus vulgaris* L.). *J. Am. Soc. Hortic. Sci.* 130, 700–706. doi: 10.21273/jashs.130.5.700

Rosas, J. C., Castro, A., Beaver, J. S., and Lepiz, R. (2000). Tolerancia a altas temperaturas y resistencia a mosaico dorado en frijol común. *Agron. Mesoam.* 11, 1–10. doi: 10.15517/am.v11i1.17327

Rossi, M., Bitocchi, E., Bellucci, E., Nanni, L., Rau, D., Attene, G., et al. (2009). Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evol. Appl.* 2 (4), 504–522. doi: 10.1111/j.1752-4571.2009.00082.x

Savchenko, G. E., Klyuchareva, E. A., Abramchik, L. M., and Serdyuchenko, E. V. (2002). Effect of periodic heat shock on the inner membrane system of etioplasts. *Russ. J. Plant Physiol.* 49, 349–359. doi: 10.1023/A:1015592902659

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713. doi: 10.1038/ng.3008

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Sgarbieri, V. C., and Whitaker, J. R. (1982). Physical, chemical, and nutritional properties of common bean (*Phaseolus*) proteins. *Adv. Food Res.* 28, 93–166. doi: 10.1016/S0065-2628(08)60111-1

Silva, J. C., Heldwein, A. B., Martins, F. B., Streck, N. A., and Guse, F. I. (2007). Risco de estresse térmico para o feijoeiro em Santa Maria, RS. *Ciência Rural* 37, 643–648. doi: 10.1590/s0103-84782007000300007

Soltani, A., MafiMoghaddam, S., Walter, K., Restrepo-Montoya, D., Mamidi, S., Schroder, S., et al. (2017). Genetic architecture of flooding tolerance in the dry bean Middle-American diversity panel. *Front. Plant Sci.* 8, 1183. doi: 10.3389/fpls.2017.01183

Soltani, A., MafiMoghaddam, S., Oladzad Abbasabadi, A., Walter, K., Kearns, P. J., Vasquez Guzman, J., et al. (2018). Genetic analysis of flooding tolerance in an Andean diversity panel of dry bean (*Phaseolus vulgaris* L.). *Front. Plant Sci.* 9, 767. doi: 10.3389/fpls.2018.00767

Soltani, A., Weraduwage, S. M., Sharkey, T. D., and Lowry, D. B. (2019). Elevated temperatures cause loss of seed set in common bean (*Phaseolus vulgaris* L.) potentially through the disruption of source–sink relationships. *BMC Genomics* 20 (1). doi: 10.1186/s12864-019-5669-2

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9, 0. doi: 10.3835/plantgenome2015.11.0120

Thornthwaite, C. (1948). *An approach toward a rational classification of climate*. Available at: https://journals.lww.com/soilsci/Citation/1948/07000/An_Approach_Toward_a_Rational_Classification_of.7.aspx (Accessed May 22, 2019).

Tohme, J., Gonzalez, D. O., Beebe, S., and Duque, M. C. (1996). AFLP analysis of gene pools of a wild bean core collection. *Crop Sci.* 36, 1375. doi: 10.2135/cropsci1996.0011183X003600050048x

Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., et al. (2007). Observations: surface and atmospheric climate change. *Phys. Sci. Basis* 1, 235–336.

Trost, G., Vi, S. L., Czesnick, H., Lange, P., Holton, N., Giavalisco, P., et al. (2014). *Arabidopsis* poly(A) polymerase PAPS1 limits founder-cell recruitment to organ primordia and suppresses the salicylic acid-independent immune response downstream of EDS1/PAD4. *Plant J.* 77, 688–699. doi: 10.1111/tpj.12421

Usadel, B., Schlüter, U., Mølhøj, M., Gipmans, M., Verma, R., Kossmann, J., et al. (2004). Identification and characterization of a UDP-*d*-glucuronate 4-epimerase in *Arabidopsis*. *FEBS Lett.* 569, 327–331. doi: 10.1016/j.febslet.2004.06.005

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* 35, 969–976. doi: 10.1038/nbt.3943

Vierling, E. (1991). The roles of heat shock proteins in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 42, 579–620. doi: 10.1146/annurev.pp.42.060191.003051

Wang, N., Yue, Z., Liang, D., and Ma, F. (2014a). Genome-wide identification of members in the YTH domain-containing RNA-binding protein family in apple and expression analysis of their responsiveness to senescence and abiotic stresses. *Gene* 538, 292–305. doi: 10.1016/J.GENE.2014.01.039

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014b). A SUPER powerful method for genome wide association study. *PLoS One* 9, e107684. doi: 10.1371/journal.pone.0107684

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* 9, 244–252. doi: 10.1016/J.TPLANTS.2004.03.006

Wantanbe, H. (1953). Studies on the unfruitfulness of beans. 3. Influences of temperature on blooming and of relative humidity on the pollen activities of beans. *J. Hortic. Ass. Jpn.* 22, 172–176. doi: 10.2503/jjshs.22.172

Warpeha, K. M., Upadhyay, S., Yeh, J., Adamiak, J., Hawkins, S. I., Lapik, Y. R., et al. (2007). The GCR1, GPA1, PRN1, NF-Y signal chain mediates both blue light and abscisic acid responses in *Arabidopsis*. *Plant Physiol.* 143, 1590–1600. doi: 10.1104/pp.106.089904

Weaver, M., and Timm, H. (1988). Influence of temperature and plant water status on pollen viability in beans. *J. Am. Soc. Hortic. Sci.* 1, 31–35.

Weis, B. L., Palm, D., Missbach, S., Bohnsack, M. T., and Schleiff, E. (2015). atBRX1-1 and atBRX1-2 are involved in an alternative rRNA processing pathway in *Arabidopsis thaliana*. *RNA* 21, 415–425. doi: 10.1261/rna.047563.114

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zuiderveen, G. H., Padder, B. A., Kamfwa, K., Song, Q., and Kelly, J. D. (2016). Genome-wide association study of anthracnose resistance in Andean beans (*Phaseolus vulgaris*). *PLoS One* 11, e0156391. doi: 10.1371/journal.pone.0156391

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 López-Hernández and Cortés. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 November 2019
doi: 10.3389/fgene.2019.01224

GWAS-Assisted Genomic Prediction to Predict Resistance to Septoria Tritici Blotch in Nordic Winter Wheat at Seedling Stage

*Firuz Odilbekov[1], Rita Armonienė[1,2], Alexander Koc[1], Jan Svensson[3] and Aakash Chawade[1]**

[1] Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden, [2] Institute of Agriculture, Lithuanian Research Centre for Agriculture and Forestry (LAMMC), Akademija, Lithuania, [3] Nordic Genetic Resource Centre, Alnarp, Sweden

OPEN ACCESS

Edited by:
Alison Bentley, National Institute of Agricultural Botany (NIAB), United Kingdom

Reviewed by:
Thomas Miedaner, University of Hohenheim, Germany
Matthew Rouse, United States Department of Agriculture, United States
Morten Lillemo, Norwegian University of Life Sciences, Norway

***Correspondence:**
Aakash Chawade
aakash.chawade@slu.se

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *01 June 2019*
Accepted: *05 November 2019*
Published: *26 November 2019*

Citation:
Odilbekov F, Armonienė R, Koc A, Svensson J and Chawade A (2019) GWAS-Assisted Genomic Prediction to Predict Resistance to Septoria Tritici Blotch in Nordic Winter Wheat at Seedling Stage. Front. Genet. 10:1224. doi: 10.3389/fgene.2019.01224

Septoria tritici blotch (STB) disease caused by *Zymoseptoria tritici* is one of the most damaging diseases of wheat causing significant yield losses worldwide. Identification and employment of resistant germplasm is the most cost-effective method to control STB. In this study, we characterized seedling stage resistance to STB in 175 winter wheat landraces and old cultivars of Nordic origin. The study revealed significant ($p < 0.05$) phenotypic differences in STB severity in the germplasm. Genome-wide association analysis (GWAS) using five different algorithms identified ten significant markers on five chromosomes. Six markers were localized within a region of 2 cM that contained seven candidate genes on chromosome 1B. Genomic prediction (GP) analysis resulted in a model with an accuracy of 0.47. To further improve the prediction efficiency, significant markers identified by GWAS were included as fixed effects in the GP model. Depending on the number of fixed effect markers, the prediction accuracy improved from 0.47 (without fixed effects) to 0.62 (all non-redundant GWAS markers as fixed effects), respectively. The resistant genotypes and single-nucleotide polymorphism (SNP) markers identified in the present study will serve as a valuable resource for future breeding for STB resistance in wheat. The results also highlight the benefits of integrating GWAS with GP to further improve the accuracy of GP.

Keywords: GWAS - genome-wide association study, genomic prediction (GP), genomic selection (GS), wheat, Septoria tritici blotch (STB), Quantitative trait loci (QTL)

INTRODUCTION

Septoria tritici blotch (STB) disease caused by fungal pathogen *Zymoseptoria tritici* is one of the devastating foliar diseases of wheat in the temperate regions worldwide. STB causes significant yield losses and additional fungicide expenses (Fones and Gurr, 2015; Torriani et al., 2015). The annual harvest losses reach 5% to 10% in the biggest EU wheat producing countries (Fones and Gurr, 2015). Cultivation of resistant cultivars in combination with fungicide application is the main strategy to control the disease. Besides, a major problem of the intensive use of fungicides is that many populations of *Z. tritici* have rapidly evolved resistance to its active agents (Torriani et al., 2009; Wieczorek et al., 2015; Cheval et al., 2017). Therefore, novel sources of resistance to STB

and their introgression into wheat breeding programs is the most economical and environmentally friendly strategy for effective management of the disease.

So far, 21 genes are mapped for resistance to STB in wheat (Brown et al., 2015). The expression pattern and effect of these genes on resistance to STB differ in seedling and adult plant stages. For example, *Stb16* is expressed and effective at the seedling and adult stages of plants while *Stb17* is expressed only at the adult stage (Tabib Ghaffary et al., 2011). *Stb18* is an isolate-specific resistance gene that shows variable resistance to *Z. tritici* at seedling and adult stages depending on the isolate (Tabib Ghaffary et al., 2011). *Stb6* and *Stb15* are the two most common STB resistance genes in the current European germplasm (Arraiano and Brown, 2006). *Stb15* was found in about 60% of cultivars tested but, unlike *Stb6, Stb15* is not known to show resistance under field conditions (Arraiano et al., 2009; Brown et al., 2015). The only qualitative gene for STB resistance *Stb6* (Saintenac et al., 2018) and recently identified avirulence gene *AvrStb6* of *Z. tritici* (Zhong et al., 2017) have been shown to be in a gene-for-gene relationship. *Stb6* is among the most frequent STB genes in European wheat germplasm and suggested as the most widespread STB gene in the contemporary wheat breeding programs (Arraiano and Brown, 2017). However, this gene alone is not sufficient to provide adequate resistance to STB, and there are no other known resistance genes contributing significantly to the reduction of *Z. tritici* populations in Europe (Arraiano et al., 2009). The majority of variation in field resistance to STB is controlled by quantitative resistance, and the progress in breeding for STB resistance over the last 30 years presumably happened by the gradual accumulation of minor genes. Recently, it was shown that the STB disease symptoms chlorosis, necrosis, and pycnidia are under varying genetic control (Odilbekov et al., 2019). Therefore, there is a need to search for new sources of durable disease resistance to STB for marker-assisted introgression into elite wheat cultivars (Fones and Gurr, 2015; McDonald and Mundt, 2016).

Wheat landraces are a valuable source of genetic diversity. They are adapted to the environmental conditions of their place of origin and thus can provide novel sources of disease resistance for developing new cultivars adapted to the changing climate (de Carvalho et al., 2012; Lopes et al., 2015). Several useful agronomic and resistance traits have been introgressed from landraces to commercial wheat cultivars including the dwarfing gene *Rht* from the Japanese landrace Shiro Daruma (Dreisigacker et al., 2005) and the high grain protein content gene *NAM-B1* in Fennoscandian wheat (Hagenblad et al., 2012). Valuable landraces and old cultivars of winter wheat consisting of more than 300 genotypes from Scandinavian countries is preserved at the Nordic Genetic Resource Centre (NordGen, Alnarp, Sweden), and part of this material was evaluated earlier for several agronomic traits and showed high diversity in morphological traits (Diederichsen et al., 2012), resistance to rust (Randhawa et al., 2016) and powdery mildew (Hysing et al., 2007). These studies prove that the material stored at NordGen is unique and a genetically diverse resource, which can be utilized for the improvement of wheat cultivars for Nordic and Baltic Sea Region countries (Chawade et al., 2018).

Genome-wide association studies (GWAS) and genomic selection (GS), both performed with genome-wide markers are important and effective tools for plant breeding. GWAS estimates marker effects across the whole genome on the target population based on prediction models (Desta and Ortiz, 2014). Based on linkage disequilibrium (LD), GWAS can identify new functional alleles (identify novel genes and QTLs) for many agriculturally important traits in diverse germplasm. Few GWAS studies were performed for STB resistance in European winter wheat accessions (Kollers et al., 2013; Miedaner et al., 2013; Vagndorf et al., 2017). Many regions associated with resistance to STB in the wheat genome were identified in these studies. In a study of 1,055 elite hybrids and their corresponding 87 parental lines, Miedaner et al. (2013) identified four significant single-nucleotide polymorphisms (SNP) associated with STB resistance located on chromosomes 1B, 2B, 5B, and 6A. Kollers et al. (2013) detected 39 SSR on 2A, 2D, 3A, 5B, 7A, 7D significantly associated with adult plant resistance in a panel of 372 European wheat lines. Four QTL, on chromosomes 1B, 2A, 5D, and 7A were highly associated with STB resistance in 164 North European cultivars and breeding lines (Vagndorf et al., 2017).

GS, on the other hand, enables the selection of superior genotypes based on genomic estimated breeding values (GEBV) to create models for the prediction of phenotypes in uncharacterized populations (Meuwissen et al., 2001). Previous studies have shown the feasibility of GS for predicting STB resistance in wheat. Juliana et al. (2017) achieved a mean genomic prediction (GP) accuracy of 0.45 for adult plant resistance to STB in a population of 333 and 314 advanced lines from Centro Internacional de Mejoramiento de Maíz y Trigo's (CIMMYT) wheat breeding program. Muqaddasi et al. (2019) investigated the potential of GP of adult stage STB infection in a European winter wheat panel of 371 elite varieties, resulting in both additive and non-additive prediction models centered around a mean GP accuracy of approximately 0.43. Spindel et al. (2016) described the new combined GS + GWAS model based only on the results of GWAS run using GS training population data. GS + GWAS has some benefits as the method does not require additional data as the same phenotype and genotype data set is used, prediction accuracy can be enhanced, and it can be more accessible to breeders as it does not require extensive knowledge on the underlying genetics of a trait of interest (Spindel et al., 2016).

Previous studies were primarily focused on resistance to STB in the adult stage of winter wheat germplasm. One of the main goals of this project was to characterize seedling stage resistance to STB in winter wheat landraces and old cultivars of Nordic origin which are well adapted to the Nordic climate. The current study relies on a collection of 175 winter wheat accessions, released between 1900 and 2012. In this work, this material was evaluated for seedling-stage resistance to STB disease. The objectives of this study were (i) to detect novel STB disease resistance loci at the seedling stage by performing GWAS analysis; (ii) to identify candidate genes to STB resistance in wheat; (iii) to evaluate GP (GP) for selection for STB resistance; and (iv) to employ GP+GWAS to further improve the accuracy of GP.

MATERIALS AND METHODS

Plant Material

The material in this study comprised of 175 winter wheat cultivars and landraces (hereafter genotypes) mainly from Scandinavian countries (**Supplementary Table 1**). The collected genotypes were released between 1900 and 2012 and representing a century of winter wheat breeding history of the region. Four genotypes originating from Germany were also included as they have been widely grown in the Scandinavian area. The seeds were obtained from Nordic Genetic Resources Centre, Alnarp, Sweden (NordGen).

Growth Conditions

The seeds were placed on a moist filter paper in Petri dishes and kept for 4 days at +4°C in dark. Afterwards, they were transferred to room temperature conditions for two days for germination. Thereafter, the germinated seeds were sown in plastic pots (8 × 8 × 8 cm) filled with peat substrate. Two seeds of each genotype were sown per pot. Plants were grown in the Biotron greenhouse chamber at 24°C with a 16-h photoperiod and 60% humidity. The light intensity was set and controlled at 250 μmol $m^{-2}{\cdot}s^{-1}$. The samples were arranged in an augmented design with eight blocks designed with the R package Agricolae (Mendiburu, 2017). Four genotypes were used as checks in each block to control block effect, namely, Nimbus (susceptible), Nelson and Target (moderately resistant), and Kranich (resistant). The entire experiment was performed twice with 1-month interval, and two replications were done at each occasion.

Inoculation and Disease Assessment

The *Z. tritici* strain was isolated from typical STB lesions on leaves of winter wheat collected in southern part of Sweden during 2015, and the inoculum was prepared as described previously (Odilbekov et al., 2018). Second and third leaves of the seedlings were marked close to the stem with a permanent marker before inoculation. The twentyone day old wheat seedlings were inoculated with *Z. tritici* inoculum using a hand sprayer with a spore concentration of 10^7 spores ml^{-1}. The inoculum was applied on the leaves three times, and leaves were allowed to dry for 1 h each time. The inoculated seedlings were transferred to fully controlled daylight chamber and kept 72 h under close to 100% relative humidity at 24°C with a 16-h photoperiod and a light intensity of 250 μmol $m^{-2}{\cdot}s^{-1}$. Relative humidity was reduced to 65% 72 h post-inoculation. Percentage of the necrotic area on the inoculated leaf surface (from 0% to 100%) was visually scored at 13, 16, and 19 days post-inoculation (dpi). The lesion development over the assessment period was summarized through the computation of the relative area under the disease progress curve (rAUDPC). The entire experiment was repeated twice.

Genotypic Data and Population Structure

The samples for DNA extraction were collected from 6-week-old seedling and the DNA extraction and genotyping of the samples was performed by TraitGenetics GmbH, Germany (http://www.traitgenetics.com/en/). The samples were genotyped with a 20K SNP wheat marker array. A total of 6,097 SNPs were used for GWAS after removing SNPs with more than 20% missing data as well as a minor allele frequency less than 5%. Principal component analysis (PCA) was done with the software Simca 14 (Umetrics, Sweden).

GWAS and GP

GWAS analysis was done with the GAPIT package (v3.0) in R (Tang et al., 2016). The primary model was constructed with the GLM algorithm (Lipka et al., 2012) with 10 principal components as covariates and MAF threshold of 0.05. A QTL was considered significant at the threshold of adjusted false discovery rate (FDR) < 0.05. New GWAS models were developed using MLM, MLMM, FarmCPU, and Super algorithm in GAPIT for verification of the QTL obtained with GLM. GP modeling was done using the R package rrBLUP (v4.6) (Endelman, 2011) for ridge-regression-based genome-wide regression. The rrBLUP model for genome-wide regression assumes the form $y = Xb + Zu$, where X and Z are the design matrices for fixed and random effects, respectively, b and u are vectors of fixed and random effects, and y is a vector of phenotypic values. Similar to the method proposed by Spindel et al. (2016), significant markers identified by GWAS results were included as fixed effects in the GS model and removed from the design matrix of random effects. To identify the best subset of GWAS-selected markers to include as fixed effects, all possible permutations of available GWAS-selected markers, were evaluated with respect to average model accuracy. Number of markers in the marker sets ranged from one (a single marker added as fixed effect) to five (all available markers). The GP models were validated on a set of 500 random 80/20 train/test set splits. Model accuracy was assessed by calculating Pearson's correlation coefficient between the predicted and observed STB resistance for each of the train/test sets and estimating the average of all correlation values for each run. The best performing model was selected on the basis of the highest average model accuracy. The GP models with markers fitted as fixed effects were compared to a GP model which did not use GWAS-selected markers as fixed effects, instead fitting all available markers as random effects, and was also compared to models that mimicked the model configuration of the fixed effect models described above, but which instead sampled random markers (as opposed to selecting markers based on highest significance in a GWAS). The subset sizes used for the models using the randomly selected markers ranged from one to five. Each subset size was evaluated five times, with a new random draw of markers. The initially described model which fit all markers as random effects, and the models fitting randomly selected markers as fixed effects, were all validated against the same 500 train/test splits as the GWAS-selected marker models.

Identification of Candidate Genes

The physical positions of the significant markers from the GLM model were identified by BLASTing their sequences against the IWGSC RefSeq v1.0 genome. The physical location of flanking markers BobWhite_c42716_71 and BS00110231_51

fell into range of 623,712,765 to 623,989,423 bp in the region of chromosome 1B. The candidate genes physically located within this range were identified, and their gene annotation was extracted from IWGSC RefSeq v1.0 genome.

RESULTS

Phenotypic Diversity

The *Z. tritici* isolate was evaluated on a differential set of wheat cultivars with known *Stb* resistance genes (**Supplementary Figure 1**). The evaluation of 175 genotypes showed that the phenotypic distribution of STB severity followed approximately a normal distribution (**Figure 1**). Highly significant ($p < 0.05$) phenotypic differences in STB severity were observed in the germplasm (**Table S1**). The mean of the rAUDPC values ranged from 0.33 for the most resistant and 2.07 for most susceptible genotypes, respectively. Tukey multiple comparison test showed that the genotypes Kranich, Starke, Galicia, and Cymbal exhibited a higher level of resistance to STB while the lower level of resistance was found in genotypes such as Penta, Sejet, Svea I, and Gluten.

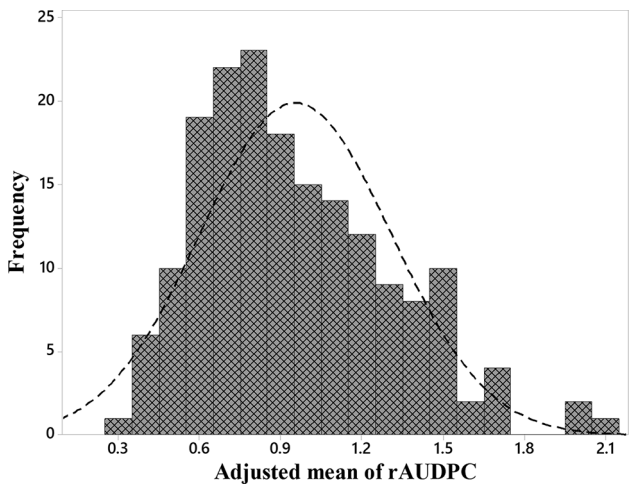


FIGURE 1 | Frequency distribution of adjusted rAUDPC mean of STB score of two greenhouse experiments.

Population Structure

To identify underlying genetic differences, PCA and Kinship analyses were performed on the genotypes based on 6,097 SNPs. The first and second principal components accounted for 12.3% and 10.03% of the variance, respectively. The genotypes were clustered into three major groups, and the clustering was mainly based on their geographic origin (**Figure 2A**). The genotypes with origin from Denmark and Finland formed two very distinct clusters, whereas the Swedish genotypes could be considered intermediate between these two clusters. The result from PCA revealed that most of the genotypes with a higher level of resistance belong to the modern wheat cultivars while most of the susceptible genotypes belonged to older released ones (**Figures 2B**, **C**). A similar result to PCA was also observed by using Kinship analysis where three different clusters were identified (**Figure 3**).

Genome-Wide Association Analysis

The GWAS was performed using the GLM model, and both population structure and kinship (K) were taken into account to control pseudo associations (**Figure 4**). As is shown by the Manhattan plot and quantile-quantile plot (QQ plot) (**Figures 4A**, **D**), six significant (FDR < 0.05) SNP markers for rAUDPC of STB were detected on chromosome 1B. The identified QTL was verified using four additional GWAS models, namely, MLM, MLMM, FarmCPU, and Super and the QTL was found to be statistically significant (FDR < 0.05) in MLMM, FarmCPU, and Super results (**Supplementary Figure 2**). All six markers are located within a 2 cM distance on chromosome 1B (97–99 cM), thus, suggesting that it could potentially be a single QTL (**Table 1**, **Figure 5**). Additional QTL were also identified on chromosome 1A, 2B, 3A, and 5A in at least two GWAS models each (**Table 1**).

Candidate Genes Located in the QTL on Chromosome 1B

In total, seven candidate genes were identified that were localized within the GWAS identified loci on chromosome 1B (**Figure 5**).

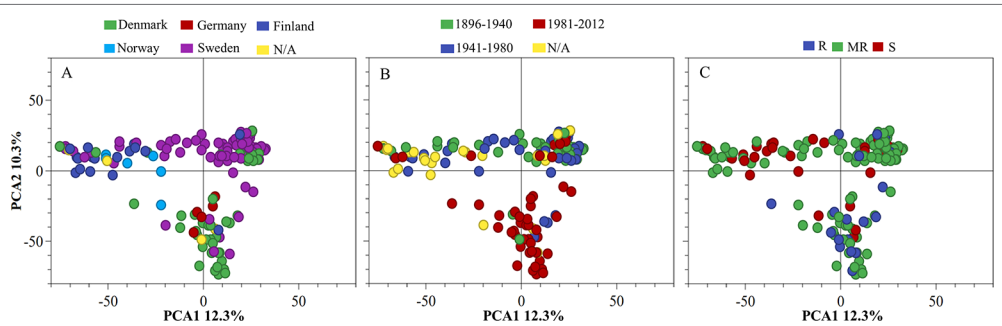


FIGURE 2 | Principal component analysis (PCA) of 175 winter wheat cultivars/landraces coloured and labelled by **(A)** country of origin, **(B)** released year and **(C)** resistance/susceptibility. PCA was based on the allele frequencies of 6097 SNP markers. R, resistant; MR, moderate resistant; S, susceptible.

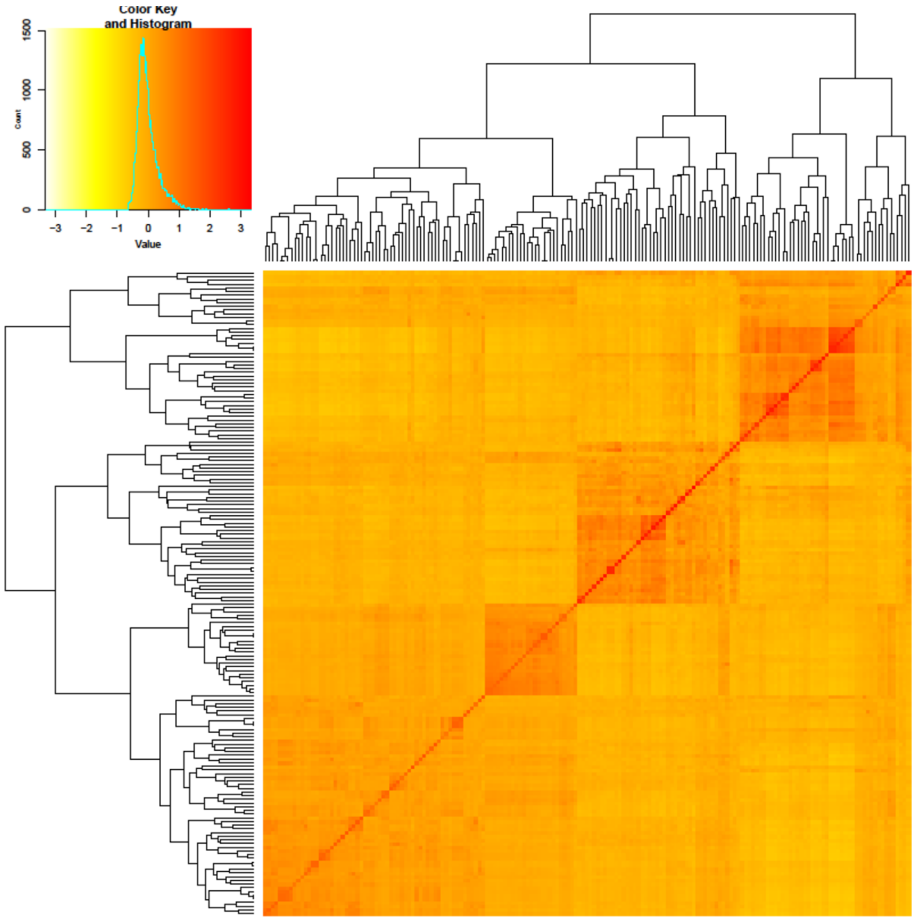


FIGURE 3 | Heatmap and dendrogram of a kinship matrix among 175 winter wheat cultivars/landraces estimated using the SNP data.

Among these genes, two genes code for F-box protein (Traes CS1B01G390100, TraesCS1B01G390500) and two genes for ATP-dependent dethiobiotin synthetase BioD (TraesCS1B01G390200, TraesCS1B01G390300). The other three genes code for B3 domain-containing protein (TraesCS1B01G390400), Rotundifolia-like protein (TraesCS1B01G390600), and Hexosyltransferase (Traes CS1B01G390000).

Genomic Prediction

GP method was applied based on all SNPs, and the prediction of the genomic breeding value for each line was evaluated using 500 randomly generated train/test sets. The average correlation between observed tolerance to STB and predicted STB by GP was 0.47 in a model with no significant markers included as fixed effects. The GWAS results were used to select markers to fit as fixed effects. Significant markers were pooled from the GLM, MLM, MLMM, FarmCPU, and Super models. The six significant SNP markers identified in proximity to each other on chromosome 1B were reduced to the marker BobWhite_c42716_71 on the basis of the lowest FDR-adjusted p-value. In total, five significant markers were used as candidates for modeling with fixed effects (**Table 1**).

All possible combinations of the five GWAS-selected SNP markers were evaluated, in subset sizes from one marker to all five used as fixed effects (**Table 2**). The highest average prediction accuracy (0.62) was obtained from a model that included all five markers as fixed effects. Among the models with reduced number of markers (1–3 markers) set as fixed effects, the models using three GWAS-selected markers performed better compared to the models using one or two markers. The prediction accuracy thus increased on average from 0.48 for one marker added as fixed effect to 0.54 for three markers. Out of the three marker models, the best performing model was a model that included the following three markers BobWhite_c1361_1187, BobWhite_c42716_71, and Excalibur_c17553_84 with a prediction accuracy of 0.59 (**Supplementary Table 2**). In comparison, the model that did not use GWAS-selected markers as fixed effects, and the models that used randomly selected markers (regardless of GWAS significance), performed on average worse than both the GWAS-assisted models and the model with all markers set as random effects (**Table 2**).

Haplotype Analysis

Haplotype analysis was performed to identify haplotype variants for the QTL identified on chromosome 1B with six significant

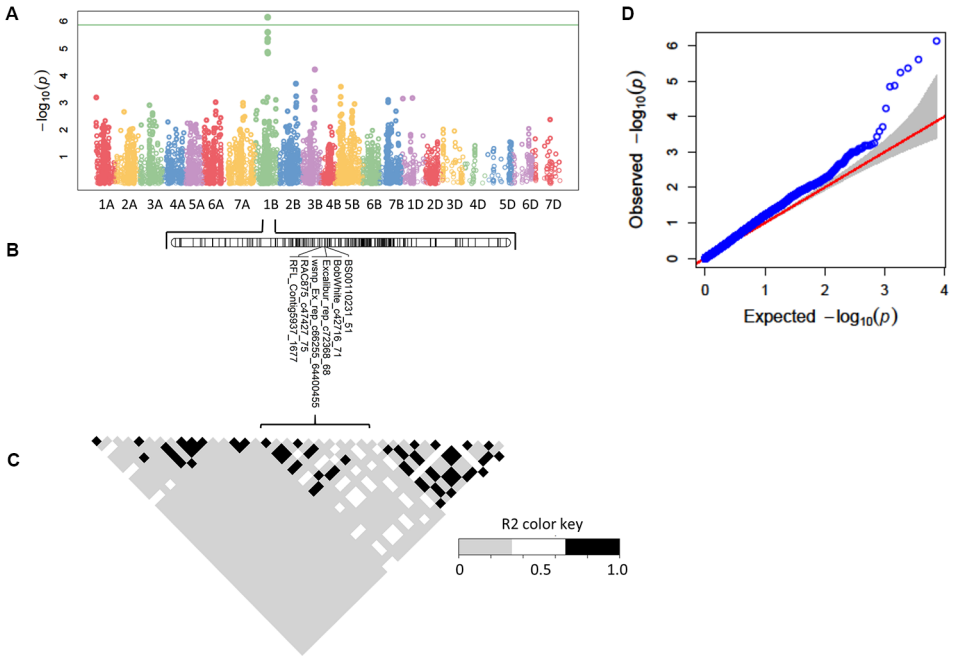


FIGURE 4 | Single nucleotide polymorphism (SNP) significantly associated with STB resistance in winter wheat identified by genome-wide association study (GWAS) with GLM model. **(A)** Manhattan plot; **(B)** Linkage map of Chromosome 1B; **(C)** Linkage disequilibrium plot; **(D)** Quantile-quantile plot.

TABLE 1 | Summary of the significant SNPs marker identified with different models which are associated with Septoria tritici blotch (STB) resistance in GWAS analysis with 175 winter wheat genotypes.

SNP marker name	Chr	Model	Position (cM)	MAF	Alleles	R^2	Allelic effecvt	Physical location
BobWhite_c1361_1187	1A	FarmCPU**** Super****	13.73	0.14	A/G	–	0.16	1525253
BobWhite_c42716_71	1B	FarmCPU**** GLM*** MLM* MLMM*** Super****	97.71	0.46	A/G	0.11	0.02	623712765
wsnp_Ex_rep_c66255_64400455	1B	GLM**	97.71	0.47	A/G	0.09	−0.01	623729791
RFL_Contig5937_1677	1B	GLM**	99.07	0.45	A/G	0.08	−0.01	623730512
RAC875_c47427_75	1B	GLM*** MLM*	99.07	0.47	A/G	0.10	−0.01	623731255
Excalibur_rep_c72368_68	1B	GLM*** MLM*	97.71	0.46	T/C	0.09	−0.003	623770763
BS00110231_51	1B	GLM**	97.36	0.43	T/G	0.09	0.01	623989423
wsnp_Ex_c22423_31615798	2B	FarmCPU*** Super***	96.99	0.37	A/C	–	0.08	215593752
wsnp_Ex_c5929_10402147	3A	FarmCPU**** Super****	86.16	0.31	T/C	–	−0.09	481018206
Excalibur_c17553_84	5A	FarmCPU*** Super***	43.27	0.35	C/T	–	0.09	375375809

*Chr, chromosome; MAF, minor allele frequency, physical location – start positions (in bp) of the markers on the chromosomes in the assembly IWGSC Refseq v1. FDR-adjusted p value *0.05, **0.01, ***0.001, ****0.0001. The percentage of variation (R^2) explained by the GLM model was calculated as the difference between the R^2 of the GAPIT model with and without the associated SNP. Allelic effect estimates the additive contribution of the tested marker and were obtained primarily from the GLM model when available else from FarmCPU model.*

markers. Haplotype variants were detected with the software DNAsp (Rozas et al., 2017). In total, 19 haplotype variants were detected with number of genotypes ranging from 1 to 71 in each variant. Of these, 3 haplotype variants were selected with at least five or more genotypic counts/genotypes (**Supplementary Table 1**). Thereafter, haplotype network was constructed with the TCS algorithm in the software PopART (Leigh et al., 2015) (**Figure 6**). The analysis revealed that Hap_2 had the lowest mean disease score of 0.77 compared to Hap_1 (0.96) and Hap_3 (0.95). Hap_2 had 11 genotypes of which 8 originated from Denmark, 2 from Sweden, and 1 from Germany. Most of the genotypes from Denmark had high resistance while one of the two genotypes from Sweden had high resistance.

DISCUSSION

STB is one of the most important winter wheat diseases in Northern Europe, and cultivars with higher levels of resistance which is stable and effective across environments are needed. Whereas individual Stb genes are not currently effective against *Z. tritici* populations in Europe (Arraiano et al., 2009), the identification of

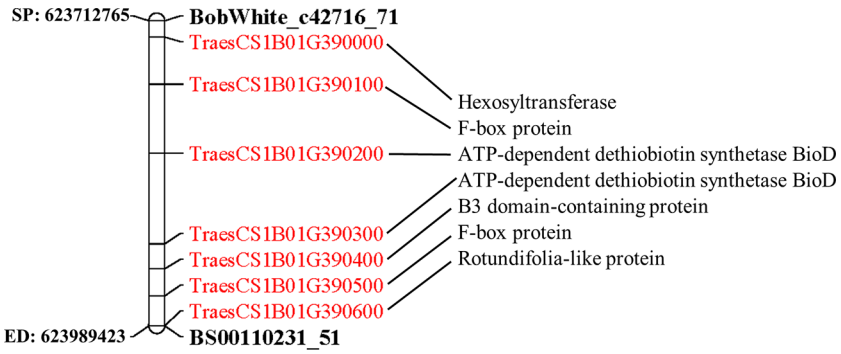


FIGURE 5 | Wheat chromosome 1B representing the physical position (in bp) of the flanking markers and genes localized within these markers. SP, start position (BobWhite_c42716_71); ED end position (BS00110231).

TABLE 2 | Summary of rrBLUP-based GWAS-assisted genomic prediction models of STB resistance scored in 175 winter wheat genotypes.

Number of markers set as fixed effects	Type of marker selection for fixed effects			
	Markers selected by significance in GWAS		Completely random selection of markers	
	Average model accuracy	95% confidence interval of the mean	Average model accuracy	95% confidence interval of the mean
0	0.47	N/A	N/A	N/A
1	0.48	[0.44, 0.51]	0.44	[0.43, 0.44]
2	0.51	[0.49, 0.53]	0.44	[0.43, 0.45]
3	0.54	[0.52, 0.56]	0.45	[0.42, 0.48]
4	0.58	[0.55, 0.61]	0.43	[0.41, 0.45]
5	0.62	N/A	0.44	[0.41, 0.47]

The models utilized permutations of 1 to 5 markers in significant association with STB resistance identified in the same population. The models were compared against a model containing no fixed effects and a series of models that sampled equally sized subsets of random markers, where each subset of random markers was repeated five times. All models were validated against the same set of 80/20 training/test sets (N = 500). The zero and five GWAS-selected marker models were only repeated once, and thus have no confidence interval data.

new QTL for STB resistance and incorporation of resistance into elite winter wheat cultivars is crucial. To this end, the current study analyzed 175 winter wheat genotypes of Nordic origin for STB resistance under controlled conditions at the seedling stage. Our results revealed that the NordGen genebank has a highly valuable and genetically diverse collection of germplasm comprising resistance to STB. This germplasm mainly originates from Sweden, Denmark, Finland, and Norway composing 56.2%, 25.5%, 9.6%, and 3.4% of all analyzed germplasm, respectively, and released approximately between 1900 and 2000. Population structure analysis revealed three clusters associated with geographical origin. Finish and Norwegian genotypes formed one cluster, the second cluster contains mainly Swedish genotypes while genotypes from Denmark and Germany segregated into the third cluster (**Figure 2A**). In addition, the result from the PCA data showed that the modern wheat cultivars exhibited a higher level of resistance in comparison to older released cultivars (**Figures 2B C**). This result indicated that the breeding progress for STB resistance over the last decades probably occurred by the gradual accumulation of genes with a minor effect, as is the case also in the American germplasm (Jlibene et al., 1994; Camacho-Casas et al., 1995). Similarly, the

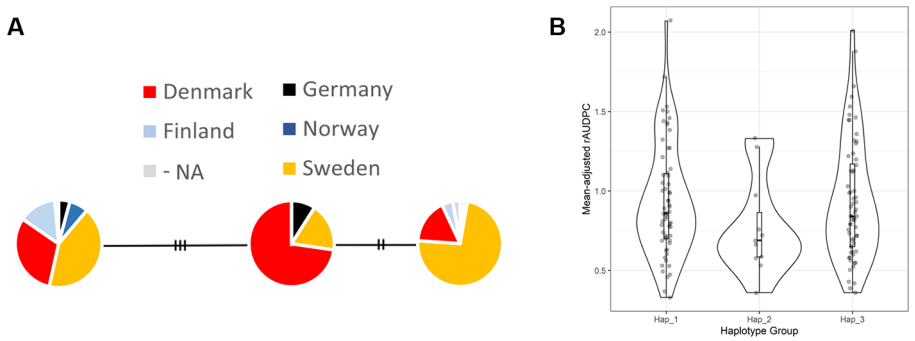


FIGURE 6 | Haplotype variants identified from the QTL on chromosome 1B. **(A)** Haplotype network with nodes denoted as pie charts and **(B)** range of distribution of STB resistance of genotypes in each variant.

characterization of old Tunisian durum wheat accessions for resistance to STB identified resistant germplasm and four new resistant genes (Ferjaoui et al., 2015). The authors, therefore, suggested that the old Tunisian durum wheat accessions harbor novel resistance genes that can be introgressed into the modern cultivars. The results from our work highlight the potential of old germplasm as novel sources of resistance to STB for winter wheat breeding programs in Northern Europe.

A QTL associated with STB resistance identified by GWAS in this study was mapped on chromosome 1B. Previous studies have mapped *Stb11* on the short arm of chromosome 1B in *TE9111* (Chartrain et al., 2005) and remapped *Stb2* was also located close to or at Stb11 locus in Vernopolis (Liu et al., 2013). *StbWW* identified in three DH populations, was also mapped on chromosome 1BS at or near *Stb11*. Raman et al. (2009) identified eight SNPs associated with STB resistance and one was mapped on chromosome 1B in European winter wheat collection. Goudemand et al. (2013) mapped two QTL on 1B (one 1BS and one 1BL) chromosome in bi-parental crosses. Recently, Vagndorf et al. (2017) identified QTL QStb.*NS-1B* located on the long arm of chromosome 1B by GWAS of Danish cultivars and breeding lines that were characterized over three years in three locations in Denmark for STB. In this study, one QTL was mapped on the long arm of chromosome 1B which is in close physical proximity to the QTL QStb.*NS-1B*. Thus, it can be postulated that it is the same QTL as identified previously. However, our study identified this QTL for quantitative resistance at the seedling stage under controlled conditions while the study by Vagndorf et al. (2017) identified the same QTL in field trials for adult plant resistance.

The other QTL associated with STB resistance identified in this study were located on chromosomes 1A, 2B, 3A, and 5A. QTL 1A, 2B, and 5A were mapped on the short arm of the respective chromosomes and QTL on 3A was mapped on the long arm. Goudemand et al. (2013) mapped two Meta-QTL (MQTL1 and MQTL6) on chromosomes 1A and 2B and another QTL (QTL8) on chromosome 5A for STB resistance which were in close physical proximity to the QTL mapped (1A, 2B, and 5A) in this study. The MQTL1 was associated with STB resistance both in adult and seedling stages whereas QTL8 was only associated with adult and MQTL6 was only associated with seedling stage resistance.

The QTL on chromosome 3A in our study was found in close physical proximity to the previously reported QTL (*QStb. risø-3A.2)* which was associated with STB resistance both in adult and seedling stages (Brown et al., 2015). Thus, our study further confirms the role of the identified QTL at the seedling stage. Introgression of these QTL in winter wheat cultivars will provide both seedling and adult plant stage resistance to STB.

In the present work, we identified seven candidate genes with putative roles in resistance to STB in wheat (**Figure 5**). Two of the identified genes (*TraesCS1B01G390100* and *TraesCS1B01G390500*) were associated with F-box proteins which plays a key role in plant immune responses through the involvement in hormone pathways (Yu et al., 2007). Two F-box proteins, *COI1* (Xie et al., 1998) and *SON1* (Kim and Delaney, 2002), have been demonstrated to have a role in plant defense in Arabidopsis plants. In our previous work, we identified candidate genes associated with STB resistance by integrating QTL mapping and transcriptome profiling, wherein, the F- box proteins were among the most represented in all identified QTL regions (Odilbekov et al., 2019). The other two genes identified in this work were related to ATP-dependent dethiobiotin synthetase BioD (*TraesCS1B01G390200* and *TraesCS1B01G390300*). ATP-dependent dethiobiotin synthetase BioD is involved in the first step of the sub-pathway that synthesizes biotin from 7,8-diaminononanoate. Li et al. (2012) demonstrated that biotin deficiency results in light-dependent spontaneous cell death and modulates defense gene expression in Arabidopsis plants. The other putative genes identified in the present work were B3 domain-containing protein (*TraesCS1B01G390400*). The B3 domain has been found in several transcription factors specific to higher plant species (Waltner et al., 2005). Wang et al. (2015) found that the B3 domain of BPH29 gene was associated with insect brown planthopper resistance in rice. Also, they have shown that during the infestation, the *RBPH54* triggers the salicylic acid signaling pathway and suppresses the jasmonic acid pathway, which is similar to biotrophic pathogens.

In the previous studies, prediction accuracy of GS models was found to be improved for example by increasing the training population size, testing the models on test populations genetically closely related to the training population, implementing a different GS algorithm, increasing the marker density or combining significantly associated markers as fixed effects (Solberg et al., 2008; Norman et al., 2018). In this work, we evaluated the prediction accuracy of GP models when GWAS markers were included as fixed effects. When GWAS markers obtained from different GWAS models were included as fixed effects, the accuracy of GP was significantly improved (**Table 2**). The results also suggest that including two or more GWAS markers as fixed effects significantly increases the accuracy of the GP models. Our results corroborate the trends in accuracy improvements seen in the previous studies integrating GWAS and GP in winter wheat (Herter et al., 2019) maize (Bian and Holland, 2017), and rice (Spindel et al., 2016).

Finally, this and the previous studies (Daetwyler et al., 2014; Crossa et al., 2016) have shown that GP can be used to obtain GEBVs for economically important traits in landraces by training models on a subset of landraces that are phenotyped. There are several hundred thousand landraces stored in genebanks worldwide, and thus, advanced methods, such as GP will enable high-throughput evaluation of landraces to identify those with superior resistance traits. The identified landraces can then be included in the wheat breeding programs to perform GP-based progeny selection.

CONCLUSIONS

This study leads to the conclusion that the wheat genotypes stored at NordGen are a genetically diverse resource. The highly resistant genotypes serve as potential donors for improving commercial cultivars in the Nordic and Baltic Sea Region countries. The significant SNP markers can be used for marker-assisted selection of STB resistance at the seedling stage in wheat breeding. The genes identified by GWAS approach can serve as candidate genes for improving STB resistance in wheat through functional studies. In addition, the results indicate that integrating GWAS with GP could

facilitate further improvement of GP accuracy thereby improving the selection efficiency of the breeding program.

DATA AVAILABILITY STATEMENT

The genotypic data can be assessed from the following link https://doi.org/10.6084/m9.figshare.10184468.

AUTHOR CONTRIBUTIONS

AC conceived and planned the study and performed GWAS, haplotype and GP analysis. FO and RA performed germplasm characterization. AK performed GWAS-assisted GP analysis. JS selected genotypes and performed genotyping. FO performed statistical analysis and wrote the first draft. All authors contributed in the interpretation of the data and in writing the manuscript.

FUNDING

This project was funded by Lantmännen Research Foundation (2016F010), Einar Nilssons Stiftelse, and SLU Grogrund.

ACKNOWLEDGMENTS

We would like to thank Ganapathi Varma Saripella for the bioinformatics support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01224/full#supplementary-material

REFERENCES

Arraiano, L., and Brown, J. (2006). Identification of isolate-specific and partial resistance to septoria tritici blotch in 238 European wheat cultivars and breeding lines. *Plant Pathol.* 55 (6), 726–738. doi: 10.1111/j.1365-3059.2006.01444.x

Arraiano, L. S., and Brown, J. K. M. (2017). Sources of resistance and susceptibility to Septoria tritici blotch of wheat. *Mol. Plant Pathol.* 18 (2), 276–292. doi: 10.1111/mpp.12482

Arraiano, L. S., Balaam, N., Fenwick, P. M., Chapman, C., Feuerhelm, D., Howell, P., et al. (2009). Contributions of disease resistance and escape to the control of septoria tritici blotch of wheat. *Plant Pathol.* 58 (5), 910–922. doi: 10.1111/j.1365-3059.2009.02118.x

Bian, Y., and Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity* 118 (6), 585–593. doi: 10.1038/hdy.2017.4

Brown, J. K., Chartrain, L., Lasserre-Zuber, P., and Saintenac, C. (2015). Genetics of resistance to Zymoseptoria tritici and applications to wheat breeding. *Fungal Genet. Biol.* 79, 33–41. doi: 10.1016/j.fgb.2015.04.017

Camacho-Casas, M., Kronstad, W., and Scharen, A. (1995). Septoria tritici resistance and associations with agronomic traits in a wheat cross. *Crop Sci.* 35 (4), 971–976. doi: 10.2135/cropsci1995.0011183X003500040006x

Chartrain, L., Brading, P., and Brown, J. (2005). Presence of the Stb6 gene for resistance to Septoria tritici blotch (Mycosphaerella graminicola) in cultivars used in wheat-breeding programmes worldwide. *Plant Pathol.* 54 (2), 134–143. doi: 10.1111/j.1365-3059.2005.01164.x

Chawade, A., Armoniene, R., Berg, G., Brazauskas, G., Frostgard, G., Geleta, M., et al. (2018). A transnational and holistic breeding approach is needed for sustainable wheat production in the Baltic Sea region. *Physiol. Plant*. 164 (4), 442–451. doi: 10.1111/ppl.12726

Cheval, P., Siah, A., Bomble, M., Popper, A. D., Reignault, P., and Halama, P. (2017). Evolution of QoI resistance of the wheat pathogen Zymoseptoria tritici in Northern France. *Crop Prot.* 92, 131–133. doi: 10.1016/j.cropro.2016.10.017

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic Prediction of Gene Bank Wheat Landraces. *G3. Genes|Genomes|Genetics* 6 (7), 1819–1834. doi: 10.1534/g3.116.029637

Daetwyler, H. D., Bansal, U. K., Bariana, H. S., Hayden, M. J., and Hayes, B. J. (2014). Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* 127 (8), 1795–1803. doi: 10.1007/s00122-014-2341-8

de Carvalho, M. A. A. P., Bebeli, P. J., Bettencourt, E., Costa, G., Dias, S., Dos Santos, T. M. M., et al. (2012). Cereal landraces genetic resources in worldwide GeneBanks. A review. *Agron. Sustain. Dev.* 33 (1), 177–203. doi: 10.1007/s13593-012-0090-0

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends In Plant Sci.* 19 (9), 592–601. doi: 10.1016/j.tplants.2014.05.006

Diederichsen, A., Solberg, S.Ø., and Jeppson, S. (2012). Morphological changes in Nordic spring wheat (*Triticum aestivum* L.) landraces and cultivars released from 1892 to 1994. *Genet. Resour. Crop Evol.* 60 (2), 569–585. doi: 10.1007/s10722-012-9858-y

Dreisigacker, S., Zhang, P., Warburton, M. L., Skovmand, B., Hoisington, D., and Melchinger, A. E. (2005). Genetic Diversity among and within CIMMYT Wheat Landrace Accessions Investigated with SSRs and Implications for Plant Genetic Resources Management. *Crop Sci.* 45 (2), 653–661. doi: 10.2135/cropsci2005.0653

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* 4 (3), 250–255. doi: 10.3835/plantgenome2011.08.0024

Ferjaoui, S., M'Barek, S. B., Bahri, B., Slimane, R. B., and Hamza, S. (2015). Identification of resistance sources to Septoria Tritici blotch in old Tunisian durum wheat germplasm applied for the analysis of the Zymoseptoria triticidurum wheat interaction. *J. Plant Pathol.* 97 (3), 1–11. doi: 10.4454/JPP.V97I3.028

Fones, H., and Gurr, S. (2015). The impact of Septoria tritici Blotch disease on wheat: An EU perspective. *Fungal Genet. Biol.* 79, 3–7. doi: 10.1016/j.fgb.2015.04.004

Goudemand, E., Laurent, V., Duchalais, L., Ghaffary, S. M. T., Kema, G. H., Lonnet, P., et al. (2013). Association mapping and meta-analysis: two complementary approaches for the detection of reliable Septoria tritici blotch quantitative resistance in bread wheat (Triticum aestivum L.). *Mol. Breed.* 32 (3), 563–584. doi: 10.1007/s11032-013-9890-4

Hagenblad, J., Asplund, L., Balfourier, F., Ravel, C., and Leino, M. W. (2012). Strong presence of the high grain protein content allele of NAM-B1 in Fennoscandian wheat. *Theor. Appl. Genet.* 125 (8), 1677–1686. doi: 10.1007/s00122-012-1943-2

Herter, C. P., Ebmeyer, E., Kollers, S., Korzun, V., Würschum, T., and Miedaner, T. (2019). Accuracy of within- and among-family genomic prediction for Fusarium head blight and Septoria tritici blotch in winter wheat. *Theor. Appl. Genet.* 132 (4), 1121–1135. doi: 10.1007/s00122-018-3264-6

Hysing, S. C., Merker, A., Liljeroth, E., Koebner, R. M., Zeller, F. J., and Hsam, S. L. (2007). Powdery mildew resistance in 155 Nordic bread wheat cultivars and landraces. *Hereditas* 144 (3), 102–119. doi: 10.1111/j.2007.0018-0661.01991.x

Jlibene, M., Gustafson, J., and Rajaram, S. (1994). Inheritance of resistance to Mycosphaerella graminicola in hexaploid wheat. *Plant Breed.* 112 (4), 301–310. doi: 10.1111/j.1439-0523.1994.tb00688.x

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Rutkoski, J. E., Poland, J. A., et al. (2017). Comparison of Models and Whole-Genome Profiling Approaches for Genomic-Enabled Prediction of Septoria Tritici Blotch, Stagonospora Nodorum Blotch, and Tan Spot Resistance in Wheat. *Plant Genome* 10 (2). doi: 10.3835/plantgenome2016.08.0082

Kim, H. S., and Delaney, T. P. (2002). Arabidopsis SON1 is an F-box protein that regulates a novel induced defense response independent of both salicylic acid

and systemic acquired resistance. *Plant Cell* 14 (7), 1469–1482. doi: 10.1105/tpc.001867

Kollers, S., Rodemann, B., Ling, J., Korzun, V., Ebmeyer, E., Argillier, O., et al. (2013). Genetic architecture of resistance to Septoria tritici blotch (Mycosphaerella graminicola) in European winter wheat. *Mol. Breed.* 32 (2), 411–423. doi: 10.1007/s11032-013-9880-6

Leigh, J. W., Bryant, D., and Nakagawa, S. (2015). popart: full-feature software for haplotype network construction. *Methods In Ecol. Evol.* 6 (9), 1110–1116. doi: 10.1111/2041-210x.12410

Li, J., Brader, G., Helenius, E., Kariola, T., and Palva, E. T. (2012). Biotin deficiency causes spontaneous cell death and activation of defense signaling. *Plant J.* 70 (2), 315–326. doi: 10.1111/j.1365-313X.2011.04871.x

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, Y., Zhang, L., Thompson, I. A., Goodwin, S. B., and Ohm, H. W. (2013). Molecular mapping re-locates the Stb2 gene for resistance to Septoria tritici blotch derived from cultivar Veranopolis on wheat chromosome 1BS. *Euphytica* 190 (1), 145–156. doi: 10.1007/s10681-012-0796-8

Lopes, M. S., El-Basyoni, I., Baenziger, P. S., Singh, S., Royo, C., Ozbek, K., et al. (2015). Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J. Exp. Bot.* 66 (12), 3477–3486. doi: 10.1093/jxb/erv122

McDonald, B. A., and Mundt, C. C. (2016). How Knowledge of Pathogen Population Biology Informs Management of Septoria Tritici Blotch. *Phytopathology* 106 (9), 948–955. doi: 10.1094/phyto-03-16-0131-rvw

Mendiburu, F. D. (2017). agricolae: Statistical Procedures for Agricultural Research. R package version 1.2-8. Available: https://CRAN.R-project.org/package=agricolae.

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

Miedaner, T., Zhao, Y., Gowda, M., Longin, C. F. H., Korzun, V., Ebmeyer, E., et al. (2013). Genetic architecture of resistance to Septoria tritici blotch in European wheat. *BMC Genomics* 14 (1), 858. doi: 10.1186/1471-2164-14-858

Muqaddasi, Q. H., Zhao, Y., Rodemann, B., Plieske, J., Ganal, M. W., and Röder, M. S. (2019). Genome-wide Association Mapping and Prediction of Adult Stage Blotch Infection in European Winter Wheat *via* High-Density Marker Arrays. *Plant Genome* 12 (1). doi: 10.3835/plantgenome2018.05.0029

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3:. Genes|Genomes|Genetics* 8 (9), 2889–2899. doi: 10.1534/g3.118.200311

Odilbekov, F., Armoniené, R., Henriksson, T., and Chawade, A. C. (2018). Proximal phenotyping and machine learning methods to identify Septoria tritici blotch disease symptoms in wheat. *Front. In Plant Sci.* 9, 685. doi: 10.3389/fpls.2018.00685

Odilbekov, F., He, X., Armoniené, R., Saripella, G. V., Henriksson, T., Singh, P. K., et al. (2019). QTL Mapping and Transcriptome Analysis to Identify Differentially Expressed Genes Induced by Septoria Tritici Blotch Disease of Wheat. *Agronomy* 9 (9), 510. doi: 10.3390/agronomy9090510

Raman, R., Milgate, A. W., Imtiaz, M., Tan, M. K., Raman, H., Lisle, C., et al. (2009). Molecular mapping and physical location of major gene conferring seedling resistance to Septoria tritici blotch in wheat. *Mol. Breed.* 24 (2), 153–164. doi: 10.1007/s11032-009-9280-0

Randhawa, M., Bansal, U., Lillemo, M., Miah, H., and Bariana, H. (2016). Postulation of rust resistance genes in Nordic spring wheat genotypes and identification of widely effective sources of resistance against the Australian rust flora. *J. Appl. Genet.* 57 (4), 453–465. doi: 10.1007/s13353-016-0345-6

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34 (12), 3299–3302. doi: 10.1093/molbev/msx248

Saintenac, C., Lee, W. S., Cambon, F., Rudd, J. J., King, R. C., Marande, W., et al. (2018). Wheat receptor-kinase-like protein Stb6 controls gene-for-gene resistance to fungal pathogen Zymoseptoria tritici. *Nat. Genet.* 50 (3), 368–374. doi: 10.1038/s41588-018-0051-x

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86 (10), 2447–2454. doi: 10.2527/jas.2007-0010

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., et al. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116 (4), 395–408. doi: 10.1038/hdy.2015.113

Tabib Ghaffary, S. M., Faris, J. D., Friesen, T. L., Visser, R. G. F., van der Lee, T. A. J., Robert, O., et al. (2011). New broad-spectrum resistance to septoria tritici blotch derived from synthetic hexaploid wheat. *Theor. Appl. Genet.* 124 (1), 125–142. doi: 10.1007/s00122-011-1692-7

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9 (2). doi: 10.3835/plantgenome2015.11.0120

Torriani, S. F. F., Brunner, P. C., McDonald, B. A., and Sierotzki, H. (2009). QoI resistance emerged independently at least 4 times in European populations of Mycosphaerella graminicola. *Pest Manage. Sci.* 65 (2), 155–162. doi: 10.1002/ps.1662

Torriani, S. F., Melichar, J. P., Mills, C., Pain, N., Sierotzki, H., and Courbot, M. (2015). Zymoseptoria tritici: A major threat to wheat production, integrated approaches to control. *Fungal Genet. Biol.* 79, 8–12. doi: 10.1016/j.fgb.2015.04.010

Vagndorf, N., Nielsen, N. H., Edriss, V., Andersen, J. R., Orabi, J., Jørgensen, L. N., et al. (2017). Genomewide association study reveals novel quantitative trait loci associated with resistance towards Septoria tritici blotch in North European winter wheat. *Plant Breed.* 136 (4), 474–482. doi: 10.1111/pbr.12490

Waltner, J. K., Peterson, F. C., Lytle, B. L., and Volkman, B. F. (2005). Structure of the B3 domain from Arabidopsis thaliana protein At1g16640. *Protein Sci.* 14 (9), 2478–2483. doi: 10.1110/ps.051606305

Wang, Y., Cao, L., Zhang, Y., Cao, C., Liu, F., Huang, F., et al. (2015). Map-based cloning and characterization of BPH29, a B3 domain-containing recessive gene conferring brown planthopper resistance in rice. *J. Exp. Bot.* 66 (19), 6035–6045. doi: 10.1093/jxb/erv318

Wieczorek, T. M., Berg, G., Semaškienė, R., Mehl, A., Sierotzki, H., Stammler, G., et al. (2015). Impact of DMI and SDHI fungicides on disease control and CYP51 mutations in populations of Zymoseptoria tritici from Northern Europe. *Eur. J. Plant Pathol.* 143 (4), 861–871. doi: 10.1007/s10658-015-0737-1

Xie, D.-X., Feys, B. F., James, S., Nieto-Rostro, M., and Turner, J. G. (1998). COI1: an Arabidopsis gene required for jasmonate-regulated defense and fertility. *Science* 280 (5366), 1091–1094. doi: 10.1126/science.280.5366.1091

Yu, H., Wu, J., Xu, N., and Peng, M. (2007). Roles of F-box proteins in plant hormone responses. *Acta Biochim. Biophys. Sin.* 39 (12), 915–922. doi: 10.1111/j.1745-7270.2007.00358.x

Zhong, Z., Marcel, T. C., Hartmann, F. E., Ma, X., Plissonneau, C., Zala, M., et al. (2017). A small secreted protein in Zymoseptoria tritici is responsible for avirulence on wheat cultivars carrying the Stb6 resistance gene. *New Phytol.* 214 (2), 619–631. doi: 10.1111/nph.14434

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Odilbekov, Armonienė, Koc, Svensson and Chawade. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 December 2019
doi: 10.3389/fgene.2019.01168

Deep Kernel and Deep Learning for Genome-Based Prediction of Single Traits in Multienvironment Breeding Trials

José Crossa[1,3], Johannes W.R. Martini[1], Daniel Gianola[2], Paulino Pérez-Rodríguez[3], Diego Jarquin[4], Philomin Juliana[1], Osval Montesinos-López[5] and Jaime Cuevas[6]**

[1] Biometrics and Statistics Unit, Genetic Resources Program, and Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [2] Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, United States, [3] Programa de Postgrado de Socioeconomia, Estadistica e Informatica, Colegio de Postgraduados, Texcoco, Mexico, [4] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, United States, [5] Facultad de Telemática, Universidad de Colima, Colima, Mexico, [6] Departamento de Ciencias, Universidad de Quintana Roo, Chetumal, Mexico

OPEN ACCESS

Edited by:
Nunzio D'Agostino,
University of Naples Federico II, Italy

Reviewed by:
Chenwu Xu,
Yangzhou University, China
Yang Hu,
Washington State University, United States
Miguel Perez-Enciso,
Autonomous University of Barcelona, Spain

***Correspondence:**
Osval Montesinos-López
oamontes2@hotmail.com
Jaime Cuevas
jaicueva@uqroo.edu.mx

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 16 September 2019
Accepted: 23 October 2019
Published: 09 December 2019

Citation:
Crossa J, Martini JWR, Gianola D, Pérez-Rodríguez P, Jarquin D, Juliana P, Montesinos-López O and Cuevas J (2019) Deep Kernel and Deep Learning for Genome-Based Prediction of Single Traits in Multienvironment Breeding Trials. Front. Genet. 10:1168. doi: 10.3389/fgene.2019.01168

Deep learning (DL) is a promising method for genomic-enabled prediction. However, the implementation of DL is difficult because many hyperparameters (number of hidden layers, number of neurons, learning rate, number of epochs, batch size, etc.) need to be tuned. For this reason, deep kernel methods, which only require defining the number of layers, may be an attractive alternative. Deep kernel methods emulate DL models with a large number of neurons, but are defined by relatively easily computed covariance matrices. In this research, we compared the genome-based prediction of DL to a deep kernel (arc-cosine kernel, AK), to the commonly used non-additive Gaussian kernel (GK), as well as to the conventional additive genomic best linear unbiased predictor (GBLUP/GB). We used two real wheat data sets for benchmarking these methods. On average, AK and GK outperformed DL and GB. The gain in terms of prediction performance of AK and GK over DL and GB was not large, but AK and GK have the advantage that only one parameter, the number of layers (AK) or the bandwidth parameter (GK), has to be tuned in each method. Furthermore, although AK and GK had similar performance, deep kernel AK is easier to implement than GK, since the parameter "number of layers" is more easily determined than the bandwidth parameter of GK. Comparing AK and DL for the data set of year 2015–2016, the difference in performance of the two methods was bigger, with AK predicting much better than DL. On this data, the optimization of the hyperparameters for DL was difficult and the finally used parameters may have been suboptimal. Our results suggest that AK is a good alternative to DL with the advantage that practically no tuning process is required.

Keywords: deep learning, deep kernel, genomic selection, kernel methods, artificial neural networks, genomic × environment interaction

INTRODUCTION

Using dense molecular markers, Meuwissen et al. (2001) were the first to propose genome-enabled prediction for implementing genomic-assisted breeding. Subsequently, an enormous number of research articles published in animal and plant breeding journals explored and studied genomic selection (GS) and genome-based prediction (GP) outcomes in a large variety of animal and plant species for different traits and measured in different environments (Crossa et al., 2017). GS combines molecular and phenotypic data in a training population to predict genomic breeding values (or genetic values) of individuals that have been genotyped but not phenotyped. The predictions can be used in a breeding program to reduce cycle length or to increase the selection precision, thus enhancing the response to selection.

GS and prediction approaches have focused on two different cases. One is predicting additive effects in early generations of a breeding program to achieve rapid selection with a short interval cycle (Beyene et al., 2015; Zhang et al., 2017). Here, breeders focus on GP of breeding values (additive values) of an infinitesimal model that assumes a linear function of allelic effects for an infinite number of loci; therefore, additive linear models that summarize the effects of the markers are sufficient. The most commonly used additive method is genomic best linear unbiased predictor (GBLUP/GB) (Van Raden, 2007). The other case considers the complete genetic values of individuals including both additive and nonadditive (dominance and epistasis) effects, thereby estimating the genetic performance of the cultivars (Crossa et al., 2017).

As pointed out by Harfouche et al. (2019), despite the fact that GS programs have provided extensive amounts of new data in crops, legumes, and tree species, the lack of predictive accuracy for many complex traits is underpinned by the complexity of modeling all of the important factors inherent to targets such as grain yield. Harfouche et al. (2019) mentioned that linking phenotypes with genotypes using high-throughput phenomics and genomics will continue to be the main challenge for plant breeding in the next decades.

The complexity of applying GS and GP in breeding is influenced by various factors acting at different levels. An important difficulty arises when predicting unobserved individuals in specific environments (site-year combinations) by incorporating genotype (genomic) × environment (G×E) interaction into statistical models. An additional layer of complexity is the G×E interactions for multitraits. Here statistical-genetic models exploit multitrait, multienvironment variance-covariance structures and correlations between traits and environments simultaneously. Understanding the complexity of traits requires a theoretical framework that accounts for often cryptic interactions.

Some of the statistical complexities can be addressed by using semiparametric genomic regression methods to account for nonadditive variation (Gianola et al., 2006; Gianola et al., 2011; Morota and Gianola, 2014; Morota et al., 2014). These methods have been used to predict complex traits in wheat with promising practical results (González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012). Semiparametric models often use kernel methods (a kernel utilizes functions that represent the inner product of many basic functions) for addressing complex gene actions (e.g., gene×gene epistatic interactions), thus capturing nonlinear relations between phenotype and genotype. Kernel-based methods for genomic regression have been used extensively in animal and plant breeding due to their capacity to produce reasonably accurate predictions (Gianola et al., 2014).

A commonly used kernel is the Gaussian kernel (GK) defined as $\exp(-hd_{ii'}^2/q)$, where h is a bandwidth parameter which controls the rate of decay of the covariance between genotypes, and q is the median of the square of the Euclidean distance, $d_{ii'}^2 = \sum_k (x_{ik} - x_{i'k})^2$ which is a measure of the genetic distance between individuals (i,i') based on molecular markers. The parameter q could also be included in the bandwidth parameter h, but standardizing the Euclidean distances by q makes it easier to apply a standardized grid search when looking for the optimal h. The GK appears as a reproducing kernel in the semiparametric reproducing kernel Hilbert spaces (RKHS) (Gianola and van Kaam, 2008; González-Camacho et al., 2012). Pérez-Elizalde et al. (2015) proposed an empirical Bayes method for estimating the bandwidth parameter h. An alternative approach to using a kernel with specific bandwidth parameters is the multikernel fitting proposed by de los Campos et al. (2010). Cuevas et al. (2016; 2017; 2018) and Souza et al. (2017) showed that using the GK within the multienvironment genomic G×E model of Jarquín et al. (2014) led to higher prediction accuracy than the same method with the linear kernel GB. Parametric alternatives for modeling epistasis have also been broadly discussed in literature (Jiang and Reif, 2015; Martini et al., 2016).

Deep learning (DL) methods are very flexible and have the potential to adapt to complex potentially cryptic data structures. In general, DL architectures are composed of three types of layers: (1) an input layer corresponding to the input information (predictors, that is, markers); (2) hidden layers, that is, the number of internal transformations performed on the original input information, which can be at least one but also a larger number; however, the number of neurons in each hidden layer needs to be tuned or specified; and (3) the output layer that produces the final predictions of the response variables we are interested in. Montesinos-López et al. (2018a; 2018b; 2019a; 2019b) recently performed extensive studies using DL methods for assessing GP for different types of traits (continuous, ordinal, and binary) accounting (or not) for G×E and comparing their prediction accuracies with those obtained by GB for single environments and multiple environments (with G×E). The authors used data from extensive maize and wheat multitrait, multienvironment trials. DL produced similar or slightly better prediction accuracies than GBLUP when G×E was not considered, but it was less accurate when G×E was included in the model. The authors hypothesized that DL may already account for G×E, so that its inclusion in the model was not required. Overall, the current drawback of applying DL for GP is the lack of a formal method for defining hyperparameters (e.g., number of neurons, number of layers, batch size) and, therefore, the time required for parameter tuning. Moreover, there may be an increased tendency towards overfitting the training data, and when important data features such as G×E interaction are known, direct modeling may lead to better predictions than modeling the structures implicitly in DL.

Recently, Cuevas et al. (2019) introduced the positive-definite arc-cosine kernel (AK) function for genome-enabled prediction.

The AK was initially proposed by Cho and Saul (2009) for exploring the option of DL in kernel machines. The nonlinear AK is defined by a covariance matrix that emulates a DL model with one hidden layer and a large number of neurons. Moreover, a recursive formula allows altering the covariance matrix stepwise, thus adding more hidden layers to the emulated deep neural network. The AK kernel method has been used in genomic single-environment models, as well as for genomic multienvironment models including genomic × environment interaction (G×E) (Cuevas et al., 2019). AK has the advantage over GK that it is computationally much simpler, since no bandwidth parameter is required, while performing similarly or slightly better than GK. The tuning parameter "number of layers" which is required for AK can be determined by a maximum marginal likelihood procedure (Cuevas et al., 2019).

Although AK has already been compared with GK (Cuevas et al., 2019), AK has not been formally compared with DL methods. Therefore, the main objective of this study was to compare the genome-based prediction accuracy of the GB, GK, AK, and DL methods using single-environment and multienvironment G×E models on two data sets from the CIMMYT Global Wheat Program. The data sets comprised two years (2015–2016 and 2016–2017) of Elite Yield Trial data, each consisting of 1052 and 1040 elite wheat lines, respectively. Lines of both Elite Yield Trials were evaluated in four environments using two irrigation levels [5 irrigations, 5IR, and 2 irrigations, 2IR] and two planting systems (flat, F, and bed, B) reflecting mega-environments defined by breeders in South Asia and Mexico.

MATERIAL AND METHODS

Genome-Based Prediction Models

The statistical methods used in this study have been described in several articles (Cuevas et al., 2016; Cuevas et al., 2017; Souza et al., 2017; Cuevas et al., 2018) for the single-environment model and the multienvironment G×E models using the GB and the GK. In addition, AK has recently been described in Cuevas et al. (2019). A brief description of the models (single-environment and G×E models) and methods (GB, GK, AK, and DL) is given below.

Single-Environment and Multiple-Environment G×E Models

For a single environment and only one kernel, the model can be expressed as:

$$\boldsymbol{y}=\mu\mathbf{1}+\boldsymbol{u}+\boldsymbol{\varepsilon} \tag{1}$$

where μ is the overall mean, $\mathbf{1}$ is the vector of ones, and $\boldsymbol{y}$ is the vector of observations of size n. Moreover, $\boldsymbol{u}$ is the vector of genomic effects $\boldsymbol{u}\sim\boldsymbol{N}(\mathbf{0},\sigma_u^2\boldsymbol{K})$, where σ_u^2 is the genomic variance estimated from the data, and matrix $\boldsymbol{K}$ is constructed as $\boldsymbol{K}=\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{Z}_g'$, with matrix $\boldsymbol{Z}_g$ a matrix of 0s and 1s with exactly one 1 in each row, and which relates the genotypes to the observations. The covariance matrix $\boldsymbol{G}$ models the genomic similarities between genotypes and varies between models: GB ($\boldsymbol{G}=\boldsymbol{X}\boldsymbol{X}'/p$) (where $\boldsymbol{X}$ is the scaled marker matrix and p denotes the number of markers); GK ($G_{ii'}=\exp(hd_{ii'}^2/q)$ where $d_{ii'}^2=\sum_k(x_{ik}-x_{i'k})^2$); and AK (see the description below). The random residuals are assumed independent with normal distribution $\boldsymbol{\varepsilon}\sim\boldsymbol{N}(\mathbf{0},\sigma_\varepsilon^2\boldsymbol{I})$, where σ_ε^2 is the error variance.

In the G×E multienvironment model of Jarquín et al. (2014), Lopez-Cruz et al. (2015), and Cuevas et al. (2016), Eq. (1) takes the form

$$\boldsymbol{y}=\mu\mathbf{1}+\boldsymbol{Z}_E\boldsymbol{\beta}_E+\boldsymbol{u}_1+\boldsymbol{u}_2+\boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{y}=[\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{nm}]'$ are the observations collected in each of the m sites (or environments) with n lines in each site. The fixed effects of the environment are modeled with the incidence matrix of environments $\boldsymbol{Z}_E$, where the parameters to be estimated are the intercept for each environment $\boldsymbol{\beta}_E$ (other fixed effects can be incorporated into the model). In this model, $\boldsymbol{u}_1\sim\boldsymbol{N}\left(\mathbf{0},\sigma_{u_1}^2\boldsymbol{K}_1\right)$ represents the genomic main effects, $\sigma_{u_1}^2$ is the genomic variance component estimated from the data, and $\boldsymbol{K}_1=\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{Z}_g'$, where $\boldsymbol{Z}_g$ relates the genotypes to the phenotypic observations. The random effect $\boldsymbol{u}_2$ represents the interaction between the genomic effects and their interaction with environments and is modeled as $\boldsymbol{u}_2\sim\boldsymbol{N}\left(\mathbf{0},\sigma_{u_2}^2\boldsymbol{K}_2\right)$, where $\boldsymbol{K}_2=\left(\boldsymbol{Z}_g\boldsymbol{G}\boldsymbol{Z}_g'\right)\circ(\boldsymbol{Z}_E\boldsymbol{Z}_E')$, where ° is the Hadamard product.

AK Method

DL architectures are generally difficult to tune. The tuning process involves, for instance, selecting the activation function, determining the number of hidden layers and the number of neurons in each hidden layer, and selecting the type of regularization. For this reason, Neal (1996) proposed a Bayesian method for deep artificial neural networks (ANN with more than one hidden layer), also called simple DL models, and, in conjunction with the results of Williams (1998) and Cho and Saul (2009), established the relationship between the AK and the deep neural networks with one hidden layer. These authors proposed a family of kernels that emulate DL models.

For AK, an important component is the angle between two vectors computed from marker genotypes $\boldsymbol{x}_i$ $\boldsymbol{x}_i$, as

$$\theta_{i,i'}=\cos^{-1}\left(\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_{i'}}{||\boldsymbol{x}_i||\,||\boldsymbol{x}_{i'}||}\right)$$

where $\cdot$ denotes the inner product and $||\boldsymbol{x}_i||$ is the norm of line i. The following kernel is positive semidefinite and related to an ANN with a single hidden layer and the ramp activation function (Cho and Saul, 2009)

$$\mathrm{AK}^1(\boldsymbol{x}_i,\boldsymbol{x}_{i'})=\frac{1}{\pi}||\boldsymbol{x}_i||\,||\boldsymbol{x}_{i'}||\,J(\theta_{i,i'}) \tag{3}$$

where π is the pi constant and $J(\theta_{i,i'})=[\sin(\theta_{i,i'})+(\pi-\theta_{i,i'})\cos(\theta_{i,i})]$. Equation (3) gives a symmetric positive semidefinite matrix (AK^1) preserving the norm of the entries such that $\mathrm{AK}(\boldsymbol{x}_i,\boldsymbol{x}_i)=||\boldsymbol{x}_i||^2$, and $\mathrm{AK}(\boldsymbol{x}_i,-\boldsymbol{x}_i)=0$ and models nonlinear relationships.

Note that the diagonal elements of the AK matrix are not identical and express heterogeneous variances of the genetic values u. This is different from the GK matrix, with a diagonal that expresses homogeneous variances. This property could represent

a theoretical advantage of AK when modeling interrelationships between individuals.

In order to emulate the performance of an ANN with more than one hidden layer (l), Cho and Saul (2009) proposed a recursive relationship of repeating l times the interior product

$$AK^{(l+1)}(\boldsymbol{x}_i,\boldsymbol{x}_{i'})=\frac{1}{\pi}[AK^{(l)}(\boldsymbol{x}_i,\boldsymbol{x}_i)AK^{(l)}(\boldsymbol{x}_{i'},\boldsymbol{x}_{i'})]^{\frac{1}{2}}\ J(\theta_{i,i'}^{(l)}) \quad (4)$$

where $\theta_{i,i'}^{(l)}=\cos^{-1}\{AK^{(l)}(\boldsymbol{x}_i,\boldsymbol{x}_{i'})[AK^{(l)}(\boldsymbol{x}_i,\boldsymbol{x}_i)AK^{(l)}(\boldsymbol{x}_{i'},\boldsymbol{x}_{i'})]^{-\frac{1}{2}}\}$. Thus, computing $AK^{(l+1)}$ at level (layer) l+1 is done from the previous layer $AK^{(l)}$. Computing a bandwidth is not necessary, and the only computational effort required is to compute the number of discrete layers. Cuevas et al. (2019) described a maximum marginal likelihood method used to select the number of hidden layers (l) for the AK kernel.

DL Neural Network

The DL for a single trait, including the multienvironment G×E situation employed in this study, follows the approach delineated by Montesinos-López et al. (2018a). In DL, the input to the model is a vector space that is subject to several complex geometric transformations that decompose into simple geometric transformations. The main objective of these geometric transformations is to map the input space to the target output space where the transformations are parameterized by the weight of the input at each neuron in each layer. A brief description of the process for tuning DL and for model selection is provided.

The implemented DL has a feedforward topology in which the information moves in only one direction (i.e., forward) from the input nodes (representing prediction variables), through the hidden nodes (located in hidden layers), and to the output nodes (representing target variables). There are no cycles or loops in this network. The three groups of nodes in this DL model are called layers. When the DL model has only one hidden layer, it reduces to a conventional artificial neural network. The lines connecting the input layer neurons, hidden layer neurons, and output layer neurons represent the network weights which need to be learned. From all input connections, the hidden neuron sums up the corresponding weight so the weighted summation is then transformed through an activation function to produce the output of each neuron. The activation functions play an important role in transforming the input and output of hidden layers so they come out in a more useful form (Chollet and Allaire, 2017).

We used the rectified linear unit (RELU) as the activation function for all neurons in the hidden and output layers because our response variables are continuous. In addition, we used a batch size of 56 for implementing the DL model with 1,000 epochs. One epoch means one pass (forward and backward) of the full training set through the neural network, and to complete an epoch, we required a certain number of iterations calculated as the size of the training set divided by 56 (batch size). We used the R statistical software (R Core Team, 2019) for implementing all the models, and the DL model was implemented in the keras library (Chollet and Allaire, 2017). In keras we used the root-mean-square propagation (RMSprop) method with its default values as an optimizer. Also, to avoid overfitting we used dropout regularization, which consists of temporarily removing a random subset (%) of neurons with their connections during training.

For selecting the number of hidden layers, the number of units (neurons) in each hidden layer and the % dropout that needs to be defined, we used a grid search method. In grid search, each hyperparameter of interest is discretized into a desired set of values to be studied, and models are trained and assessed for all combinations of the values across all hyperparameters (that is, a "grid"). The grid search looked for the optimal combination of these three hyperparameters; the values used in the grid were 1, 2, 3, and 4 hidden layers. With regard to the number of units, we tried 80, 160, 240, 320, and 400 units, while for the % dropout (% neurons removed from the DL network), we tried 0%, 5%, 10%, 20%, 25%, and 35%. To select the optimal combination of these three hyperparameters, we implemented a fivefold cross-validation. After obtaining the optimal combination of hyperparameters, the model was refitted using the complete training data.

Random Cross-Validations for Assessing Model Prediction Accuracy

The cross-validation strategy used in this study was a fivefold random cross-validation where 20% of the wheat lines were predicted by 80% of the other lines. This is the random cross-validation CV2 (Burgueño et al., 2012) that mimics a prediction problem faced by breeders in incomplete field trials where lines are evaluated in some, but not all, target environments (usually called *sparse testing,* when not all breeding lines are included for testing in all the environments). In this case, 20% of the lines are not observed in some environments and thus predicted in those environments, but are observed in other environments. When the main purpose of the model is prediction, a reasonable criterion of model quality is the mean squared error of prediction (MSEP) that measures the mean squared distance between the prediction value and the observed value.

Predictions were made for each environment for both the single-environment model (G) and the G×E multienvironment model, using GB, GK, and AK constructed with molecular markers. To make the models comparable in their prediction accuracy as well as their computing time, exactly the same random cross-validations were used for the four methods: GB, GK, AK, and DL.

Experimental Data

We used data from CIMMYT's Global Wheat Program (GWP) consisting of a set of elite wheat lines evaluated under differently managed environmental conditions at CIMMYT's main wheat breeding experiment station in Cd. Obregon, Mexico. These environmental conditions simulated target areas of megaenvironments for the CIMMYT GWP. The wheat lines included in this study were later included in screening nurseries that were distributed worldwide.

Phenotypic Data

The phenotypic data consist of grain yield (ton/ha) records collected during two evaluation years (year 2015–2016 including 1,052 elite wheat lines, and year 2016–2017 including 1,040 elite wheat lines). All trials were established using an alpha-lattice

design with three replicates per line and environment. Each environment was defined by a combination of a planting system (BED = bed planting; FLAT = planting on the flat) and an irrigation intensity (2IR = two irrigations giving moderate drought stress; 5IR = five irrigations representing an optimally irrigated crop). In the 2IR and 5IR regimes, irrigation was applied without measuring soil moisture, and each irrigation added 100 mm of water. Thus, for each of the years (2015–2016 and 2016–2017), four environments BED5IR, FLAT5IR, BED2IR, and FLAT2IR were established. The phenotype used in the analysis was the best linear unbiased estimate (BLUE) of grain yield obtained from a linear model applied to the alpha-lattice design of each year-environment combination. The data included in the present study represent two years of field trials under the same environmental conditions and using similar experimental designs. However, the wheat lines included in both data sets are different and the environmental conditions of the two years were relatively different during the growing season. We therefore decided not to consider a joint analysis of the two data sets.

Genotypic Data

Genotypes were derived using genotyping-by-sequencing technology (GBS; Poland et al., 2012). GBS markers with a minor allele frequency lower than 0.05 were removed. As is typical of GBS genotypes, all markers had a high uncalling rate. In our quality control pipeline, we applied thresholds for the incidence of missing values aimed at maintaining relatively large and similar numbers of markers per data set. We removed markers with more than 60% missing values; this left 15,744 GBS markers available for analysis. Finally, only lines with more than 2,000 called GBS markers were used in the data analysis; this left 515 and 505 wheat lines in years 2015–2016 and 2016–2017, respectively.

Data Repository

The phenotypic and genotypic data for both data sets, year 2015–2016 and year 2016–2017, are available at the following link: http://hdl.handle.net/11529/10548273. Furthermore, basic codes for running the DL and AK kernel methods can be found in the **Appendix**.

RESULTS

Phenotypic Data

A box plot of the grain yield of the four environments in each of the years (2015–2016 and 2016–2017) is displayed in **Figures 1A**, **B**. The two irrigated environments (BED5IR and FLAT5IR) in year 2015–2016 had similar productivity as in year 2016–2017, but the two drought environments (BED2IR and FLAT2IR) produced less grain yield in year 2015–2016 than in year 2016–2017, reflecting the year

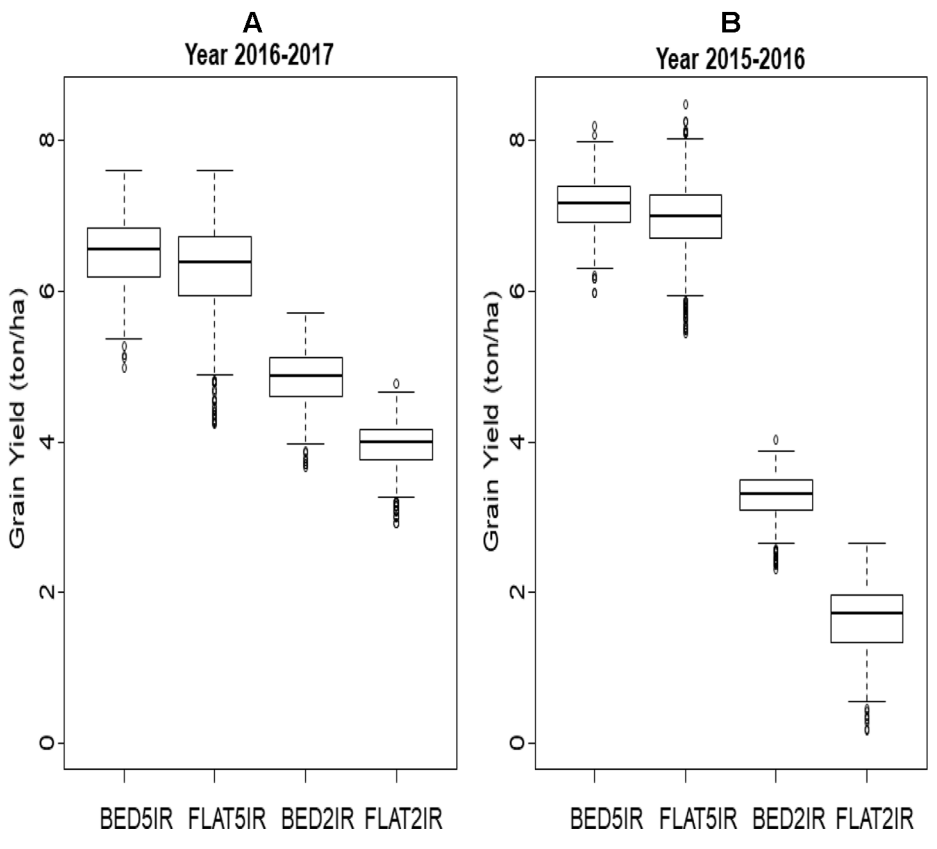


FIGURE 1 | Box plot of grain yield (ton/ha) for four environments (BED5IR, FLAT5IR, BED2IR, and FLAT2IR) for **(A)** year 2016–2017 and **(B)** year 2015–2016.

effect in the drought environments. The narrow-sense heritabilities based on the full model in Eq. (2) for grain yield of environments in year 2015–2016 were BED5IR=0.595, FLAT5IR=0.446, BED2IR=0.590, and FLAT2IR=0.744, and for environments in year 2016–2017 the narrow-sense heritability were BED5IR=0.547, FLAT5IR=0.603, BED2IR=0.565, and FLAT2IR=0.500.

In general, the phenotypic correlations between the four environments in each year were low except for the two drought environments BED2IR and FLAT2IR (0.609 in year 2015–2016 and 0.585 in year 2016–2017) (**Table 1**). The phenotypes of environment FLAT5IR were correlated with those obtained in environments BED2IR and FLAT2IR in year 2016–2017 at ~0.44. The narrow-sense heritability of grain yield in all environment and year combinations was relatively high. Note that these heritability estimates were obtained using genomic markers for the single-environment and the multienvironment models. The heritability of grain yield for years 2016–2017 and 2015–2016 across all four environments were 0.72 and 0.82, respectively. The heritability for year 2016–2017 for the four environments ranged from 0.50 (FLAT2IR) to 0.60 (FLAT5IR), whereas for year 2015–2016, the heritability was 0.45 for FLAT5IR and 0.59 for BED5IR.

Genome-Based Prediction of the Single-Environment and Multienvironment Models

The results for year 2016–2017 for single-environment and multienvironment accuracies are shown in **Table 2** and **Figure 2**, whereas results for year 2015–2016 for single-environment and multienvironment accuracies are shown in **Table 3** and **Figure 3**.

TABLE 1 | Phenotypic correlations among four environments (BED5IR, BED2IR, FLAT5IR, and FLAT2IR) based on grain yield for year 2016–2017 (upper triangle) and year 2015–2016 (lower triangle).

Lower triangle\ upper triangle	BED5IR	FLAT5IR	BED2IR	FLAT2IR
BED5IR	1.000	0.098	0.131	0.006
FLAT5IR	0.260	1.000	0.443	0.446
BED2IR	0.214	0.093	1.000	0.585
FLAT2IR	0.094	0.113	0.609	1.000

Year 2016–2017 Single-Environment Accuracy

The range of MSEP for the single-environment model (G) was between 0.0718 (AK for FLAT2IR) and 0.3883 (DL for FLAT2IR) (**Table 2** and **Figure 2**). Of the four methods implemented (GB, GK, AK, and DL), and the four environments, we found that the lowest MSEPs were obtained with the AK method in three environments, BED5IR, BED2IR, and FLAT2IR and the worst predictions were obtained with DL (except for FLAT5IR, where the best model was DL). The second best model was GK, which performed very similarly to AK (**Table 2**) for all the environments. Environments FLAT5IR and BED2IR had the same MSEP for both GK and AK (0.2297 and 0.0914, respectively).

The average MSEP for method GB was higher than for methods GK and AK, and the average MSEP of DL was also higher than that of any of the other three methods for all environments, except for environment FLAT5IR, where DL had the best prediction accuracy with an MSEP of 0.1589 (**Table 2** and **Figure 2B**). In addition, it is clear from **Figure 2C** that for environment BED2IR, the four methods had very similar prediction accuracies for the single-environment model (G) (GB=0.0977, GK=0.0914 AK=0.0914, and DL=0.1110).

Year 2016–2017 Multienvironment Accuracy

The best method in terms of MSEP was GK for all the environments under the G×E genomic model, while the lowest MSEP of 0.0624 was for environment FLAT2IR. The environment with the highest average MSEP was FLAT5IR for the DL method (0.2797) (**Table 2** and **Figure 2**). The AK kernel closely followed GK in terms of MSEP accuracy, ranging from 0.0625 (FLAT2IR) to 0.2048 (FLAT5IR). Methods GB and DL were the worst in terms of MSEP accuracy. Interestingly, except for GB, GK, and AK for environment BED5IR, and DL for environment FLAT5IR, the MSEP for model E+G+GE were smaller than the MSEP for model G for all four methods. The models including G×E are more precise than those including only the genomic effect (G), regardless of the method used. The differences between MSEP of method DL versus the MSEP of the other methods were much less for the E+G+GE model than those found for the single-environment model and especially for environments BED5IR and FLAT2IR, where the DL methods had high values for MSEP (see **Figures 2A**, **D**).

TABLE 2 | Average mean-squared-error prediction (MSEP) for year 2016–2017 of single environment (G) and multienvironment G×E models (E+G+GE) for predicting each environment comprising a combination of irrigation level (five irrigation, 5IR; two irrigations, 2IR) under two planting systems (FLAT and BED) for methods GBLUP (GB), Gaussian kernel (GK), arc-cosine (AK), (*l* is the number of layers of the deep kernel), and deep learning (DL).

		GB	GK	AK		DL
Model	Environment	MSEP	MSEP	MSEP	*l*	MSEP
E+G+EG	BED5IR	0.1719 (0.006)	**0.1656** (0.009)	0.1659 (0.009)	1	0.1924 (0.010)
E+G+EG	FLAT5IR	0.2144 (0.025)	**0.2040** (0.028)	0.2048 (0.028)	1	0.2797 (0.018)
E+G+EG	BED2IR	0.0867 (0.009)	**0.0807** (0.008)	0.0811 (0.008)	1	0.1066 (0.004)
E+G+EG	FLAT2IR	0.0669 (0.007)	**0.0624** (0.007)	0.0625 (0.007)	1	0.0977 (0.007)
G	BED5IR	0.1627 (0.019)	0.1545 (0.019)	**0.1544** (0.019)	5	0.3806 (0.012)
G	FLAT5IR	0.2415 (0.033)	0.2297 (0.037)	0.2297 (0.038)	4	**0.1589** (0.013)
G	BED2IR	0.0977 (0.010)	**0.0914** (0.008)	**0.0914** (0.008)	5	0.1110 (0.003)
G	FLAT2IR	0.0749 (0.012)	0.0723 (0.011)	**0.0718** (0.011)	5	0.3883 (0.012)
	Average	0.1396	**0.1326**	0.1327	–	0.2144

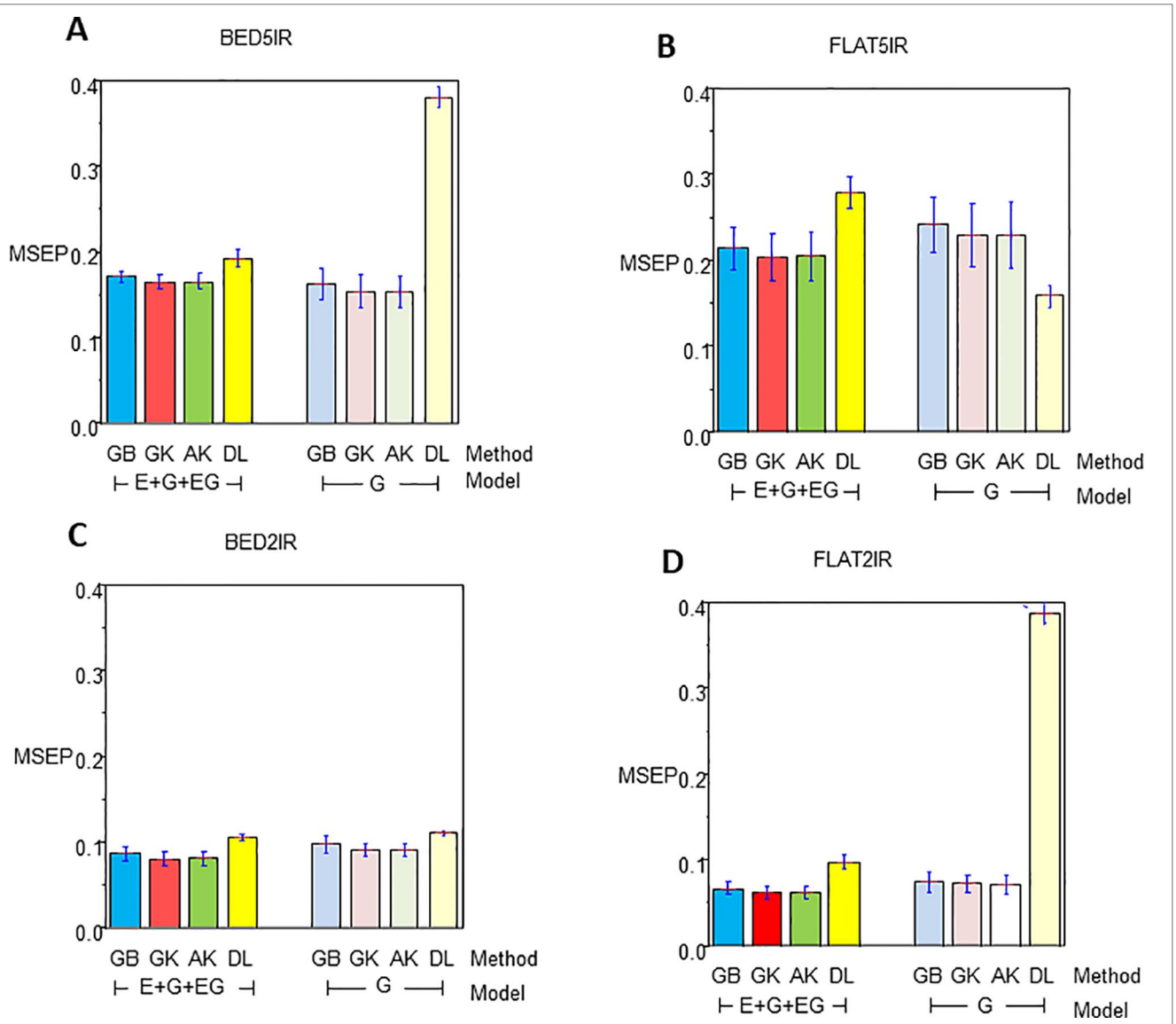


FIGURE 2 | Mean squared error of the prediction for year 2016–2017 for single environment (G) and multienvironment (E+G+GE) models with kernels GB, GK, and AK and the deep learning (DL) method for environments **(A)** BED5IR, **(B)** FLAT5IR, **(C)** BED2IR, and **(D)** FLAT2IR.

TABLE 3 | Average mean-squared-error prediction (MSEP) for year 2015–2016 of single environment (G) and multienvironment G×E models (E+G+GE) for predicting each environment comprising a combination of irrigation level (five irrigation, 5IR; two irrigations, 2IR) under two planting system (FLAT and BED) for methods GBLUP (GB), Gaussian kernel (GK), arc-cosine (AK), (*l* is the number of layers of the deep kernel), and deep learning (DL).

		GB	GK	AK		DL
Model	**Environment**	**MSEP**	**MSEP**	**MSEP**	***l***	**MSEP**
E+G+EG	BED5IR	0.1048 (0.009)	**0.1007** (0.010)	**0.1007** (0.010)	1	0.2403 (0.007)
E+G+EG	FLAT5IR	0.1898 (0.032)	**0.1719** (0.032)	0.1729 (0.032)	1	0.3749 (0.023)
E+G+EG	BED2IR	0.0632 (0.004)	**0.0601** (0.004)	**0.0601** (0.004)	1	0.1355 (0.011)
E+G+EG	FLAT2IR	0.1349 (0.012)	**0.1318** (0.012)	0.1321 (0.012)	1	0.2931 (0.009)
G	BED5IR	0.1095 (0.011)	**0.1031** (0.011)	0.1036 (0.012)	5	0.3307 (0.0124)
G	FLAT5IR	0.1901 (0.010)	0.1819 (0.012)	**0.1792** (0.013)	4	0.4316 (0.025)
G	BED2IR	0.0729 (0.011)	**0.0690** (0.010)	0.0693 (0.010)	5	0.1495 (0.008)
G	FLAT2IR	0.1415 (0.012)	**0.1369** (0.008)	0.1377 (0.007)	5	0.2452 (0.009)
	Average	0.1288	**0.1194**	0.1195	–	0.2751

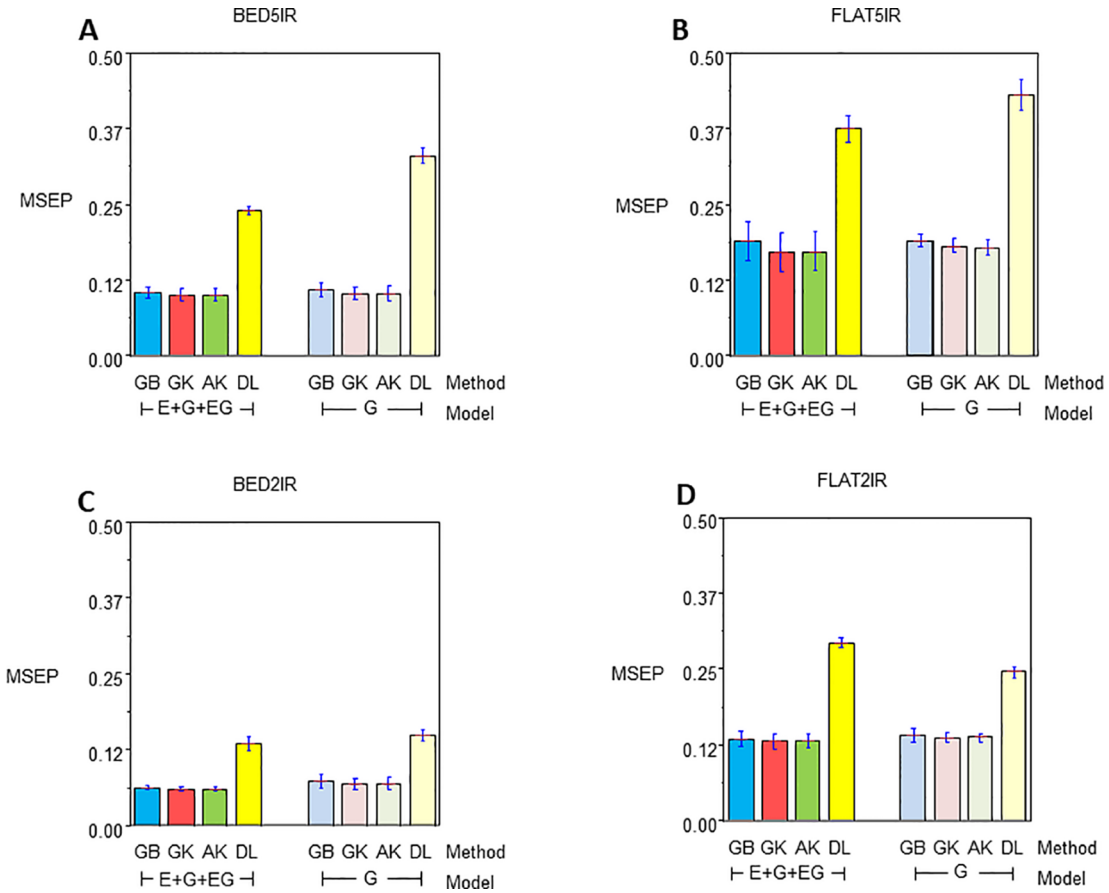


FIGURE 3 | Mean squared error of the Prediction for year 2015–2016 for single environment (G) and multienvironment (E+G+GE) models with kernels GB, GK, and AK and the deep learning (DL) method for environments **(A)** BED5IR, **(B)** FLAT5IR, **(C)** BED2IR, and **(D)** FLAT2IR.

Year 2015–2016 Single-Environment Accuracy

Genome-enabled predictive abilities for the single-environment and multienvironment G×E models are given in **Table 3** and **Figure 3**. For the single-environment models (G), GK had the lowest MSEP in three environments (0.1031 for BED5IR, 0.0690 for BED2IR, and 0.1369 for FLAT2IR) but not for FLAT5IR, where AK was best (**Figure 3B**). The prediction accuracy of the linear kernel GB was lower than that of the nonlinear kernels (GK and AK), ranging from 0.0729 in BED2IR to 0.1901 in FLAT5IR. The DL accuracies of genome-based prediction were the worst, ranging from 0.1495 in BED2IR to 0.4316 in FLAT5IR.

Figure 3 illustrates that the prediction accuracy of DL was not competitive with that of the other methods, which showed a very similar MSEP. The values of MSEP in environment BED2IR were the lowest across all the environments. The highest MSEP values were found in environment FLAT5IR.

Year 2015–2016 Multienvironment Accuracy

The best model in terms of MSEP was GK in all the environments under the G×E genomic model, with the lowest MSEP of 0.0601 in environment BED2IR. The environment with the highest average MSEP was FLAT5IR for the DL method (0.3749) (**Figure 3B**). AK had, together with GK, the two best prediction accuracies, in BED5IR (0.1007) and in BED2IR (0.0601) (**Table 3**). As already mentioned, kernel GK was also the best in FLAT5IR and in FLAT2IR (0.1318). Similar to previous cases, methods GB and DL were the worst in terms of MSEP accuracy. Results show that in all four environments except for FLAT2IR and DL, the MSEP for model E+G+GE were smaller than the MSEP for model G, for all four methods. The models including G×E were more precise than those that only included the genomic effect G.

Furthermore, in general, genome-based accuracy for year 2016–2017 was lower than genomic accuracy computed in year 2015–2016 (**Figure 2** vs. **Figure 3**). The DL method seemed to have more difficulties for learning from the data of year 2015–2016 than from the data of year 2016–2017. This may be partially due to the year effect and to the difficulty of optimizing the hyperparameters of the DL method in this year.

DISCUSSION

The two data sets included in this study represent two years of data with different wheat lines included in each year, but evaluated

under the same experimental environments. Results show that the prediction accuracy of the same models, for instance DL, were very different across years. This may be a result of the different lines used in the two data sets, but more likely the year effects and differences in the G×E interaction. Using the average performance of the lines in each year and performing a two-year analysis may confound the year effect with the different line effects in each year. In order to avoid this possible confounding effect, we performed genomic G×E analyses across environments within each year.

DL Method

DL is a branch of machine learning inspired by the functioning of the human brain. It is helping to automate many tasks that until some time ago only humans were able to perform (e.g., artificial intelligence and autonomous driving). Applications of DL are found in many domains, from social sciences to natural sciences. It is used for classifying exoplanets in astrophysics, for selecting human resources in enterprises, for detecting frauds in banks, and for detecting and classifying many types of cancers, among other things (Chollet and Allaire, 2017). In plant breeding, DL has been used to predict phenotypes of hybrids or lines for which only genomic information is available (Montesinos-López et al., 2018a; Montsinos-López, 2018b; Montsinos-López, 2019a; Montsinos-López, 2019b). However, the training process of DL models is challenging because successful implementation requires large data sets and a tuning process of many hyperparameters (number of hidden layers, number of neurons in each layer, type of activation function, number of epochs, batch size, learning rate, optimizer, etc.). For this reason, when a data set is not large enough, DL training is cumbersome and difficult, because part of the training data must be used to select the optimal combination of hyperparameters.

DL algorithms are flexible and generic and have attracted the interest of researchers working on genome-based predictions. However, the predictive ability of DL versus GBLUP has not been very convincing and not well studied, as pointed out by a recent review by Pérez-Enciso and Zingaretti (2019). Those authors mentioned that initial shallow single-layer neural networks are very competitive with penalized linear methods. However, what has not been addressed are the main difficulties of DL methods when appropriately tuning the hyperparameters and finding an optimal combination of them in order to achieve good genomic-enabled prediction accuracy without overfitting the data. In this study, authors have dedicated important efforts to fitting DL to the two data sets; however, the tuning process has been very difficult and cumbersome, and the results were not completely satisfying. Especially for the data set of the wheat lines from 2015–2016, the prediction accuracy was much smaller for DL than for any of the other models. We can speculate that investing a significant amount of extra time would have led to another set of hyperparameters resulting in better prediction accuracy.

Optimization of the DL Algorithm

The network implemented in this study has no cycles or loops but is a feedforward topology where information moves in only one direction (forward) from the input nodes (prediction variables), through the hidden nodes, and to the output nodes (target variables). As previously described (see the *Material and Methods* section), we performed, for each of the 50 random partitions of the data, an optimization process for selecting the hyperparameters consisting of a grid search method to select the "optimal" set of hyperparameters for that specific partition of the random cross-validation; therefore, it was not possible to give one unique final set of estimated hyperparameters for implementing the DL method. Furthermore, the genomic-enabled prediction accuracy of the DL method will change for every random partition of the data due to the different ranges of the estimated hyperparameters.

Therefore, since the tuning of the DL algorithm is complex and biased for the different range of values of hyperparameters obtained in each of the 50 random partitions, it is reasonable to say that the optimization process for selecting the hyperparameters is suboptimal. This is related to the fact that the optimization process does not guarantee finding a global minimum but may end at a local minimum. This circumstance makes it difficult to tune DL methods.

Deep Kernel Method

Due to the abovementioned difficulties, deep kernel methods that imitate DL methods are an appealing alternative because deep kernels also capture nonlinearity and complex interactions but do not need a complex tuning process, as does conventional DL. The kernel function induces nonlinear mapping from inputs x to feature vectors $\mathbf{\Phi}(x_i)$ by using the kernel trick function: $k(x_i, x_i)=\mathbf{\Phi}(x)\cdot(x_{i'})$ that mimics a single hidden layer or ANN model. Therefore, the iterated mapping of the following equation:

$$k^{(l)}(\boldsymbol{x}_i,\boldsymbol{x}_{i'})=\overbrace{\Phi(\Phi(\ldots\Phi(\boldsymbol{x}_i))}^{l\ times}\cdot\overbrace{\Phi(\Phi(\ldots\Phi(\boldsymbol{x}_{i'})))}^{l\ times} \tag{5}$$

emulates the computation of a DL model (ANN with more hidden layers) where "·" represents the inner product. However, this iterative mapping does not lead to interesting results in linear kernels $[k(\boldsymbol{x}_i,\boldsymbol{x}_{i'})= \boldsymbol{x}_i\boldsymbol{x}_{i'}]$, homogeneous polynomial kernels $[k(\boldsymbol{x}_i,\boldsymbol{x}_{i'})= (\boldsymbol{x}_i\boldsymbol{x}_{i'})^d]$ and Gaussian kernels $[k(\boldsymbol{x}_i,\boldsymbol{x}_{i'})=e^{-\lambda\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|2}]$ (Cho and Saul, 2009). Applying the exponential function twice leads to a kernel which is different from GK, but the qualitative behavior will not be changed (Cho and Saul, 2009). However, in the AK, the recursion $k^{(l)}(\boldsymbol{x}_i,\boldsymbol{x}_{i'})=\overbrace{\Phi(\Phi(\ldots\Phi(\boldsymbol{x}_i)))}^{l\ times}\cdot\overbrace{\Phi(\Phi(\ldots\Phi(\boldsymbol{x}_{i'})))}^{l\ times}$, also alters the kernel qualitatively and mimics an ANN with more than one hidden layer. The results we obtained with AK were similar to those obtained with GK, but with the main advantage that a complex tuning process for choosing the bandwidth parameter across a grid is not required. We also found that GK and AK outperformed the DL method, which might be due to the fact that our data sets are not large enough for successful training of DL and that the main interaction structures within the data were known (G×E) and thus modeled directly.

It is important to point out that the AK deep kernel method is not completely exempt from a tuning process, since one needs to define the depth of the kernel (equivalent to the number of hidden layers). However, choosing such values is straightforward, since we only need to choose integers 1, 2, 3, 4, 5, etc. (Cho and Saul, 2009). We used the maximum marginal likelihood proposed by Cuevas et al. (2019) to select this parameter. As has been the case in many other studies, our results are not definitive, since we only compared the methods with two real data sets. For this reason, we encourage other scientists to do this benchmarking process with other types of data in order to increase the evidence of the prediction performance of these methods. Although our results are not conclusive, there is evidence that the AK (deep kernel) method competes well with DL and the GK, but with the main advantage that the tuning process is considerably less costly. For example, for cycle 2016–2017 with a marker matrix of 1040×8311, the average time for computing the squared distance for the basic GK was 105 s, whereas the computing time (using the same server) for the basic deep kernel AK1 (one layer) was 7 s. Similarly, the average computing time for selecting the bandwidth h for GK was, for each partition, 80 s. In contrast, the average time for selecting the number of layers for AK was 10 s. These differences increase (or decrease) exponentially as the size of the matrices to be used increases (or decreases). This advantage means that the AK method can be implemented in many statistical or machine learning software even by users with no background in statistics, computer science, or machine learning. The deep kernel method can be implemented and used more easily than DL models.

On the Marginal Likelihood and the Number of Hidden Layers (Or Levels) of the AK Deep Kernel Method

To illustrate how the marginal likelihood changed with the number of hidden layers used in the AK deep kernel, we give the example of the marginal likelihood of the observations for environment BED5IR for year 2016–2017 for layers (l) 1 to 8. The corresponding values were -2109.017, -2104.825, -2102.632, -2101.585, -2101.228, -2101.305, -2101.669, and -2102.232, respectively. The maximum likelihood is reached at l=5 (-2101.228). Note that for method GB, the marginal likelihood is -2116.175, which is even lower than the first level (l=1) of the AK deep kernel (-2109.017).

CONCLUSIONS

We performed a benchmarking study comparing a DL model with the AK deep kernel method, with the conventional GBLUP and with the nonlinear GK. We found that AK and GK performed very similar, but when taking the G×E interaction into account, GK constantly predicted best across all four environments and with both data sets. In general, AK and GK were better than GBLUP and DL. Our findings suggest that AK is an attractive alternative to DL and GK, since it offers competitive predictions at low costs in the tuning process. AK is a computationally simple model that makes it possible to emulate the behavior of DL networks with a large number of neurons. In general, the results of this study with respect to DL are not conclusive because the low performance of DL for year 2015–2016 may be partially a result of suboptimal hyperparameters.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

We are grateful for the financial support provided by the Bill & Melinda Gates Foundation and CIMMYT's CGIAR CRP (maize and wheat), as well as the USAID projects (Cornell University and Kansas State University) that generated the CIMMYT wheat data analyzed in this study. We acknowledge the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806.

REFERENCES

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., and Babu, R. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55, 154. doi: 10.2135/cropsci2014.07.0460

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

Cho, Y., and Saul, L. K. (2009). Kernel methods for deep learning, in: NIPS'09 *proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS)*. (Neural Information Processing Systems Conference). 342–350.

Chollet, F., and Allaire, J. J. (2017). Deep Learning with R. Manning Publications, Manning Early Access Program, Ed. 1st. (New: New Delhi, India).

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O. A., Jarquín, D., de Los Campos, G., et al (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22 (11), 961–975. doi: 10.1016/j.tplants.2017.08.011

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., et al. (2016). Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* 9 (3), 1:20. doi: 10.3835/plantgenome2016.03.0024

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian Genomic prediction with genotype × environment kernel models. *G3 (Bethesda)* 7 (1), 41–53. doi: 10.1534/g3.116.035584

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., et al. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3 (Bethesda)* 8 (4), 1347–1365. doi: 10.1534/g3.117.300454

Cuevas, J., Montesinos-López, O. A., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep kernel for genomic and near infrared prediction in multi-environments breeding trials. *G3 (Bethesda)* 9, 2913–2924. doi: 10.1534/g3.119.400493

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

Gianola, D., and van Kaam, J.B.C.H.M. (2008). Reproducing Kernel Hilbert spaces regression methods for genomic-assisted prediction of quantitative traits. *Genet.* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semi-parametric procedures. *Genet.* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gianola, D., Okut, H., Weigel, K. A., and Rosa, G. J. M. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12, 87. doi: 10.1186/1471-2156-12-87

Gianola, D., Morota, G., and Crossa., J.Gianola_et_al.,_2014Gianola, D., Morota, G., and Crossa., J. (2014). Genome-enabled prediction of complex traits with kernel methods: What have we learned? Proc. 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada. 17–22 Aug. 2014. Amer. Soc. Animal Sci., Champaign, IL.

González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., and Cairns, J. E. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

Harfouche, A. L., Jacobson, D. A., Kainer, D., Romero, J. C., Harfouche, A. H., Mugnozza, G. S., et al. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotechnol.* S0167-7799 (19), 30114–30113. doi: 10.1016/j.tibtech.2019.05.007

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

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 759–768. doi: 10.1534/genetics.115.177907

Lopez-Cruz, M. A., Crossa, J., Bonnet, D., Dreisigacker, S., Poland, J., Janink, J.-L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a markers × environment interaction genomic selection model. *G3 (Bethesda)* 5, 569–582. doi: 10.1534/g3.114.016097

Martini, J. W. R., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet.* 129 (5), 963–976. doi: 10.1007/s00122-016-2675-5

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

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018a). Multi-environment genomic prediction of plant traits using deep learners with a dense architecture. *G3 (Bethesda)* 8 (12), 3813–3828. doi: 10.1534/g3.118.200740

Montesinos-López, O. A., Montesinos-López, A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018b). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3 (Bethesda)* 8 (12), 3829–3840. doi: 10.1534/g3.118.200728

Montesinos-López, O. A., Vallejo, M., Crossa, J., Gianola, D., Hernández-Suárez, C., M., Montesinos-López, A., et al. (2019a). A benchmarking between deep learning, support vector machine and bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3 (Bethesda)* 9 (2), 601–618. doi: 10.1534/g3.118.200998

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., and Hernández-Suárez, C. M. (2019b). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3 (Bethesda)* 9, 1545–1556. doi: 10.1534/g3.119.300585

Morota, G., and Gianola., D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet.* 5, 363. doi: 10.3389/fgene.2014.00363

Morota, G., Boddhireddy, P., Vukasinovic, N., and Gianola, D. (2014). Kerbel-based variance component estimations and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Front. Genet.* 5, 56. doi: 10.3389/fgene.2014.00056

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. (Cham, Switzerland: Springer-Verlag) 6330. doi: 10.1007/978-1-4612-0745-0

Pérez-Elizalde, S., Cuevas, J., Pérez-Rodríguez, D., and Crossa, J. (2015). Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. *J. Agric. Biol. Environ. Stat.* 20, 512–532. doi: 10.1007/s13253-015-0229-y. (JABES).

Pérez-Enciso, M., and Zingaretti, L. M. (2019). A guide for using deep learning for complex trait genomic prediction. *Genes* 10, 553. doi: 10.3390/genes10070553

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manes, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric models for genome-enabled prediction in wheat. *G3 (Bethesda): Genes G3 (Bethesda)* 2, 1595–1605. doi: 10.1534/g3.112.003665

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S. Y., and Manes, Y. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

R Core Team. (2019). *R: A language and environment for statistical computing*. (Vienna, Austria: R Foundation for Statistical Computing).

Souza, M. B., Cuevas, J., de O. Couto, E. G., Pérez-Rodríguez, P., and Jarquín, D. (2017). Genomic-enabled prediction in maize using kernel models with genotype environment interaction. *G3 (Bethesda): Genes Genomes Genet* 7, 1995–2014. doi: 10.1534/g3.117.042341

Van Raden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull. Bull.* 37, 33– 36.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Comput.* 10 (5), 1203–1216. doi: 10.1162/089976698300017412

Zhang, X., Perez-Rodriguez, P., Burgueño, J., Olsen, M., Buckler, E., and Atlin, G. (2017). Rapid cycling genomic selection in a multiparental tropical maize population. *G3 (Bethesda)* 7, 2315–2326. doi: 10.1534/g3.117.043141

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Crossa, Martini, Gianola, Pérez-Rodríguez, Jarquin, Juliana, Montesinos-López and Cuevas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

Basic Codes for AK

```
######## Equation (3)
### AK1.fun:Build the base kernel (AK1) of AK with level one
AK1.fun<-function(X){
   n<-nrow(X)
   cosalfa<-cor(t(X))
   angulo<-acos(cosalfa)
   mag<-sqrt(apply(X,1,function(x) crossprod(x)))
   sxy<-tcrossprod(mag)
                    AK1<-(1/pi)*sxy*(sin(angulo)+(pi*matrix(1,
n,n)-angulo)*cosalfa)
   AK1<-AK1/median(AK1)
   colnames(AK1)<-rownames(X)
   rownames(AK1)<-rownames(X)
   return(AK1)
    }
#######  ### marg.AK function: Select the optimal recursion
level
marg.AK <- function(y,AK1,ml){
        lden.fun<-function(phi,nr,Uh,Sh,d){
      lden <- -1/2*sum(log((1+phi*Sh)))-(nr-1)/2*log(sum(d^2/
((1+phi*Sh))))
        lden <- -(lden)
        return(lden)
        }
        vero<-function(y,GC) {
        Kh <- GC
        eigenKh <- eigen(Kh)
        nr<- length(which(eigenKh$val>1e-10))
        Uh <- eigenKh$vec[,1:nr]
        Sh <- eigenKh$val[1:nr]
        d <- t(Uh)%*%scale(y,scale=F)
        sol <-optimize(lden.fun,nr=nr,Uh=Uh,Sh=Sh,d=d,lower=
c(0.0005),upper=c(200))
        phi<-sol[[1]]
   log.vero<--1/2*sum(log((1+phi*Sh)))-(nr-1)/2*log(sum(d^2/
((1+phi*Sh))))
        return(log.vero)
        }
        GC<-AK1
        l<-1
        GC2<-GC
        vero1<-vero(y=y,GC=GC2)
        m<-0
        while( m==0 && (l<ml)){
           l<-l+1
           GC<-AK.fun(AK1=GC2,nl=1)
           GC2<-GC
           vero2<-vero(y=y,GC=GC2)
           if(vero2<vero1) m=1
           vero1<-vero2
           }
       return(l-1)
       }
######### Equation (4)
### Kernel.function: Build the AK kernel, with the base kernel
(AK1) and the recursion level (nl)
AK.fun<-function(AK1,nl){
   n<-nrow(AK1)
   AK<-AK1
    for ( l in 1:nl){
       Aux<-tcrossprod(diag(AK))
       cosalfa<-AK*(Aux^(-1/2))
       cosa<-as.vector(cosalfa)
       cosa[which(cosalfa>1)]<-1
       angulo<-acos(cosa)
       angulo<-matrix(angulo,n,n)
            AK<-(1/pi)*(Aux^(1/2))*(sin(angulo)+(pi*matrix(1,
n,n)-angulo)*cos(angulo))
     }
   AK<-AK/median(AK)
   rownames(AK)<-rownames(AK1)
   colnames(AK)<-colnames(AK1)
   return(AK)
 }
################################ Fitting the G (single
enviroment) model
### Inputs: Matrix markers (X), observations  (y)
library (BGGE)
AK1<-AK1.fun(X)
trn<-!is.na(yna)
tst<-is.na(yna)
AKtrn<-AK1[trn,trn]
l<-marg.AK(y=y[trn],AK1=AKtrn,ml=30)
AK<-AK.fun(AK1=AK1,nl=l)
K<-list(list(Kernel=AK,Type="D"))
fit<-BGGE(y=yna,K=K,ne=1,ite=12000,burn=2000,thin=2,verb
ose=T)
cor(fit$yHat[tst],y[tst],use="pairwise.complete.obs")

Basic codes for DL
####Input and response variable
  X_trn=
   X_tst=
   y_trn=
   y_tst=
   Units_O=400
   Epoch_O= 1000
   Drop_O=0.05

      ###########specification of the Deep neural network
#################
   model_Sec<-keras_model_sequential()
   model_Sec %>%
    layer_dense(units =Units_O , activation ="relu", input_shape
= c(dim(X_trn)[2])) %>%
    layer_dropout(rate =Drop_O) %>%
    layer_dense(units =Units_O , activation ="relu") %>%
    layer_dropout(rate =Drop_O) %>%
    layer_dense(units =Units_O , activation ="relu") %>%
    layer_dropout(rate =Drop_O) %>%
    layer_dense(units =Units_O , activation ="relu") %>%
```

```
  layer_dropout(rate =Drop_O) %>%
  layer_dense(units =1)
###########Compiling the model #################
model_Sec %>% compile(
  loss = "mean_squared_error",
  optimizer = optimizer_adam(),
  metrics = c("mean_squared_error"))
###########Fitting the model #################
ModelFited <-model_Sec %>% fit(
  X_trn, y_trn,
  epochs=Epoch_O, batch_size =56, verbose=0)

####Prediction of testing set ########################
Yhat=model_Sec %>% predict(X_tst)
y_p=Yhat
y_p_tst=as.numeric(y_p)
```

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 January 2020
doi: 10.3389/fgene.2019.01384

Genomic Diversity Evaluation of *Populus trichocarpa* Germplasm for Rare Variant Genetic Association Studies

Anthony Piot[1,2,3], *Julien Prunier*[1,2,3], *Nathalie Isabel*[4], *Jaroslav Klápště*[5], *Yousry A. El-Kassaby*[6], *Juan Carlos Villarreal Aguilar*[3,7,8] *and Ilga Porth*[1,2,3]*

[1] Department of Wood and Forest Sciences, Université Laval, Quebec, QC, Canada, [2] Institute for System and Integrated Biology (IBIS), Université Laval, Quebec, QC, Canada, [3] Centre for Forest Research, Université Laval, Quebec, QC, Canada, [4] Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Quebec, QC, Canada, [5] Scion, Rotorua, New Zealand, [6] Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, Vancouver, BC, Canada, [7] Smithsonian Tropical Research Institute (STRI), Ancon, Panama, [8] Department of Biology, Université Laval, Quebec, QC, Canada

OPEN ACCESS

Edited by:
Charles Chen, Oklahoma State University, United States

Reviewed by:
Lan Zhu, Oklahoma State University, United States
Deborah Weighill, Harvard University, United States

***Correspondence:**
Ilga Porth
ilga.porth@sbf.ulaval.ca

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *19 September 2019*
Accepted: *18 December 2019*
Published: *28 January 2020*

Citation:
Piot A, Prunier J, Isabel N, Klápště J, El-Kassaby YA, Villarreal Aguilar JC and Porth I (2020) Genomic Diversity Evaluation of Populus trichocarpa Germplasm for Rare Variant Genetic Association Studies. Front. Genet. 10:1384. doi: 10.3389/fgene.2019.01384

Genome-wide association studies are powerful tools to elucidate the genome-to-phenome relationship. In order to explain most of the observed heritability of a phenotypic trait, a sufficient number of individuals and a large set of genetic variants must be examined. The development of high-throughput technologies and cost-efficient resequencing of complete genomes have enabled the genome-wide identification of genetic variation at large scale. As such, almost all existing genetic variation becomes available, and it is now possible to identify rare genetic variants in a population sample. Rare genetic variants that were usually filtered out in most genetic association studies are the most numerous genetic variations across genomes and hold great potential to explain a significant part of the missing heritability observed in association studies. Rare genetic variants must be identified with high confidence, as they can easily be confounded with sequencing errors. In this study, we used a pre-filtered data set of 1,014 pure *Populus trichocarpa* entire genomes to identify rare and common small genetic variants across individual genomes. We compared variant calls between *Platypus* and *HaplotypeCaller* pipelines, and we further applied strict quality filters for improved genetic variant identification. Finally, we only retained genetic variants that were identified by both variant callers increasing calling confidence. Based on these shared variants and after stringent quality filtering, we found high genomic diversity in *P. trichocarpa* germplasm, with 7.4 million small genetic variants. Importantly, 377k non-synonymous variants (5% of the total) were uncovered. We highlight the importance of genomic diversity and the potential of rare defective genetic variants in explaining a significant portion of *P. trichocarpa*'s phenotypic variability in association genetics. The ultimate goal is to associate both rare and common alleles with poplar's wood quality traits to support selective breeding for an improved bioenergy feedstock.

Keywords: annotation, genes, genetic architecture, missing heritability, rare defective alleles, small genetic variants, variant calling comparisons

INTRODUCTION

In tandem, phenotypic and genomic diversity assessments are key to understand the genetic regulation and architecture of quantitative traits. Genetic association studies in the form of genome-wide association studies (GWAS) have been used extensively to associate genome-wide polymorphisms to phenotypic variation (Visscher et al., 2017). Typical GWAS are only including common genetic variations. Most of these studies, however, failed to explain most of the observed heritability which is coined the missing heritability problem (Manolio et al., 2009; Brachi et al., 2011). It has been suggested that the missing heritability could be found in other forms of hereditary information such as epigenetic factors, epistasis, and rare genetic variation (Maher, 2008). For over a decade, human geneticists have questioned the role of rare genetic variants in complex diseases (Pritchard, 2001; McCarthy et al., 2008; Manolio et al., 2009). Consequently, the first association studies including rare genetic variants and the associated statistical tests originated in the field of human genetics (Cohen et al., 2004; Hoffmann et al., 2010; Wu et al., 2011).

Generally, most genetic polymorphisms in natural population are rare (*i.e.* found at frequencies lower than 5% in populations). In addition, deleterious variants tend to exist at low frequency in populations because of their negative impact on the phenotype. Non-synonymous genetic variants especially, may have important effects on phenotypes as they alter the amino acid sequence. For instance, a genetic variation leading to a stop codon gain can have drastic impacts on gene products (*i.e.* RNA and protein). Non-synonymous variants can either be missense or nonsense variants. Missense variants result in a codon change that code for a different amino acid while nonsense variants result in truncated or incomplete gene products. Including rare genetic variants in GWAS along with common genetic variants represents a unique opportunity to explain a significant part of the missing heritability (McClellan and King, 2010). Prior to genetic association studies, however, high confidence identification of the genetic polymorphisms within the studied population is required.

Due to their low frequency, rare genetic variants are challenging to identify. Genetic information for a substantial number of individuals is required to find those genetic variants that are rare in a population. In addition, rare genetic variants can easily be confounded with sequencing errors as high-throughput technologies have sequencing error rates between 0.1 to 1% (Fox et al., 2014). Therefore, rare genetic variants must be identified with high confidence before use in GWAS.

High-throughput sequencing permit the resequencing of large numbers of individuals at reasonable cost. Thanks to this technological advancement, genetic data for model species are now sufficiently large to identify rare genetic variants. Currently, the lack of computing resources remains one of the most important challenges to analyze these overwhelming data sets.

To decrease the confusion of low-frequency genetic variants with sequencing errors, strict quality filters are applied from processing of raw sequencing reads to variant discovery to discard bad quality reads and other chimeras. In addition, comparison between variant calling software resulting in a consensus set of Single Nucleotide Polymorphisms (SNP) lead to increased variant detection accuracy (Baes et al., 2014; Fahrenkrog et al., 2017). This approach minimizes the identification of false genetic variants, even though it will discard true genetic variants that were not identified by all variant callers. Using strict quality filtering and variant caller comparison, it is possible to evaluate both common and rare genetic diversity with high confidence. Sensitivity (the number of true positives) and specificity (the number of false positives) of the data processing and variant calling steps should be optimized according to the objectives of the genomic diversity evaluation.

Some populations are expected to contain a higher number of low-frequency genetic variants than others. Natural, outbreeding, and wide-ranging populations are expected to possess higher heterozygosity and a larger number of low-frequency variants (Petit and Hampe, 2006; Evans et al., 2014). On the contrary, domesticated species typically have reduced genetic diversity because of repeated cycles of artificial selection using a few performant breeders with common genetic backgrounds. Because of this high expected number of low-frequency genetic variants, natural forest tree species represent good candidates for rare variant association studies. In forest trees, rare nonsense variants associated to complex traits have been successfully identified. So far, these variants were found in the following genes and species: a *CAD* (*Cinnamyl alcohol dehydrogenase*) in *Pinus taeda* (MacKay et al., 1997), a *CCR* (*Cinnamoyl-CoA reductase*) in two *Eucalyptus* species (Thumma et al., 2005), an *HCT1* (*Hydroxycinnamoyl transferase*) in *Populus nigra* (Vanholme et al., 2013), and a *KANADI* in a *P. trichocarpa* x *P. deltoides* pseudo backcross (Muchero et al., 2015). Other studies also highlighted the ubiquity of rare genetic variants and their role in complex trait regulation in poplar species (Evans et al., 2014; Fahrenkrog et al., 2017).

Populus trichocarpa (Torr. & Gray), is a deciduous forest tree species with important ecological and economical aspects. This fast-growing tree mainly ranges along the North American west coast, from Alaska to Baja California Norte (latitude 31°N to 62°N) (**Figure 1**). The tree is used for pulp and oriented strand board production and represents a good candidate for second-generation biofuel feedstock (Porth and El-Kassaby, 2015). Additionally, *P. trichocarpa* was the first tree species to have its whole genome sequenced with a genome size close to 500Mbp (Tuskan et al., 2006). Since then, hundreds of whole genome resequencing efforts were conducted (Evans et al., 2014; Muchero et al., 2015; McKown et al., 2017) and numerous phenotypic traits related to phenology and wood properties have been measured in common garden experiments (Porth et al., 2013; Evans et al., 2014; McKown et al., 2014; Muchero et al., 2015).

Contrary to other better-studied model species, forest trees have not been subject to extensive genomic evaluation using whole genome resequencing data. Only a handful of such studies have been performed on economically important trees. Silva-Junior and collaborators used pooled resequencing of 240 Eucalyptus tree genome to develop a SNP chip able to identify 60K SNPs (Silva-Junior et al., 2015). *P. trichocarpa* is by far the

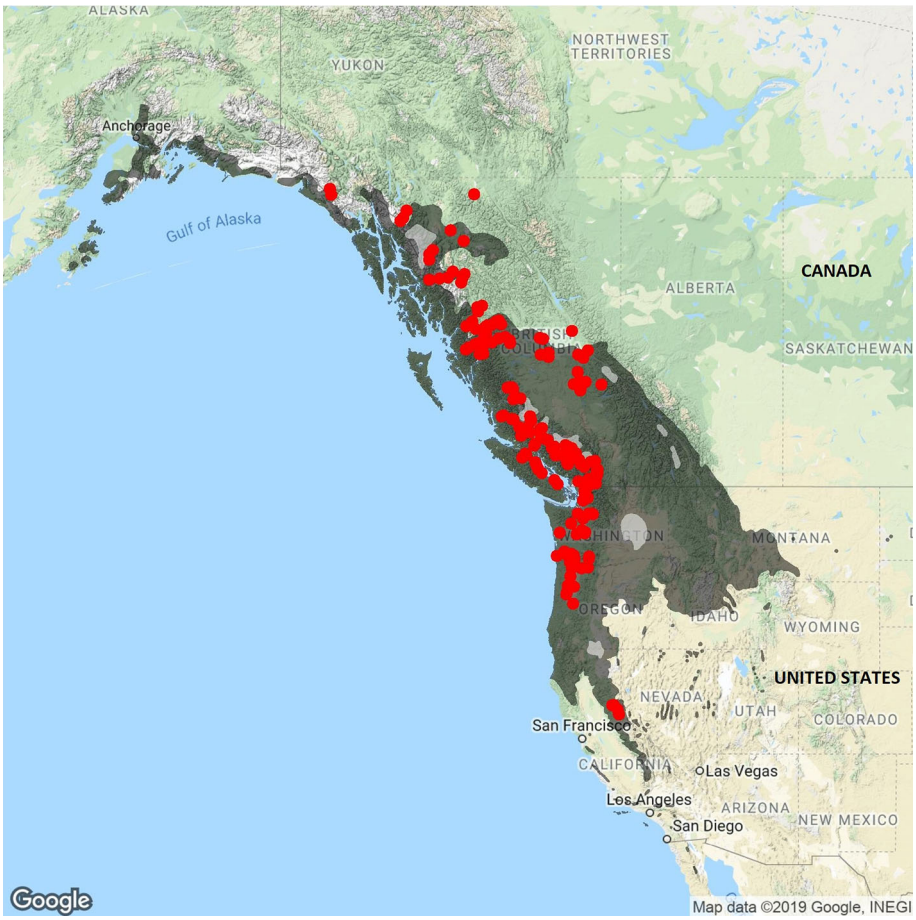


FIGURE 1 | The 1,038 *P. trichocarpa* individuals retrieved from the JGI Genome Portal are represented by red dots across northwestern America. *P. trichocarpa* natural range is defined in dark grey. *P. trichocarpa* natural range was drawn from Little (1971).

forest tree species with the most available genetic resources. In 2014, Evans and colleagues evaluate the genomic diversity across a data set of 544 WGS of *P. trichocarpa* individuals identifying 17M variants (Evans et al., 2014). A second data set developed by the US Department Of Energy (DOE) BioEnergy Science Center (BESC) used 882 WGS of *P. trichocarpa* to identify 28M genetic variants genome-wide (https://bioenergycenter.org/besc/gwas/). To our knowledge, these few studies were the largest genomic evaluation studies performed to date using WGS. It must be noted that large number of individuals have been used in conifers to perform genomic evaluation, but these studies relied on exome or targeted sequencing constrained by the enormous and complex nuclear genome of these species. In 2016, Suren and collaborators used 579 interior spruce samples and 631 lodgepole pine samples to identify 10M SNPs and insertions/deletions (INDELs) in each species using exome capture (Suren et al., 2016).

The goal of the present study was to characterize the genomic diversity of *P. trichocarpa* individuals across its geographic range. The specific objective was to identifying low frequency genetic variants with high confidence that could be used in GWAS including both common and rare genetic variants. We present here a detailed evaluation of small genetic variants using strict quality filtering and comparison between two variant callers. In addition, we provide functional information obtained from the annotation of the discovered genetic variants. Finally, we performed a Gene Ontology (GO) enrichment of genes in which nonsense variants were found. This is the first study in a plant species aiming at rare allele discovery using a large sampling size from whole genome sequencing (over 1,000 individuals).

EXPERIMENTAL PROCEDURES

P. trichocarpa Sequencing Reads

A total of 1,038 unique *P. trichocarpa* individuals were sequenced by the US DOE's BESC (Xie et al., 2009; Slavov et al., 2012). These individuals were sampled across most of *P. trichocarpa*'s geographic range in California, Oregon, and Washington, USA, as well as in British Columbia, Canada (**Figure 1**). These 1,038 *P. trichocarpa* accessions were retrieved online in fall 2017 from the

Joint Genome Institute Genome Portal (https://genome.jgi.doe.gov/portal/) in the form of raw sequencing read files. Whole genome sequencing (WGS) were performed using short paired end reads (100 bp) on an Illumina HiSeq 2000 platform. The sampled individuals were checked for hybrids status after variant discovery based on comparison with closely related species in Principal Component Analysis (PCA; see Results).

Sequencing Reads Quality Filtering

Rare genetic variants and sequencing errors are both found at low frequencies in raw sequencing reads. To differentiate between true genetic variants and sequencing errors, we set stringent quality control on the raw read files. All bioinformatics manipulations were performed on *Cedar* and *Graham* computing servers from Compute Canada and on *Katak* and *Manitou* computing servers at the Institute of Integrative Biology and Systems, Université Laval (Quebec, Canada). First, we trimmed low-quality reads and sequencing adapters using *Trimmomatic* (Bolger et al., 2014) (**Figure 2**). Only bases having a Phred quality score higher than 27 (two chances out of 1,000 that the base is a sequencing error) were kept for further analyses. In addition, reads presenting a mean base quality score below 27 and/or shorter than 50 bases were discarded. The high quality of the cleaned read files was then ensured using *FastQC* (Andrew, 2010) before the alignment and variant calling steps. The mean number of paired reads per accession was about 66M (range: 24 to 321M) after quality filtering.

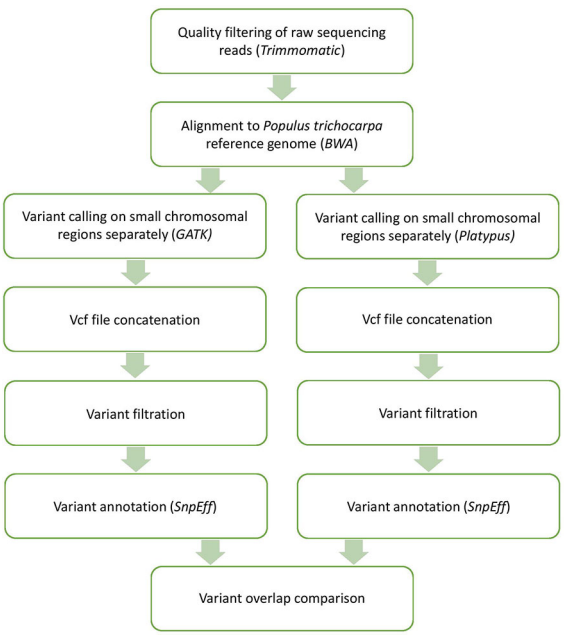


FIGURE 2 | Methodology followed to obtain annotated filtered genetic variants from raw sequencing reads. Software used during each step are in parentheses.

Sequence Alignments

After the quality control steps, each individual accession was align to the reference genome of *P. trichocarpa* version 3.0 using the *Burrow Wheeler Aligner* (BWA; Li and Durbin, 2009) (**Figure 2**). We used the BWA-MEM algorithm that uses seedling alignments with maximal exact matches (MEMs) and then extending seeds with the affine-gap Smith-Waterman algorithm. Twenty-one genomes with average coverage lower than 5X were discarded in order to retain high confidence alignments. Ultimately, 1,017 alignments corresponding to the same number of unique individuals were used in the following analyses.

Variant Calls From Two Different Software Pipelines

In order to obtain high confidence genetic variants, we used two types of variant calling software for result comparison (**Figure 2**): *Platypus* version 0.8.1 (Rimmer et al., 2014) and *HaplotypeCaller* from the *Genome Analysis Tool Kit* version 3.8 (*GATK*; DePristo et al., 2011; Poplin et al., 2017). These two variant calling software are widely used for variant discovery therefore facilitating data reproducibility. They also perform well in terms of sensitivity and precision of discovered variants while being computationally efficient thanks to the implementation of multithreading (Sandmann et al., 2017). *Platypus* enables the user to apply numerous quality filters during variant discovery, while *GATK* offers a filtering tool for use after variant discovery.

Platypus was used to perform single nucleotide variants (SNV) and INDEL calling on the 1,017 alignment files. As suggested by *Platypus* default parameter, bases with quality scores below 20 and reads with mapping quality below 20 were ignored during variant calling. The following custom parameters have been used to address rare variant calling: 1) only variants supported by at least 10 reads were considered; 2) reads having less than 40 bases with a quality lower than 20 were discarded; 3); variants where the median minimum quality in a window of 20 nucleotides around the variant fell below 20 were labelled as "bad reads".

HaplotypeCaller was also used to perform SNV and INDEL calling on the 1,017 alignment files. The filtering tool *VariantFiltration* from *GATK* (DePristo et al., 2011) allowed us to apply quality filters to variants discovered by *HaplotypeCaller*. Parameters for filtering SNPs were set according to *GATK* recommendations for hard filtering. Variants were filtered out when: 1) their quality divided by nucleotide site depth was lower than 2; 2) they were located on a read with an approximate depth lower than 10; 3) their root mean square mapping quality was lower than 40; 4) their phred-scaled p-value using Fisher's exact test was greater than 60; their symmetric odds ratio of 2x2 contingency table to detect strand bias was greater than 3; their Z-score from Wilcoxon rank sum test of alternative vs. reference read mapping qualities was lower than -8; their z-score from Wilcoxon rank sum test of alternative vs. reference read position bias was lower than -12.5. In addition to the recommended parameters for hard-filtering, variants were filtered out if not supported by at least 10 reads.

Using custom python scripts, we filtered out *vcf* files obtained by *Platypus* and *HaplotypeCaller*. More precisely, variants that were attributed the "bad reads" flag were discarded from the *vcf* files obtained by the two types of software. Retained variants were therefore validated by each quality criteria settled during the variant calling phase. Additionally, only INDELs smaller than four nucleotides were included.

Parallelization

Given the large size of our data set, we took advantage of task parallelization in order to minimize computation time for the analyses of the two variant calling software (**Figure 3**). Both *Platypus* version 0.8.1 and *GATK's HaplotypeCaller* version 3.8 allow task parallelization within the software, using multiprocessing for *Platypus* and multithreading for *HaplotypeCaller*. In addition, we used task parallelization outside the software using a scatter-gathering approach. With this method, large files are divided into smaller regions (scattering) analyzed in parallel, then, the results are collected and merged together (gathering). Both approaches are based on task parallelization, but multiple tasks are run within the software using multiprocessing and multithreading, whereas task parallelization is done by the user and happens outside the software for the scatter-gathering approach. A combination of these two approaches was used in order to minimize both analysis and queue time on calculation servers. Portions of the genome were analyzed in parallel (scatter-gathering) while multiple tasks were also running in parallel on each portion (multiprocessing, multithreading). Using *Platypus*, we ran the analysis on each chromosome separately, while we had to divide the analyses on smaller chromosomal regions using *HaplotypeCaller*. The computing resources used for each analysis varied considerably according to the studied chromosomal regions.

Variant Annotations

Filtered variants discovered by the two variant callers were annotated using *SnpEff* (Cingolani et al., 2012). To annotate variants based on the same reference genome used during read alignment we built a custom *SnpEff* database of the annotated genome of *P. trichocarpa* version 3.0.

GO Enrichment

Based on variants recovered by the two variant calling software, we performed a gene ontology (GO) enrichment test using *PANTHER* version 14.1 (Mi et al., 2019). We retrieved the names of *P. trichocarpa* genes in which stop-gained genetic variants were found. Stop-gained variants can have drastic

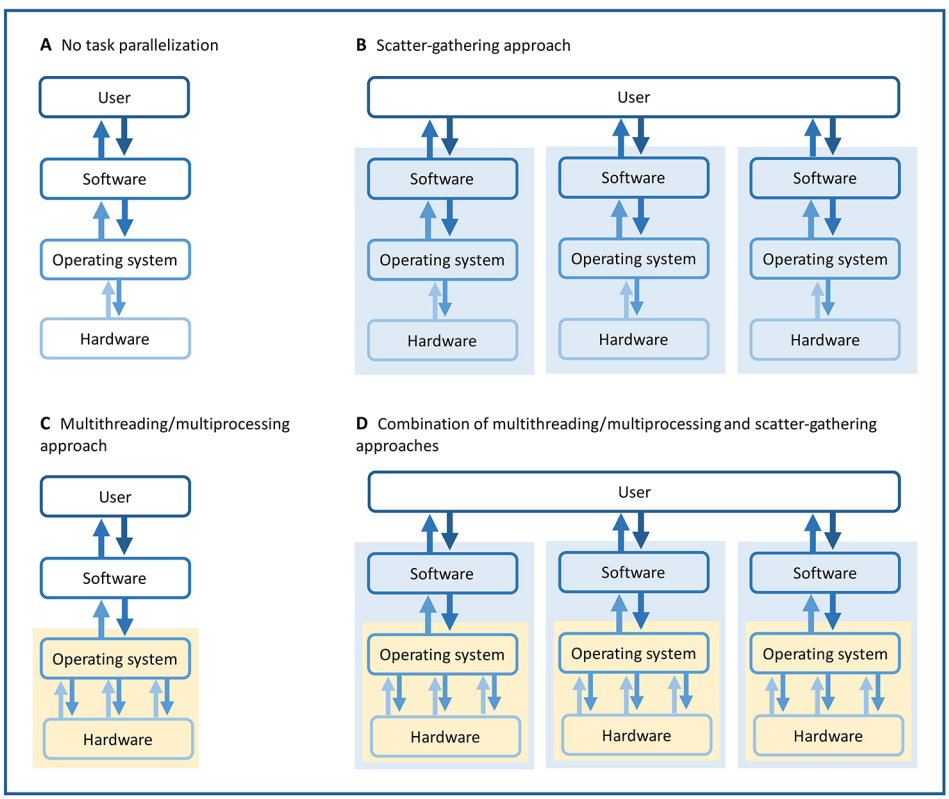


FIGURE 3 | Illustration of different task parallelization approaches. **(A)** Simplest approach with no task parallelization. **(B)** Scatter-gathering approach, where task parallelization is done by the user and happens outside the software. **(C)** Multithreading/multiprocessing approaches, where task parallelization is done by and happens within the software. **(D)** Combination of multithreading/multiprocessing and scatter-gathering, where task parallelization happens outside and within the software. Yellow backgrounds highlight the multithreading and multiprocessing parallelization. Blue backgrounds highlight the scatter-gathering parallelization.

impact on phenotypes and have been found to affect wood composition in poplars (Muchero et al., 2015). We tested whether stop-gained variants are enriched in specific gene functions with respect to biological processes. *PANTHER* version 14.1 does not include *P. trichocarpa* annotations; therefore, we retrieved names of the closest *Arabidopsis thaliana* genes from *P.* trichocarpa genes possessing this type of nonsense variants. The closest *Arabidopsis* genes were determined during the annotation of the *P. trichocarpa* reference genome v3.1 by aligning *A. thaliana* TAIR10 proteins to the *P. trichocarpa* genome (the detailed procedure is available on the *P. trichocarpa* v3.1 Phytozome page). The closest *A. thaliana* gene can be found in the gene annotation file of the *P. trichocarpa* reference genome v3.1 (available on the JGI Genome Portal). This information is available for 84% of the *P. trichocarpa* genes. We used the *PANTHER* classification system to perform a statistical overrepresentation test in GO biological processes, using a Fisher's exact test with the names of *A. thaliana* genes most similar to the targeted *P. trichocarpa* genes. Fisher's exact test was used rather than the binomial test because the former assumes a hypergeometric distribution, which is more accurate for smaller gene lists. Finally, we applied False Discovery Rate (FDR) correction to the obtained p-values. FDR correction was designed to control the false positive rate in the statistical test results and is generally considered a better choice than Bonferroni correction in enrichment analysis (Mi et al., 2019).

RESULTS

Variant Calling From Platypus and HaplotypeCaller

Before filtering, 31,607,230 genetic variants were identified by Platypus in our data set of 1,017 *P. trichocarpa* individuals. After filtering by quality and variant size, this number reduced to 15,734,785 variants, no longer than three consecutive nucleotides and distributed across 14,539,625 polymorphic sites. The majority of these variants (64%) showed a frequency in the population lower than 0.05 (**Figure 4**), *i.e.* found in less than 51 individuals.

Before filtering, 35,597,076 genetic variants have been identified by *HaplotypeCaller* across the 1,017 *P. trichocarpa* individuals. After filtering by quality and variant size, this number reduced to 19,971,499, no longer than three consecutive nucleotides and distributed across 19,478,954 polymorphic sites. Most of these variants (66%) had a frequency lower than 0.05 in the population (**Figure 4**).

Variant Annotation

We used *SnpEff* to annotate genetic variants discovered by *Platypus* and *HaplotypeCaller*. Variant annotation uses information from reference genome annotations to describe genetic variants, such as the variants' inter- or intra-genic locations, and for variants located inside gene, the respective gene name and the effect of the variant on the entire nucleotide

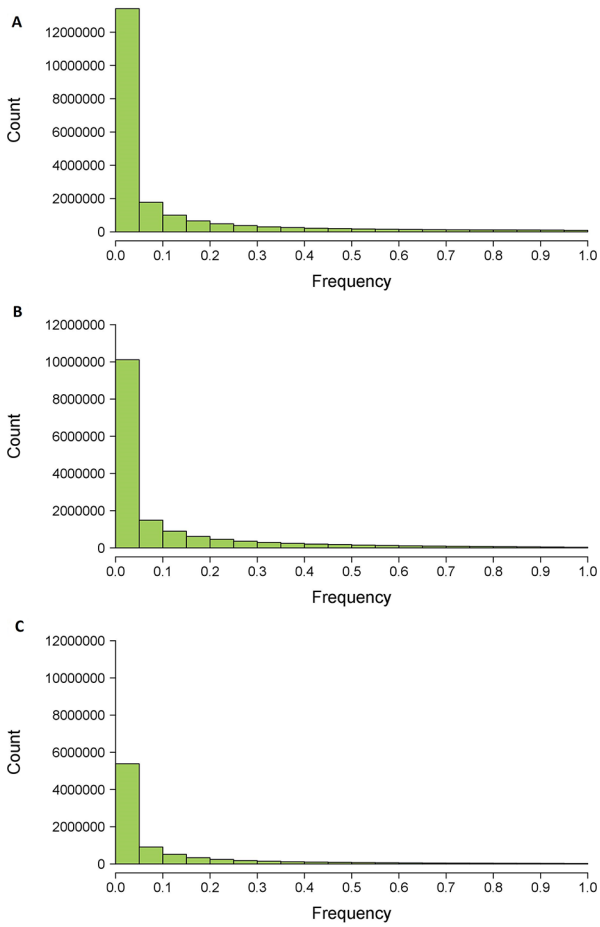


FIGURE 4 | Histograms of variant frequencies after filtering from *HaplotypeCaller* **(A)**, *Platypus* **(B)** and consensus data set **(C)** between both software.

or gene coding sequence. We must note that the total number of variant annotation greatly exceed the total number of genetic variants. The reason is that some variants belong to more than one gene (*i.e.* overlapping genes) and here we report annotations for the effect of variants on each gene they belong to because the same variant can have different effects on different genes. On the contrary, we refer to the total number of genetic variants as the total number of nucleotide variation in the genome.

After annotation of variants discovered by *Platypus*, we found that most of the variants (86%) were located outside of genes, with nearly 11M variants found in intergenic regions and 8.1M and 7.8M variants found in upstream and downstream gene regions, respectively (**Table 1**). Upstream and downstream regions correspond to 5-kb-long regions around genes in *SnpEff* default parameter. The remaining variants (about 4.3M) were located in genic regions, with more than 633K non-synonymous variants (**Table 1**), accounting for 2% of the total.

For annotation of variants discovered by *HaplotypeCaller*, we found that most of the variants were located outside of genes, with 15M variants found in intergenic regions and 11.4M and

TABLE 1 | Annotations obtained from variant calling by *Platypus* and *HaplotypeCaller*.

Annotations	Current study			
	1,014 (1,017) individuals			
	Consensus		*Platypus*	*HaplotypeCaller*
Polymorphic sites	7,313,551	(8,368,838)	14,539,625	19,478,954
Total	7,441,340	(8,497,509)	15,734,785	19,971,499
intergenic variant[a]	5,254,503	(5,645,996)	10,886,077	15,149,344
downstream gene variant[b]	3,955,249	(4,607,452)	7,883,178	11,059,573
upstream gene variant[b]	3,955,094	(4,478,850)	8,086,954	11,413,237
intron variant[c]	1,341,551	(1,762,003)	2,427,258	3,463,071
missense variant[d]*	333,036	(418,974)	559,277	787,053
3 prime UTR variant[e]	269,591	(345,432)	484,519	672,092
synonymous variant[f]	231,894	(324,970)	410,776	554,853
5 prime UTR variant[e]	136,098	(175,634)	245,084	349,285
splice region variant[g]	54,271	(71,655)	95,479	128,316
5 prime UTR premature start gain[h]*	19,099	(24,989)	32,639	45,849
frameshift variant[i]*	9,766	(11,103)	16,937	31,172
stop gained[j]*	8,365	(9,226)	12,967	20,146
splice donor variant[k]*	2,694	(3,208)	4,387	6,237
splice acceptor variant[k]*	2,284	(2,689)	3,807	5,315
stop lost[l]*	1,082	(1,335)	1,994	2,612
start lost[m]*	821	(981)	1,511	2,123
stop retained variant[n]	535	(695)	925	1,246
initiator codon variant[o]	115	(142)	215	280
non_coding_transcript_variant[p]	66	(70)	368	221
intragenic_variant[q]	2	(5)	13	21
exon loss variant[r]	2	(3)	3	3
5 prime UTR truncation[s]	2	(2)	2	2
non canonical start codon[t]	1	(1)	2	2
3 prime UTR truncation[s]	0	(1)	1	1

For the consensus data set, numbers in brackets indicate the number of variants before suspected hybrids removal, while the number outside the brackets indicates the number of variants after suspected hybrids were already removed. *Non-synonymous variants corresponding to genetic variants inside coding regions altering the amino acid sequence of a protein and identified in both caller analyses. [a]Intergenic variant: located in intergenic regions and outside upstream and downstream gene regions. [b]Upstream and downstream variant: located in 5kb regions before and after a gene, respectively. [c]Intron variant: located in non-translated introns of genes. [d]Missense variant: located inside coding regions and resulting in an amino acid change. [e]5 and 3 prime UTR variant: located in 5′ and 3′ untranslated region of a gene, respectively. [f]Synonymous variant: located inside coding regions and not resulting in an amino acid change. [g]Splice region variant: located within the region of the splice site. [h]5 prime UTR premature start gain: resulting in an initiator codon inside the 5′ untranslated region. [i]Frameshift variant: resulting in a reading frame change, because the number of nucleotides inserted or deleted is not a multiple of three. [j]Stop gained: resulting in a premature stop codon in the coding sequence. [k]Splice donor and acceptor variant: changing the 2 nucleotide regions at the 5′ and 3′ end of an intron, respectively. [l]Stop lost: resulting in an elongated gene product because of stop codon loss. [m]Start lost: resulting in initiator codon loss. [n]Stop retained variant: change in one base in the terminator codon, but the terminator remains. [o]Initiator codon variant: change in at least one base of the first codon of a transcript. [p]Non-coding transcript variant: located in a non-coding RNA gene. [q]Intragenic variant: occurs within a gene but falls outside of all transcript features. [r]Exon loss variant: resulting in the loss of an exon from a transcript. [s]5 and 3 prime UTR truncation: causing the reduction of the 5′ and 3′ untranslated region, respectively. [t]Non-canonical start codon: a start codon that is not the usual AUG sequence. The total number of variant annotations does not equal the total number of variants. The reason is that some variants are part of several overlapping genes and may have different effect on different genes.*

11M variants found in upstream and downstream gene regions, respectively (**Table 1**), accounting for 86.1% of the total. The remaining variants (about 6.1M) were located in genic regions, with nearly 901K non-synonymous variants (**Table 1**), accounting for 2.1% of the total.

Variant Calling Overlap

In order to add a further quality criterion to the filtering process of genetic variants, we retained only variants recovered by the two variant calling software used in this study (*i.e. Platypus* and *Haplotype Caller*). We used the *isec* command from *bcftools* to find common variants between the two *vcf* files leading to a consensus variant set (Li, 2011). As a result, 8.5M genetic variants were recovered by both variant calling software, distributed across 8.4M polymorphic sites (**Figure 5**).

Most of the variants were located outside of genes, with 5.7M variants found in intergenic regions and 4.6M and 4.5M variants found in downstream and upstream gene regions, respectively (**Table 1**), accounting for 82.4% of the total. The remaining variants (3.2M) were located in genic regions, with nearly 473K non-synonymous variants (**Table 1**), accounting for 2.6% of the total. Missense variants (419K) accounted for 2.3% of the total and nonsense variants (54K) for 0.3% of the total.

We found 45% more non-synonymous variants compared to synonymous variants. Furthermore, among these non-synonymous variants, missense variants even exceeded synonymous variants by 29%. The total number of genetic variants was only 1.5% higher than the number of polymorphic sites.

To explore result disparities between genomic evaluation studies that used different methods and data set sizes, we also identified the genetic variants commonly found between our study and two other genomic evaluation studies on *P. trichocarpa* (Evans et al., 2014; https://bioenergycenter.org/besc/gwas/). When comparing our results with the study of Evans and collaborators that used 544 P. *trichocarpa* individuals and

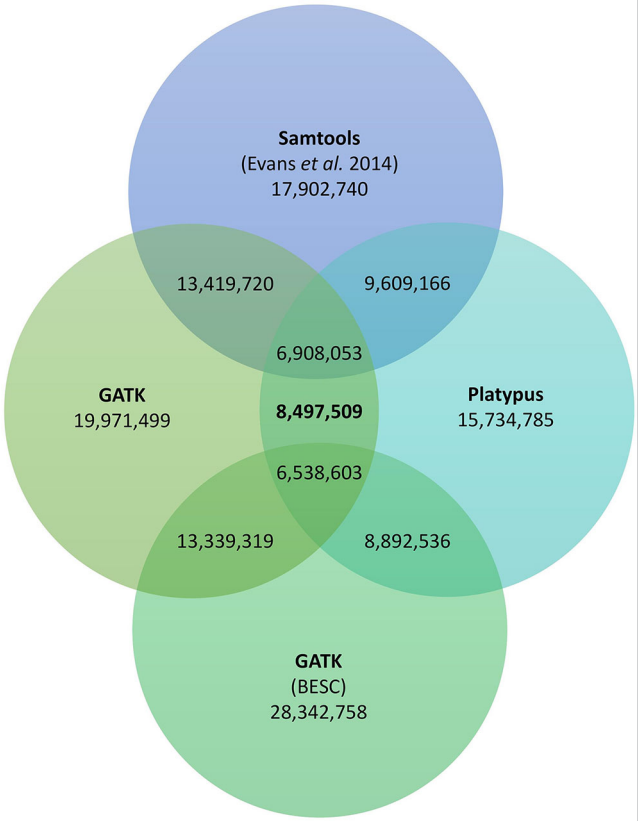


FIGURE 5 | Venn diagram of the number of genetic variants identified in this study in addition to variants identified by Evans and colleagues (Evans et al., 2014) and variants in the BESC data set (https://bioenergycenter.org/besc/gwas/). We used *Platypus* and *GATK's HaplotypeCaller* with 1,017 WGS of *P. trichocarpa* individuals. Evans et al., 2014 used *samtools* and *bcftools* conjointly and 544 WGS of *P. trichocarpa* individuals. The BESC data set was called using multiple tools from *GATK* and 882 WGS of *P. trichocarpa* individuals. The number of variants identified by more than one software are indicated where circles overlap. We note here that some of the individuals used in Evans et al. and the BESC data set are the same as the 1,017 used in the present study.

Samtools as a variant caller we found that 81% (6,908,053) of the variants they identified are also present in our consensus data set. When comparing our results with the BESC data set that used 882 P. *trichocarpa* individuals and tools from *GATK* we found that 77% (6,538,603) of the variants they identified are also present in our consensus data set (**Figure 5**). Interestingly, we found less variants in common with the study using 882 individuals comparing to the study using only 544 individuals. Details regarding individual SNP sets from the two variant callers overlap with the SNPs from Evans et al. and the BESC data set and a summary indicating which variants occur within each SNP set are provided in the Supplement (**Tables S1** and **S2**).

Hybrid Identification

In order to identify potential hybrids in our data set, we also identified the genetic variation across two *Populus balsamifera*, two *Populus deltoides*, one *Populus angustifolia*, and one *Populus fremontii* individuals for comparison. These species are closely related to *P. trichocarpa* and co-occur naturally in some parts of its natural range (Wang et al., 2019). These four species therefore hybridize naturally with *P. trichocarpa*. Raw WGS reads were downloaded from the JGI Genome Portal for *P. balsamifera* and *P. deltoides* (https://genome.jgi.doe.gov/portal/) and from the Genome Sequence Archive of the BIG Data Center for *P. angustifolia* and *P.* fremontii (https://bigd.big.ac.cn/gsa/, accession number CRA001510). Genetic variants from this six genomes were identified with the same bioinformatic pipeline used for *P. trichocarpa* individuals (see Experimental procedures). We used visual identification from a PCA to identify potential *P. trichcocarpa* hybrids. To do so, we filtrated the genetic variants identified across the genomes of these four species and the consensus variant set identified in this study using *plink* (–geno 0.01 –maf 0.1 –hwe 0.01 –LD 50 10 0.1) (Purcell et al., 2007). This filtration step yielded 12,001 variants with which we performed the PCA using *plink* (–pca 2) (**Figure 6**).

The fractions of the genetic variation explained by the two PCAs were 22% and 16%, respectively. Graphical representation of Principal Components (PC) 1 and 2 (**Figure 6**) clearly separated *P. balsamifera, P. deltoides*, *P. angustifolia* and *P. fremontii* from the 1,017 individuals of our consensus variant set. Every individual from the Tahoe population was slightly separated from the core of *P. trichocarpa* individuals. The Tahoe population is the southernmost population of our data set, geographically quite distant from the other *P. trichcocarpa* populations. This suggests that individuals from the Tahoe population differ genetically from other *P. trichocarpa* populations because of geographic distance and not because of hybridization with other *Populus* species. One individual each, from the Skagit, Vancouver Island and Willamette populations, respectively, were located halfway between the core of *P. trichocarpa* individuals and other *Populus* species, strongly suggesting that these individuals may be hybrids or introgressed. These three individuals were therefore removed from the consensus variant set, hence lowering the number of *P. trichocarpa* individuals to 1,014.

Biological Pathways Overrepresented Among Functionally Defective Alleles

Among the consensus variants from *Platypus* and *GATK* variant calling (based on the finalized 1,014 individuals data set), we found that 8,365 stop-gained variants were distributed in 6,327 *P. trichocarpa* genes. These genes corresponded to 3,829 synonymous genes for *A. thaliana*. Analyses of gene function classification in *PANTHER* show that the set of genes containing stop-gained variants was enriched in 106 GO terms with respect to biological processes. Interestingly, multiple GO terms related to wood properties such as cell wall polysaccharide metabolism, cellulose biosynthesis, phenylpropanoid metabolism and plant-type cell wall biogenesis are enriched in genes possessing stop-gained variants (**Table 2**).

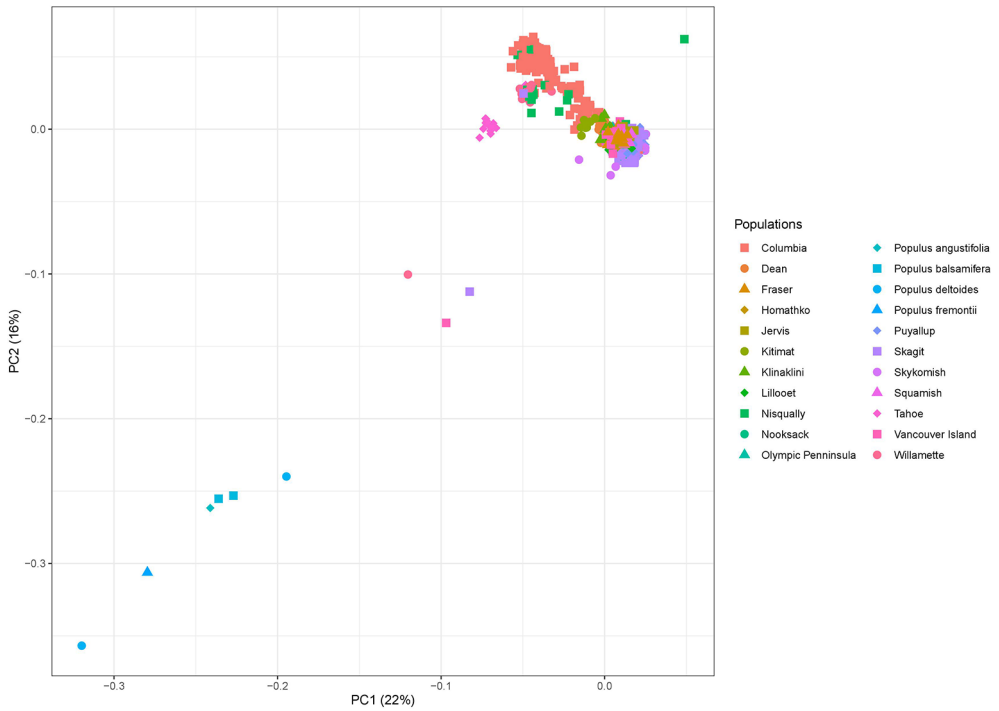


FIGURE 6 | Principal component analysis showing the first two principal components of the genetic variation found across 1,017 *P. trichocarpa*, two *P. balsamifera*, two *P. deltoides*, one *P. angustifolia,* and one *P. fremontii* individual genomes retrieved from various databases for comparison. Note Nisqually-1 (Tuskan et al., 2006) in the upper right corner used as the overall reference.

DISCUSSION

Tool comparison for genomic variant calling has become the standard when using Next Generation Sequencing in clinical diagnostics (Sandmann et al., 2017, *e.g.*). To our knowledge, this approach has never been used in plant sciences when performing a large scale genomic diversity evaluation using WGS. In our study, we evaluated the genomic diversity across 1,017 individuals of *P. trichocarpa* in the form of small genetic variation using an existing set of whole genome sequences. Our goal was to identify rare and common genetic variation in the form of SNPs and small INDELs for subsequent use in GWAS. Using stringent filtering steps and variant calling comparison between two software we identified a set of high confidence genetic variants.

Performance Comparison Between Platypus and HaplotypeCaller

Our data set was computationally heavy with more than one thousand *P. trichocarpa* genomes (~450 Mbp). For this reason, we opted to use variant calling software enabling multithreading to speed up variant identification analyses. *HaplotypeCaller* from *GATK* version 3.8 was considerably slower at identifying variants compared to *Platypus* version 0.8.1. Multithreading for current versions of *GATK* (version 4) is still under development and not safe for production work, therefore, we used a previous version of *GATK* (version 3.8). Both software identify variants based on haplotype reconstruction while *Platypus* also integrates a Bayesian statistical framework for variant discovery. The two software ran on the same data set, but the number of variants identified between each software differed substantially. *HaplotypeCaller* identified 27% more variants in comparison to *Platypus*. This discrepancy in the number of variants identified by the two variant calling software highlight the importance of result comparison between variant callers.

Bioinformatic Approaches

We used a scatter-gathering approach coupled to multithreading to perform variant calling on smaller parts of the data separately. We either conducted the variant calling on chromosomes or smaller chromosomal regions separately to reach acceptable running time and computing resource use. This approach allowed us to identify variants across more than one thousand complete genomes of *P. trichocarpa* within reasonable time. An approach based on a single thread would not have permitted to reach our goal with current computing technologies. The combination of multithreading and scatter-gathering proved very efficient for variant discovery on a large data set.

Consensus Variant Set

Nearly 8.5M genetic variants were identified by the two software and represent high confidence genetic variation. The vast majority of the identified variants had a frequency lower than 0.05 in our data set. Our results are in close agreement with other

TABLE 2 | Results from the Gene ontology (GO) enrichment test performed with PANTHER are presented using a list of *A. thaliana* genes closest to the *P. trichocarpa* genes and related to wood formation and in which stop-gained variants where found querying 1,014 black cottonwood individuals.
Results are sorted hierarchically to better understand the hierarchical relations between over-represented functional classes.

GO biological complete	List of genes with stop-gained variants					
	#	Expected	Enrichment	+/–	raw P-value	FDR
cellulose biosynthetic process	20	6.66	3	+	1.19E-04	8.89E-03
beta-glucan biosynthetic process	22	8.33	2.64	+	2.69E-04	1.82E-02
glucan biosynthetic process	36	15.69	2.29	+	4.52E-05	4.28E-03
cellular polysaccharide biosynthetic process	49	23.6	2.08	+	2.56E-05	2.55E-03
macromolecule metabolic process	1092	792.57	1.38	+	1.15E-26	1.14E-23
organic substance metabolic process	1556	1145.6	1.36	+	1.82E-39	5.41E-36
metabolic process	1798	1334.13	1.35	+	5.28E-47	3.15E-43
organic substance biosynthetic process	452	370.95	1.22	+	6.11E-05	5.13E-03
biosynthetic process	475	393.99	1.21	+	9.41E-05	7.69E-03
cellular biosynthetic process	441	363.73	1.21	+	1.09E-04	8.44E-03
cellular metabolic process	1487	1109.93	1.34	+	3.52E-34	4.20E-31
cellular process	2021	1610.26	1.26	+	8.40E-36	1.67E-32
cellular macromolecule metabolic process	823	606.54	1.36	+	1.10E-17	7.31E-15
cellular polysaccharide metabolic process	73	39.57	1.85	+	9.51E-06	1.07E-03
polysaccharide metabolic process	93	60.67	1.53	+	2.86E-04	1.91E-02
carbohydrate metabolic process	188	138.27	1.36	+	1.38E-04	1.02E-02
primary metabolic process	1430	1049.26	1.36	+	9.69E-36	1.45E-32
cellular carbohydrate metabolic process	92	56.64	1.62	+	5.83E-05	5.04E-03
polysaccharide biosynthetic process	51	28.6	1.78	+	5.06E-04	2.99E-02
cellular carbohydrate biosynthetic process	53	28.74	1.84	+	1.77E-04	1.27E-02
cellular glucan metabolic process	56	29.57	1.89	+	5.06E-05	4.51E-03
glucan metabolic process	56	29.57	1.89	+	5.06E-05	4.44E-03
cell wall polysaccharide metabolic process	39	19.16	2.04	+	1.99E-04	1.38E-02
plant-type cell wall biogenesis	41	20.41	2.01	+	1.89E-04	1.33E-02
cell wall biogenesis	56	31.93	1.75	+	3.56E-04	2.24E-02
cellular component biogenesis	228	171.45	1.33	+	8.33E-05	6.90E-03
cellular component organization or biogenesis	547	426.48	1.28	+	2.54E-08	4.59E-06
phenylpropanoid metabolic process	34	17.08	1.99	+	8.21E-04	4.62E-02
organic cyclic compound metabolic process	548	369.84	1.48	+	1.69E-17	1.01E-14
cellular aromatic compound metabolic process	518	355.81	1.46	+	2.52E-15	1.37E-12

Results are sorted hierarchically to better understand the hierarchical relations between over-represented functional classes. We provide for each GO term (up to seven levels): the number of genes present within the analyzed list (#), the expected number of genes under no GO enrichment (Expected), the enrichment value (Enrichment), the sign of the enrichment (+/–), the P-value associated with the enrichment test without multiple testing correction (raw P-value) and multiple testing corrected using False Discovery Rate (FDR).

genetic diversity evaluation studies of *P. trichocarpa* or closely related species (Evans et al., 2014; Fahrenkrog et al., 2017) and is expected in outcrossing, wide ranging, and undomesticated tree populations (Petit and Hampe, 2006; Fahrenkrog et al., 2017). Most genetic variants are located outside the gene space where nucleotide substitutions are expected to have lower effect on the phenotype and therefore are less subject to purifying selection.

Non-Synonymous/Synonymous Variant Ratio

More surprisingly, inside coding regions, non-synonymous genetic variants were more numerous than synonymous mutations. This pattern has already been observed in a similar study on *P. trichocarpa* (Evans et al., 2014). Given their higher impact on protein sequence, purifying selection is expected to be stronger on non-synonymous variants compared to synonymous ones. A positive ratio of non-synonymous to synonymous genetic substitutions is associated with positive selection (Yang and Bielawski, 2000). *P. trichocarpa* is wide-ranging across the west coast of North America and across a large latitudinal gradient from Alaska to southern California. Individuals in our data set were collected across most of *P. trichocarpa*'s range. Consequently, individuals in this study adapted to different environmental conditions and likely exhibit high genetic diversity in response to local adaptation (Evans et al., 2014). Populations genomic studies are needed to evaluate selection pressures and especially adaption acting across *P. trichocarpa* geographic range.

Comparison With Other Genomic Evaluations on Poplars

Previous studies evaluated the genomic diversity in *P. trichocarpa* (Evans et al., 2014; BESC SNP data set: https://bioenergycenter.org/besc/gwas/) and *Populus deltoides* (Fahrenkrog et al., 2017). Fahrenkrog and colleagues (2017) used targeted resequencing and variant calling overlap between three different software to identified 358K SNPs in 391 unrelated individuals of *P. deltoides*, which is much lower than the 8.5M variants we found. Their final data set included variants found in a subset of genes, thus reducing the size of the analyzed genome. Intergenic variants were also excluded while most genetic variations are usually found in intergenic regions. Moreover,

the variant calling comparison between three different software further decreased the number of identified variants. This approach resulted in a set of high confidence rare and common genetic variants, although less numerous than for studies based on WGS.Using 544 WGS of *P. trichocarpa* individuals, Evans and colleagues (2014) identified 17M SNPs using one variant caller. This number is more than two times higher than the 8.5M variants we identified using comparison between two variant calling software and stringent quality criteria on two times the number of individuals. Evans and colleagues performed no variant filtration, however, and found that stringent filtering had minimal impact on the sensitivity of known SNP discovery while reducing substantially the number of known SNPs passing the filtering threshold (*i.e.* specificity). For the targeted identification of rare genetic variants and for sequencing data with low to moderate sequencing depth we believe that variant filtration is highly beneficial. The DOE's BESC also released a SNP data set (a description of how the SNPs were called is available in the method section of the following study: Weighill et al., 2018). This data set included 28M variants identified across 882 WGS of *P. trichocarpa.* Genetic variants were called using *GATK* tools. First, variants were called independently for each individual using *HaplotypeCaller* and merged afterward. Biallellic SNPs were then extracted and filtered using the *VariantQualityScoreRecalibration* (VQSR) tool. This latter tool uses machine learning to filter variants using a set of known genetic variants (see Weighill et al., 2018 for more information). Similarly, to Evans and colleagues, the BESC data set identified a lot more genetic variants than our study using less individuals. The number of identified variants seems to increase when using only one variant caller. On the contrary, using variant caller comparison the number of individuals scanned does not seem to increase the number of identified variants. Indeed, the number of common variants between our study and the 882 individual data set is slightly lower than the number of common variants between our study and the 544 individuals data set (**Figure 5**). The number of commonly identified variants can even be greater between two different variant callers than between the same variant caller, *i.e. HaplotypeCaller*. These observations show that the use of a certain variant caller is not the main factor determining which variants will be identified, instead parameters used during variant discovery and for filtering along with the comparison between variant caller seem to be of considerable importance.

Quality Filtering and Variant Caller Comparison

Application of stringent filtering criteria before and after variant discovery and the result overlap between variant calling software are key factors for genomic diversity evaluation. With current sequencing technologies and variant calling algorithms, a balance must be found between sensitivity and specificity of variant discovery. Increasing severity in quality filters and increasing the number of variant calling software tend to increase the quality of the identified variants while decreasing the total number of variants. Therefore, the goal of genomic diversity evaluation studies must be clearly stated to ensure that optimal parameters for variant identification are used. Common genetic variants can be identified easily with high confidence without using strict quality filters or comparison between variant calling software. On the contrary, rare genetic variants are difficult to identify with high confidence and require strict quality filtering and overlap between results from various variant calling softwares for reliable identification. When identifying both common and rare genetic variants, as in this study, confidence in the identified variants should be prioritized.

Predicting Models for Increased Specificity

When high quality sets of genetic variants are already available as in model species, one can build models to better detect true and false genetic variants using sets of known genetic variants to increase the specificity of variant identification [*e.g. VQSR* from *GATK* (McKenna et al., 2010)]. Although these models are very useful for human and some other model species, they do not apply to every study. Large sequence data sets such as WGS and Whole Exome Sequencing (WES) and high quality sets of known genetic variants must be used in order to build accurate predicting models. WGS and WES are now widely used in *P. trichocarpa* and high quality sets of known common genetic variants are available. Known high quality sets of rare genetic variants, however, are scarce or even non-existent when considering both genic and intergenic regions. Consequently, we did not use such models to increase the specificity of our variant discovery. The consensus set of 8.5M genetic variants, common and rare, identified in this study will be available as a high quality set of known variants to build models aiming at increasing variant specificity in future genomic diversity evaluations of *P. trichocarpa* and closely related species.

GO Enrichment

We used a GO enrichment test to identify biological pathways overrepresented with genes containing stop-gained genetic variants. A multitude of biological process were overrepresented with genes containing stop-gained variants. Among them, biological processes related to wood properties, and especially secondary cell-wall polysaccharides are of great interest. Previous studies already highlighted the role of functional variants (premature or abolished stop codon, altered start codon, frameshift variant or alternative splice sites) on genes involved in lignin biosynthesis (MacKay et al., 1997; Thumma et al., 2005; Vanholme et al., 2013; Muchero et al., 2015). The lignin and other secondary cell-wall polymers (*i.e.* cellulose and hemicellulose) biosynthesis pathways may be largely affected by functional variants. Thus, these overrepresented biopathways will help us select candidate genes for further analyses. Recent functional mutations are expected to show greater effects on a phenotype, since such functional allelic variants have not undergone selection to much extent. For example, we detected a stop gain mutation at 2.3%

minor allele frequency and with 4.3% carriers in the population for the poplar orthologue of the Arabidopsis *irx10* gene (aka *PtrGUT2B*; Potri.001G068100). Its protein is known to be implicated in xylan backbone formation, and thus a prime target for improving cell wall traits (Porth et al., 2018). Therefore, such variants are important candidates for the purpose of rare variant association studies and ultimately, selective breeding with rare defective alleles (Vanholme et al., 2013; Porth and El-Kassaby, 2015).

CONCLUSION

We identified 8.5M small genetic variants, common and rare, across more than one thousand *P. trichocarpa* individuals sampled throughout the species' range. Use of appropriate quality filtering and variant comparison between two variant callers resulted in high-quality sets of genetic variants. With a data set of 1,017 complete genomes, this is the first time that a genomic diversity evaluation of this magnitude has been conducted in *P. trichocarpa* and, to our knowledge, in any tree species. The high-quality set of known genetic variants identified will be directly available to support other genomic diversity evaluations of *P. trichocarpa* and other closely related species. Moreover, GWAS including rare and common genetic variants will be conducted using those high-quality variants. Thus, starting out from a wealth of genetic variants uncovered in the present study, we will be able to further narrow down the set of important variants for poplar selective breeding.

DATA AVAILABILITY STATEMENT

Raw sequence data are available under https://phytozome.jgi.doe.gov/pz/portal.html.

AUTHOR CONTRIBUTIONS

AP obtained data, analyzed all data, and wrote the manuscript. JP supported data analysis. NI, JK, and YE-K provided valuable insights on poplar genomics. JA obtained co-funding. IP designed the study, obtained funding, and helped in drafting the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). We also acknowledge funding from NSERC Discovery Grants to IP (RGPIN/04748-2017) and JA (RGPIN/05967-2016), respectively, to support this study. WGS data were produced by the US Department of Energy Joint Genome Institute https://www.jgi.doe.gov/in collaboration with the user community.

Support for the poplar GWAS data set was provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER) *via* the Bioenergy Science Center (BESC) under Contract No. DE-PS02-06ER64304. The poplar GWAS project used resources from the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01384/full#supplementary-material

REFERENCES

Andrew (2010). FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., et al. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics* 15, 948. doi: 10.1186/1471-2164-15-948

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12 (10), 232. doi: 10.1186/gb-2011-12-10-232

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Aust.)* 6, 80–92. doi: 10.4161/fly.19695

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305, 869–872. doi: 10.1126/science.1099870

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., et al. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat. Genet.* 46, 1089–1096. doi: 10.1038/ng.3075

Fahrenkrog, A. M., Neves, L. G., Resende, M. F. R., Vazquez, A. I., de los Campos, G., Dervinis, C., et al. (2017). Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. *New Phytol.* 213, 799–811. doi: 10.1111/nph.14154

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., and Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next Gener. Seq. Appl.* 1, 106. doi: 10.4172/jngsa.1000106

Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS One* 5, e13584. doi: 10.1371/journal.pone.0013584

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Little, E. L. (1971). *Atlas of United States trees. Volume 1. Conifers and important hardwoods*. Misc. Publ. 1146. Washington, DC: U.S. Department of Agriculture, Forest Service. 320 p.

MacKay, J. J., O'Malley, D. M., Presnell, T., Booker, F. L., Campbell, M. M., Whetten, R. W., et al. (1997). Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. *Proc. Natl. Acad. Sci.* 94 (15), 8255–8260. doi: 10.1073/pnas.94.15.8255

Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi: 10.1038/nrg2344

McClellan, J., and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell* 141, 210–217. doi: 10.1016/J.CELL.2010.03.032

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., and DePristo, M. A. (2010). The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

McKown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C. B., et al. (2014). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in Populus trichocarpa. *New Phytol.* 201, 1263–1276. doi: 10.1111/nph.12601

McKown, A. D., Klápště, J., Guy, R. D., Soolanayakanahally, R. Y., La Mantia, J., Porth, I., et al. (2017). Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in Populus. *Sci. Rep.* 7 (1), 1831. doi: 10.1038/s41598-017-01893-z

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., et al. (2019). Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* 14, 703–721. doi: 10.1038/s41596-019-0128-8

Muchero, W., Guo, J., DiFazio, S. P., Chen, J.-G., Ranjan, P., Slavov, G. T., et al. (2015). High-resolution genetic mapping of allelic variants associated with cell wall chemistry in populus. *BMC Genomics* 16, 24. doi: 10.1186/s12864-015-1215-z

Petit, R. J., and Hampe, A. (2006). Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Evol. Syst.* 37, 187–214. doi: 10.1146/annurev.ecolsys.37.091305.110215

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178. doi: 10.1101/201178

Porth, I., and El-Kassaby, Y. A. (2015). Using populus as a lignocellulosic feedstock for bioethanol. *Biotechnol. J.* 10, 510–524. doi: 10.1002/biot.201400194

Porth, I., Klápště, J., Skyba, O., Hannemann, J., McKown, A. D., Guy, R. D., et al. (2013). Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. *New Phytol.* 200, 710–726. doi: 10.1111/nph.12422

Porth, I., Maghuly, F., El-Kassaby, Y. A., and Mansfield, S. (2018). Localization of gene expression, tissue specificity of Populus xylosyltransferase genes by isolation and functional characterization of their promoters. *PCTOC* 134, 503–508. doi: 10.1007/s11240-018-1426-5

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137. doi: 10.1086/321272

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., et al. (2017). Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci. Rep.* 7, 43169. doi: 10.1038/srep43169

Silva-Junior, O. B., Faria, D. A., and Grattapaglia, D. (2015). A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytol.* 206, 1527–1540. doi: 10.1111/nph.13322

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol.* 196, 713–725. doi: 10.1111/j.1469-8137.2012.04258.x

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., et al. (2016). Exome capture from the spruce and pine giga-genomes. *Mol. Ecol. Resour.* 16, 1136–1146. doi: 10.1111/1755-0998.12570

Thumma, B. R., Nolan, M. F., Evans, R., and Moran, G. F. (2005). Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. *Genetics* 171, 1257–1265. doi: 10.1534/genetics.105.042028

Tuskan, G. A., Difazio, S., Putnam, N., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., et al. (2006). The genome of black cottonwood, populus trichocarpa (Torr. & Gray). *Science* 80313, 1596–1604. doi: 10.2307/20031305

Vanholme, B., Cesarino, I., Goeminne, G., Kim, H., Marroni, F., Van Acker, R., et al. (2013). Breeding with rare defective alleles (BRDA): a natural Populus nigra HCT mutant with modified lignin as a case study. *New Phytol.* 198, 765–776. doi: 10.1111/nph.12179

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/J.AJHG.2017.06.005

Wang, M., Zhang, L., Zhang, Z., Li, M., Wang, D., Zhang, X., et al. (2019). Phylogenomics of the genus populus reveals extensive interspecific gene flow and balancing selection. *New Phytol.* 225, 1370–1382. doi: 10.1111/nph.16215

Weighill, D., Jones, P., Shah, M., Ranjan, P., Muchero, W., Schmutz, J., et al. (2018). Pleiotropic and epistatic network-based discovery: integrated networks for target gene discovery. *Front. Energy Res.* 6, 30. doi: 10.3389/fenrg.2018.00030

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Xie, C.-Y., Ying, C. C., Yanchuk, A. D., and Holowachuk, D. L. (2009). Ecotypic mode of regional differentiation caused by restricted gene migration: a case in black cottonwood (Populus trichocarpa) along the Pacific Northwest coast. *Can. J. For. Res.* 39, 519–525. doi: 10.1139/X08-190

Yang, Z., and Bielawski, J. P. (2000). Statistical tests of adaptive molecular evolution. *Trends Ecol. Evol.* 15, 496–502. doi: 10.1016/S0169-5347(00)01994-7

Conflict of Interest: Author JK was employed by the company New Zealand Forest Research Institute Limited (Scion).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer LZ and handling Editor declared their shared affiliation.

Copyright © 2020 Piot, Prunier, Isabel, Klápště, El-Kassaby, Villarreal Aguilar and Porth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 February 2020
doi: 10.3389/fgene.2019.01345

Genome-Wide Association Study Uncovers Novel Genomic Regions Associated With Coleoptile Length in Hard Winter Wheat

Jagdeep Singh Sidhu[1†], *Dilkaran Singh*[2†], *Harsimardeep Singh Gill*[1†], *Navreet Kaur Brar*[1], *Yeyan Qiu*[1], *Jyotirmoy Halder*[1], *Rami Al Tameemi*[1], *Brent Turnipseed*[1] *and Sunish Kumar Sehgal*[1*]

[1] Department of Agronomy, Horticulture & Plant Science, South Dakota State University, Brookings, SD, United States, [2] Department of Biology and Microbiology, South Dakota State University, Brookings, SD, United States

OPEN ACCESS

Edited by:
Nunzio D'Agostino, Università degli Studi di Napoli Federico II, Italy

Reviewed by:
Francesca Taranto, Council for Agricultural and Economics Research, Italy
Alessandro Tondelli, Council for Agricultural and Economics Research, Italy

***Correspondence:**
Sunish Kumar Sehgal
sunish.sehgal@sdstate.edu

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 20 September 2019
Accepted: 09 December 2019
Published: 05 February 2020

Citation:
Sidhu JS, Singh D, Gill HS, Brar NK, Qiu Y, Halder J, Al Tameemi R, Turnipseed B and Sehgal SK (2020) Genome-Wide Association Study Uncovers Novel Genomic Regions Associated With Coleoptile Length in Hard Winter Wheat. Front. Genet. 10:1345. doi: 10.3389/fgene.2019.01345

Successful seedling establishment depends on the optimum depth of seed placement especially in drought-prone conditions, providing an opportunity to exploit subsoil water and increase winter survival in winter wheat. Coleoptile length is a key determinant for the appropriate depth at which seed can be sown. Thus, understanding the genetic basis of coleoptile length is necessary and important for wheat breeding. We conducted a genome-wide association study (GWAS) using a diverse panel of 298 winter wheat genotypes to dissect the genetic architecture of coleoptile length. We identified nine genomic regions associated with the coleoptile length on seven different chromosomes. Of the nine genomic regions, five have been previously reported in various studies, including one mapped to previously known *Rht-B1* region. Three novel quantitative trait loci (QTLs), *QCL.sdsu-2AS*, *QCL.sdsu-4BL*, and *QCL.sdsu-5BL* were identified in our study. *QCL.sdsu-5BL* has a large substitution effect which is comparable to *Rht-B1*'s effect and could be used to compensate for the negative effect of *Rht-B1* on coleoptile length. In total, the nine QTLs explained 59% of the total phenotypic variation. Cultivars 'Agate' and 'MT06103' have the longest coleoptile length and interestingly, have favorable alleles at nine and eight coleoptile loci, respectively. These lines could be a valuable germplasm for longer coleoptile breeding. Gene annotations in the candidate regions revealed several putative proteins of specific interest including cytochrome P450-like, expansins, and phytochrome A. The QTLs for coleoptile length linked to single-nucleotide polymorphism (SNP) markers reported in this study could be employed in marker-assisted breeding for longer coleoptile in wheat. Thus, our study provides valuable insights into the genetic and molecular regulation of the coleoptile length in winter wheat.

Keywords: *Triticum aestivum*, coleoptile length, semi-dwarf wheat, genome-wide association study, quantitative trait loci, SNP (Single-nucleotide polymorphism), marker-assisted selection

INTRODUCTION

Successful crop stand establishment is the first critical step for achieving a high yield potential (Rebetzke et al., 2007b; Rebetzke et al., 2014). Temperature and moisture are two major environmental factors that determine the success of seedling emergence out of the soil (Jame and Cutforth, 2004; Hunt et al., 2018). Therefore, to ensure that ideal temperature and moisture are available to the seed, optimum planting depth is critical. In regions with dry soils and higher temperatures, deep seed placement ensures optimum temperature and moisture (Mahdi et al., 1998). Deep sowing of seeds also minimizes winter injury and prevents seed damage caused by animals (Brown et al., 2003), however, it delays emergence.

The coleoptile is a sheath that facilitates the emergence of the shoot through the soil crust in monocots. The length of the coleoptile dictates the maximum depth at which seed can be sown. Thus, genotypes with longer coleoptile can be sown deeper to circumvent dry and high-temperature conditions. Whereas genotypes having shorter coleoptiles may fail to emerge if sown too deep and thus result in a poor stand and eventually leading to production losses (Mahdi et al., 1998; Rebetzke et al., 2005; Rebetzke et al., 2007b). Further, an increase in temperature affects coleoptile length negatively. Thus, such genotype*environmental interactions can be devastating on crop yield (Jame and Cutforth, 2004; Rebetzke et al., 2016). Extremely dry situations during the fall season (Budak et al., 1995; Schillinger et al., 1998) and dry spring in the northern Great Plains lead to a poor establishment of hard winter and hard spring wheat, respectively. Extreme fluctuations in weather with changing climate necessitate an adjustment in the breeding programs towards developing crop varieties having longer coleoptiles to ensure better plant stands and establishment.

Present-day wheat varieties' genetic potential for coleoptile length cannot adequately meet the requirements of deep-sowing farming practices and of changing climate. Two reasons responsible for the poor genetic makeup for coleoptile length are; (1) no dedicated breeding effort has been made for improving coleoptile length of wheat varieties; (2) development of semi-dwarf wheat varieties using dwarfing genes *Rht-B1b* and *Rht-D1b* which suppresses or have association with a locus which suppresses coleoptile length (Allan et al., 1962; Allan, 1980; Yu and Bai, 2010; Li et al., 2011; Rebetzke et al., 2016).

Molecular markers linked to genes or quantitative trait loci (QTLs) can facilitate simultaneous marker-assisted breeding and pyramiding for several traits, avoiding laborious and time-consuming phenotyping. Recently, a few QTL mapping studies in spring wheat have mapped several QTLs that control coleoptile length on chromosomes 1A, 1B, 1D, 2B, 2D, 3A, 3B, 3D, 4A, 4BS (*Rht-B1b*), 4DS (*Rht-D1b*), 5A, 5B, 5D, 6A, 6B, and 7B (Rebetzke et al., 2007a; Spielmeyer et al., 2007; Yu and Bai, 2010; Rebetzke et al., 2014; Singh et al., 2015; Li et al., 2017) However, linkage mapping studies have lower power in identifying QTLs with smaller effect and typically demarcate the QTLs to large genomic regions of 15-20 cM (Tuberosa et al., 2002; Korte and Farlow, 2013).

Nearly all previous studies (Spielmeyer et al., 2007; Yu and Bai, 2010; Rebetzke et al., 2014; Singh et al., 2015) consistently mapped QTLs close to *Rht-B1b* and *Rht-D1b*, however, the diverse populations used in those studies led to the identification of distinct novel loci; on chromosomes 1B, 3D, 4DL, and 5AS using a Chinese wheat variety (Yu and Bai, 2010); on chromosomes 1D, 3A, 6A, and 7B using a population derived from Australian cultivars (Spielmeyer et al., 2007; Rebetzke et al., 2014); on chromosomes 3BS and 3BL using Indian cultivars (Singh et al., 2015); and on chromosomes 1BS, 2DS, 4BS, and 5BL using diverse 893 accessions collected from around the world (Li et al., 2017). This suggests that there are a number of QTLs for coleoptile length and therefore, the potential of utilizing these distinct loci in the development of varieties suitable to specific regions.

Genome-wide association (GWAS) is a powerful tool for dissecting genetic architecture of complex traits with the availability of high-density SNP arrays (Wang et al., 2014) and next-generation sequencing technologies (Poland et al., 2012; Ayana et al., 2018; Ramakrishnan et al., 2019; Sidhu et al., 2019). Further, GWAS can effectively identify many natural allelic variations in a large set of unrelated individuals as compared to the traditional QTL mapping (Huang and Han, 2014). Li et al. (2017) conducted GWAS using a global wheat collection of 893 accessions and identified two major QTLs for coleoptile length. These two QTLs are present on chromosome 4B and 4D, independent of *Rht-B1b* and *Rht-D1b* respectively, but their physical locations are unknown. Though a number of QTLs have been mapped in spring wheat and a few in winter wheat, they may not cover the entire variation for coleoptile length. Further, most of the QTLs cover a large genomic region and information on functional characterization of these QTLs is lacking. The functions of candidate genes have only been reported in one study (Singh et al., 2015) where cell wall expansion genes were found in two QTL regions. The functional characterization of genes is necessary to use them efficiently at the molecular and genetic level. Furthermore, understanding the function of genes will also help in navigating the complexity that arises due to breeding for longer coleoptiles, but shorter shoots simultaneously.

Allan et al. (1962) reported the correlation between coleoptile length and final stand establishment in fall sown winter wheat varieties. However, no study has been done to explore the genetic regions controlling coleoptile length in winter wheat varieties of the USA, even though regions of low-precipitation in the Great Plains and Pacific Northwest necessitates deep sowing to ensure moisture for germination (Budak et al., 1995; Schillinger et al., 1998) and better winter survival. Identification and characterization of QTLs by exclusively using winter wheat varieties will shed light on the underlying diversity for coleoptile length, and provide linked markers to facilitate marker-assisted selection. Further, annotation of genes associated with coleoptile length in the candidate regions will help understand the molecular mechanism of coleoptile length in wheat and other monocots.

The objectives of this study were; (i) mapping QTLs that control the length of coleoptile by conducting genome-wide association analysis in a hard winter wheat panel of 298 winter wheat accessions; (ii) identifying SNP markers linked to QTLs for marker-assisted selection; (iii) identifying candidate genes located in the QTL regions.

MATERIALS AND METHODS

Plant Materials

In the present study, we used a hard winter wheat association mapping panel (HWWAMP) of 298 winter wheat accessions developed under the USDA TCAP project (Guttieri et al., 2015). The total collection of 298 accessions consists of released varieties since the 1940s and breeding lines from the US hard winter wheat growing region including Colorado, Kansas, Michigan, Montana, Nebraska, North Dakota, Oklahoma, South Dakota, and Texas. Additional physiological and agronomic data about the HWWAMP accessions is available in the T3/Wheat database (https://triticeaetoolbox.org/wheat/pedigree/pedigree_info.php).

Experimental Setup

Seed for all 298 HWW accessions were harvested from the field and dried to 11–13% moisture content. The seeds of each line were then carefully cleaned with a Carter Day dockage tester, and clean uniform seeds from the #2 middle sieve were collected for this experiment. Coleoptile lengths of 298 accessions were evaluated in three independent experiments with two replications in each experiment. In each experiment, 10 healthy-looking seeds of each genotype were placed and germinated on a wet paper towel measuring 15 cm x 10 cm (SGB1924B, Anchor Paper Co., USA). Seeds were placed about 1 cm apart with germ end downwards on wet germination paper leaving a 1 cm margin at the bottom. Another wet germination towel of the same size was placed on top. These two germination papers enclosing the seeds were carefully placed in a plastic bag and kept at 4°C for 48 h to break the seed dormancy. Later the plastic bags were hanged vertically in a growth chamber for 14 days at 18°C. After 14 days, coleoptile lengths were measured using a ruler. Distance between the tip of coleoptile and scutellum was considered as the length of coleoptile.

Data Analysis

The phenotypic data was analyzed using the linear mixed model (LMM) approach, considering all factors as random. The analysis was conducted in R environment (R Core Team, 2016) using R package 'minque' (Wu, 2014) based on the model:

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + R_{i(j)} + e_{ijk} \quad (1)$$

where "μ" stands for population mean, "G" stands for genotypes, "E" for experiments, "R" for replications nested under experiments, and "e" for the random error. Broad-sense heritability (H^2) was calculated using equation 2:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2/n + \sigma_{G*E}^2/nr} \quad (2)$$

Where, σ^2_G = *genotype*, σ^2_E = *experiment*, σ^2_{G*E} = *genotype * experiment*, r = *number of replications*, and n = *number of experiments*.

Genotyping

The HWWAMP was genotyped using the wheat Infinium 90K iSelect array (Illumina Inc. San Diego, CA) under the USDA-TCAP (Cavanagh et al., 2013) and the genotypic data (21,555 SNPs) was obtained from the T3 Toolbox (https://triticeaetoolbox.org/wheat/genotyping/display_genotype.php?trial_code=TCAP90K_HWWAMP). To avoid any spurious marker-trait associations, the SNP markers with a minimum allele frequency (MAF) < 0.05 and more than 10% missing SNP data were excluded from further analyses, leaving 15,590 SNP markers. The genetic positions of the wheat Infinium 90K iSelect SNP markers used in the study were obtained from the consensus genetic map of 46,977 SNPs (Wang et al., 2014). The SNP flanking sequences were mapped to wheat Chinese Spring RefSeq v1.1 assembly (IWGSC et al., 2018) using BLASTN to identify the physical location of the mapped SNPs.

Population Structure And Linkage Disequilibrium

Population structure among the 298 winter wheat accessions was studied to determine any relationship between breeding programs and coleoptile length. We used a set of 15,590 SNP markers with MAF > 0.05 and less than 10% missing genotypic data to estimate the population structure using a model-based Bayesian cluster analysis program, STRUCTURE v2.3.4 (Pritchard et al., 2000). The admixture model was used with 10 independent replicates for each value of genetic groups (*K = 1-10*) followed by 10,000 iterations of burn-in and 10,000 Markov Chain Monte Carlo (MCMC) iterations. Structure Harvester (Earl and vonHoldt, 2012) was used to extract the output of the structure analysis. The optimum number of clusters was inferred using statistic ΔK (delta K) (Evanno et al., 2005), which is based on the rate of change in the log probability of given data, between successive K values. Furthermore, we conducted principal component analysis (PCA) in TASSEL 5.0 (Bradbury et al., 2007) using the same set of markers and used the PCA covariates for GWAS analysis. Linkage disequilibrium (LD) decay distances for the HWWAMP were calculated using TASSEL v5.0 (Bradbury et al., 2007) with only 1,842 markers taking out non-informative markers in our previous study (Ayana et al., 2018). The estimated r^2 values were plotted against the genetic distance (cM) to elucidate the LD decay for all as well as individual genomes. The LD (r^2 > 0.1) decay distance of about 4.5 cM was estimated for the whole genome (Ayana et al., 2018).

Marker Trait Associations

Genome-wide association mapping was conducted using 15,590 SNPs and coleoptile data from 298 HWWAMP accessions using the mixed linear model (MLM) (Yu et al., 2006) implemented in

TASSEL (Trait Analysis by association, Evolution, and Linkage) v 5.0 software (Bradbury et al., 2007). MLM is mathematically represented as:

$$y = X\beta + \mathrm{Zu} + e \quad (3)$$

where y represents the vector of the phenotypic values, β represents fixed effects due to the marker and population structure, u represents the vector of the random effects, e represents the vector of residuals, and X and Z are the incidence matrices for β and u, respectively.

MLM was used as it incorporates kinship and population structure as covariates to minimize the confounding effects, reducing the probability of type-I error when compared to the general linear model (GLM). Kinship (K) was estimated using the Centered IBS (identity by state) method in TASSEL v 5.0 (Endelman and Jannink, 2012). By default, TASSEL v5.0 uses PCA as covariates to adjust for the population stratification. We incorporated the first four PCAs as covariates in the MLM model to reduce the confounding effects. As the false discovery rate (FDR) correction for multiple testing was too stringent, markers with a −log10(p-value) > 3 were considered as significant associations. Furthermore, MLM results from TASSEL v5.0 were confirmed using MLM and SUPER in the genome association and prediction integrated tool (GAPIT) (Lipka et al., 2012) implemented in the R environment (R Core Team, 2016). Further, the identified QTLs were also subjected to five-fold validation (Ramakrishnan et al., 2019). Briefly, the population was randomly divided into five subsets of equal size and process was repeated five times. Out of each of the five subsets, four (240 lines) were used for marker-trait association analysis and the last set (60 lines) was used to cross-validate the significant markers using t-test among different alleles of each significant SNP marker.

Identification and Annotation of the Candidate Genes in the QTL Regions

We used the flanking sequence of significant SNPs to physically map them on Chinese Spring Refseqv1.1 (IWGSC, 2018) using BLASTN search with an E-value cut off $1e^{-50}$. To demarcate the candidate QTL regions, the SNP markers with $P < 0.005$, both up- and downstream of the most significant marker, were identified. The coding sequences (CDS) of high confidence genes (https://urgi.versailles.inra.fr/jbrowseiwgsc) from each of these QTL regions were extracted in the FASTA format and Blast2Go software (https://www.blast2go.com) was used for functional gene annotation. Consequently, we identified the candidate genes that may be associated with coleoptile length based on the LD Decay in the region (Ayana et al., 2018) and their putative functions after a thorough review of the literature.

RESULTS

Phenotypic Variance

Coleoptile length within 298 winter wheat accessions varied from 49.40 to 111.00 mm with an overall mean of 74.65 mm (**Supplementary Table S1**). LMM analyses revealed that the three experiments were consistent (**Figure 1**, **Supplementary Table S2**). Average coleoptile length for the three independent experiments (further referred to as Exp1, Exp2, and Exp3) was 76.10, 73.50, and 74.00 mm, respectively (**Figure 1**). Overall, only 1.24% of the variation was contributed by experiments and replications together. The estimated broad-sense heritability for coleoptile length was 73.4%. The median coleoptile length was 71.75 mm. About 25% of the genotypes were less than 66.33 mm and 25% were above 81.17 mm. The majority of the genotypes in all the experiments reached a coleoptile length of ≥ 65 and ≤ 70 mm (**Figure 1**). An accession from Oklahoma 'OK05723W' had the shortest coleoptile (49.40 mm) while the cultivar 'AGATE' had the longest coleoptile (111.00 mm). We also evaluated if the seed source (location) may have an impact on the coleoptile length by comparing the coleoptile length of two varieties from four different locations. The genotype and location effects were found to be significant for two genotypes. However, genotype*location interaction was non-significant, with the ranking of two varieties being the same across four locations. Thus, the growing environment did not significantly impact the ranking of the genotypes for coleoptile length.

LD Analysis and Population Structure

The hard winter wheat association-mapping panel was characterized for LD in our previous study (Ayana et al., 2018). LD decay was calculated based on the r^2 values for the whole genome and within each genome of the association panel. The distance where LD value (r^2) decreases below 0.1 or half strength of D' (D' = 0.5) was estimated based on the curve of the nonlinear logarithmic trend line. LD dropped to 0.5 at about 4.5 cM for whole-genome; whereas, LD extent in A and B and D genomes was around 3.4 and 3.6 cM, but much larger in D genome (14.2 cM) owing to fewer markers.

The association-mapping panel used in this study is comprised of 298 winter wheat cultivars/breeding lines from different regions of the USA. We investigated the population structure to reveal if the association-mapping panel is structured, based on the breeding programs/origin; and figure out any relationship of structure with the coleoptile length. We identified four sub-populations in the HWWAMP, namely: P1, P2, P3, and P4 (**Supplementary Figure S1**). Populations P1, P2, P3, and P4 consist of 120, 34, 33, and 111 genotypes, respectively with a corresponding average coleoptile length of 79.13, 75.18, 69.91, and 72.20 mm. The average coleoptile length of population P1 was higher than the populations P2, P3, and P4; however, it was statistically different only from P3 and P4 (**Supplementary Table S3**).

Marker Trait Associations (MTAs)

In total, GWAS analysis using MLM in TASSEL v5.0 identified 46 significant SNPs ($P < 0.001$) in nine genomic regions present on seven different chromosomes (**Supplementary Table S4**). Based on the threshold value of −log10 (p-value) > 3, we identified 14, 1, 1, 2, 18, 6, and 4 significant SNPs on chromosomes 2A (*QCL.sdsu-2AS*), 2B (*QCL.sdsu-2BS*), 2D

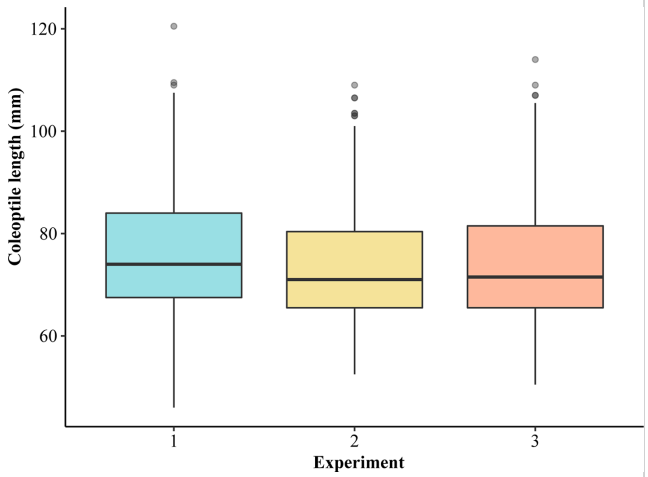


FIGURE 1 | Boxplots showing the distribution of average coleoptile length of 298 genotypes of hard winter wheat association mapping panel (HWWAMP) in three experiments.

(*QCL.sdsu-2DS*), 3B (*QCL.sdsu-3BS*), 4B (*QCL.sdsu-4BS* and *QCL.sdsu-4BL*), 5B (*QCL.sdsu-5BL*), and 6B (*QCL.sdsu-6BL*), respectively (**Figure 2**). Like previous studies (Rebetzke et al., 2007a; Rebetzke et al., 2014; Li et al., 2017), we also found *Rht-B1*, a Gibberelin (GA) insensitive dwarf allele to be associated with coleoptile length. Out of 298 genotypes, 201 (67.4%) carried the dwarf allele (allele 2) and 84 (28.2%) carried the tall allele (allele 1) of *Rht-B1*. In the current study, *Rht-B1* linked SNP was highly significant with a $-\log_{10}$ (p-value) of 9.69 and explained 16.7% of the variation. The average coleoptile length of genotypes carrying allele 1 of *Rht-B1* was 13.50 mm longer than genotypes carrying allele 2. Another dwarfing gene, *Rht-D1*, was not found to be associated with coleoptile length in the current study as only 14 (4.7%) of 298 individuals carried the dwarf allele for this gene.

In total, the eight QTLs, in addition to *Rht-B1* explained 42.2% of variation in coleoptile length (**Table 1**). After *Rht-B1, QCL.sdsu-4BS* explained the highest variation (10.6%), followed by *QCL.sdsu-5BL and QCL.sdsu-2AS*, explaining 5.26% and 5.00% variation, respectively. The most significant SNPs linked to QTLs, *QCL.sdsu-2AS*, *QCL.sdsu-2BS*, *QCL.sdsu-2DS*, *QCL.sdsu-3BS*, *QCL.sdsu-4BS*, *QCL.sdsu-4BL*, *QCL.sdsu-5BL,* and *QCL.sdsu-6BL,* were *D_F1BEJMU02JILPD_53, BS00067280_51, D_contig17313_245, Tdurum_contig43252_1407, IAAV971, RAC875_rep_c82932_407, Tdurum_contig67535_391,* and *BS00065357_51*, respectively (**Table 1**). All eight QTLs identified using TASSEL v5.0 were validated using MLM (P+K model) and SUPER algorithms implemented in GAPIT to further ascertain the significance. However, the QQ plots from different algorithms revealed that MLM model has the better fit than SUPER (results not shown).

In addition, five-fold cross-validation was used to ascertain the significance of the identified SNP markers in each genomic region. After dividing the HWWAMP into five subsets, we used four sets for the marker-trait association and the remaining set of 60 accessions were used for cross-validation of significant markers. The cross-validation confirmed that six SNPs linked to QTLs, *QCL.sdsu-2AS, QCL.sdsu-2DS, QCL.sdsu-3BS, QCL.sdsu-4BS, QCL.sdsu-4BL, and QCL.sdsu-5BL,* were significantly associated with coleoptile length (Based on p-value for T-test, **Table 1**). Another QTL, *QCL.sdsu-2BS* had p-value of 0.06 from the respective t-test; thus, marginally out at 5% level of significance.

Pairwise comparison among the alleles of the significant SNPs also verified their association with coleoptile length (**Figure 3**, **Supplementary Table S5**). Positive allele (allele 1) increases the coleoptile length and its counterpart, negative allele (allele 2) decreases the coleoptile length. Allele 1 and allele 2 for each of the most significant SNP on each chromosome is given in **Supplementary Table S5**. Individually, coleoptile length difference between the allele 1 and allele 2 of the SNP on chromosomes 2A, 2B, 2D, 3B, 4BS, 4BL, 5B, and 6B was 8.62, 3.51, 7.13, 8.25, 10.70, 5.76, 10.94, and 4.56 mm, respectively. All the differences were significant at a p-value < 0.05. Overall, *QCL.sdsu-5BL* has the largest substitution effect (10.94 mm) for coleoptile length following *Rht-B1*.

Genotypes With Longer Coleoptiles

We found eight genotypes with coleoptile length longer than 100 mm, namely: 'CRIMSON', 'SCOUT66', 'GENOU', 'KIRWIN', 'KAW61', 'LONGHORN', 'MT06013', and 'AGATE' (**Table 2**, **Supplementary Table S6**). 'AGATE' had the longest coleoptile length (average 111 mm) followed by 'MT06103' (average 110.6 mm). Interestingly, 'MT06103' carried positive alleles (allele 1) for all the SNPs except *Rht-B1*. 'AGATE' was positive for all the SNPs. Significant SNP data for the other six genotypes are given in **Table 2**. From the perspective of most significant SNPs, all of the eight genotypes with the longest coleoptiles carried positive alleles for SNPs on chromosomes 2A, 2B, 4B, and 6B. On the contrary, SNP "Tdurum_cotig67535_391" on chromosome 5B was only positive in 'GENOU', 'AGATE', and 'MT06103'.

Identification of Candidate Genes and Putative Functions

To facilitate the identification of candidate genes governing coleoptile length, the chromosome regions were first delimited based on the consensus genetic map (Wang et al., 2014) and LD decay distance from our previous study (Ayana et al., 2018). Subsequently, these demarcated regions were identified by BLASTN, searching the flanking sequence of significant SNPs against CS RefSeqv1.1 (IWGSC, 2018). We then delimited the QTLs region to a 5.3, 5.9, 7, 2, 5.5, and 1.6 Mb region on chromosomes 2AS, 3BS, 4BS, 4BL, 5BL, and 6BL, respectively. Contrarily, the significant markers on chromosomes 2BS and 2DS were localized on the terminal regions of respective chromosomes, with no flanking marker available on the terminal end in the consensus genetic map (Wang et al., 2014). Therefore, the terminal regions, 6.9 and 10.3 Mb from 1bp extending up to the flanking marker on the distal end were identified as a candidate region on chromosome 2BS and 2DS, respectively. The putative genes from these regions were further

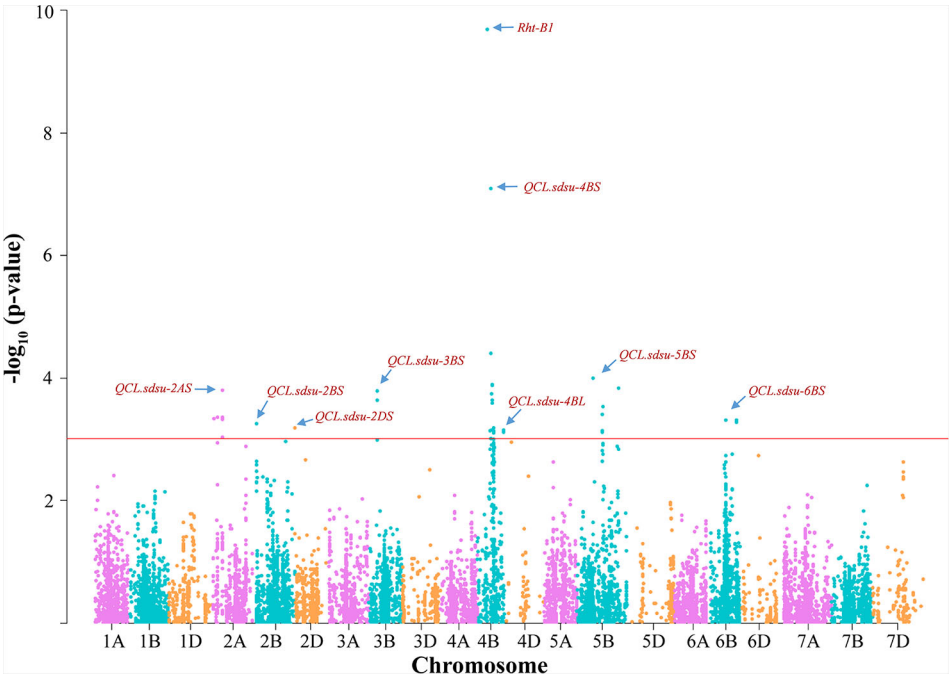


FIGURE 2 | Distribution of marker-trait associations for coleoptile length in hard winter wheat association mapping panel (HWWAMP) based on their –log(10) p-values. Manhattan plot was developed using a mixed linear model (MLM) in TASSEL v.5. The -log10 (p-values) from a genome-wide scan are plotted against particular position on each of the 21 wheat chromosomes. Horizontal line indicate genome-wide significance thresholds.

narrowed down based on the LD decay distance and proximity to the most significant SNP. Finally, we annotated the coding sequences of high confidence (HC) genes in these candidate regions using the Blast2Go (Conesa et al., 2005).

Overall, 825 high confidence genes from the eight candidate regions were annotated. Among these genes, we identified candidate genes with possible involvement in coleoptile length based on proximity to the most significant SNP and a thorough review of the literature. Accordingly, we found 28 genes predicted to encode 10 different putative proteins that can play a role in governing the coleoptile length (**Table 3**). In the 5.3 Mb region spanning *QCL.sdsu-2AS*, we found five genes that encode 1-aminocyclopropane-1-carboxylate oxidase homolog 1-like protein, which have possible involvement in coleoptile length. Another gene, *TraesCS2A02G033900*, is predicted to have a jacalin-like lectin domain, found to be a coleoptile specific lectin in barley (Grunwald et al., 2007). For QTL *QCL.sdsu-2BS*, we identified two genes encoding a cytochrome P450 87A3-like, and a probable indole-3-pyruvate monooxygenase *YUCCA5*- like proteins. Similarly, two different genes were identified in the region harboring *QCL.sdsu-2DS* encoding for the same two protein. The 2DS region also harbors four other genes predicted to encode cytochrome P450 85A1-like proteins. In these two regions (2BS and 2DS), genes encoding cytochrome P450 87A3-like and cytochrome P450 85A1-like proteins are of specific interest-based on their established role in other species. Another QTL, *QCL.sdsu-3BS* in the 5.9 Mb region of chromosome 3BS harbored 10 genes of specific interest, all

TABLE 1 | Most significant SNP markers linked to the eight QTLs for coleoptile length detected from genome-wide association analysis of 298 winter wheat genotypes.

QTL	Marker	Chromosome	Mb[a]	-log$_{10}$(p-value)	R^2 (%)	T-test[b]
QCL.sdsu-2AS	D_F1BEJMU02JILPD_53	2A	15.61	3.80	5.00	6.64E-03
QCL.sdsu-2BS	BS00067280_51	2B	6.10	3.25	4.10	6.76E-02
QCL.sdsu-2DS	D_contig17313_245	2D	93.44	3.18	4.15	1.78E-05
QCL.sdsu-3BS	Tdurum_contig43252_1407	3B	23.78	3.79	5.03	2.74E-04
QCL.sdsu-4BS	IAAV971	4B	40.75	7.10	10.56	1.15E-06
QCL.sdsu-4BL	RAC875_rep_c82932_407	4B	666.04	3.14	3.93	1.45E-03
QCL.sdsu-5BL	Tdurum_contig67535_391	5B	536.63	4.00	5.26	5.18E-02
QCL.sdsu-6BS	BS00065357_51	6B	705.75	3.31	4.19	1.25E-01
Rht-B1	*Rht-B1*	4B	30.86	9.69	16.69	–

[a]*The SNP position (Mb) is based on the CS RefSeq v1.1 (IWGSC, 2018).*
[b]P-value *obtained from the 5-fold cross validation.*

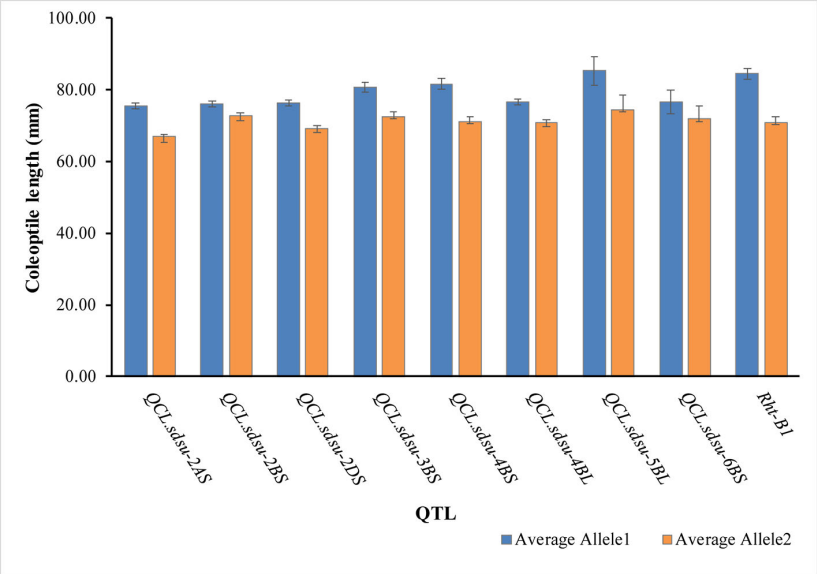


FIGURE 3 | Average coleoptile length of hard winter wheat association mapping panel genotypes corresponding to each allele of the most significant marker on the respective chromosome. Error bars are also shown at top of the bars.

predicted to encode an expansin-like protein. The fifth QTL, *QCL.sdsu-4BS* was delimited to a 7 Mb region with 65 annotated genes including two genes of interest viz. *TraesCS4B02G052000* and *TraesCS4B02G049800* putatively encoding phytochrome A-like and receptor protein kinase *TMK1*-like proteins, respectively. In the region harboring *QCL.sdsu-4BL*, a gene annotated as putative 2-oxoglutarate-dependent dioxygenase seems a likely candidate as it catalyzes several metabolic pathways in plants such as a gibberellins pathway. Most of the identified genes from the *QCL.sdsu-5BL* region were annotated as "predicted proteins", with no clear differentiation into protein families. Thus, only one gene with a likely role in coleoptile length was discovered in a 5.5 Mb region harboring this novel QTL (**Table 3**). Further, we were unable to select any candidate genes in the region harboring QTL *QCL.sdsu-6BS* based on the available literature.

DISCUSSION

Breeding Wheat for Longer Coleoptiles

Winter wheat is grown in a range of harsh environments around the globe, (Stockton et al., 1996; Bai et al., 2004) and challenges are further elevated by rising temperatures and unpredictable droughts. In conditions like hard and dry grounds (drought), and unpredicted freezing and thawing, early wheat establishment is challenged, potentially leading to lower yields (Stockton et al., 1996; Bai et al., 2004). One of the solutions to increase seedling establishment is deep sowing in order to exploit the leaching moisture regime. Coleoptile length is the limiting factor for deep planting since it affects the emergence capacity of seedlings planted deep, especially in fields with thicker stubble (No-till) and/or crusted soil surfaces (Rebetzke et al., 2014). Furthermore, around 90% of the modern semi-dwarf wheat varieties have GA-

TABLE 2 | Hard winter wheat association mapping panel (HWWAMP) genotypes with coleoptile length longer than 100 mm, along with their genotype for the most significant markers related to coleoptile length.

SNP on	2A	2B	2D	3B	4BS	4BL	5B	6B	*Rht-B1*	CL*	CSE‡
Substitution effect	8.6	3.5	7.1	8.2	10.7	5.8	10.9	4.6	13.5		
Allele 1/Allele 2	C/T	T/C	C/A	T/C	C/T	A/G	C/A	C/T	a/b		
CRIMSON	1	1	1	1	1	2	2	1	1	101.00	56.27
SCOUT66	1	1	1	1	1	1	2	1	1	101.50	62.04
GENOU	1	1	N	2	1	1	1	1	1	101.80	57.59
KIRWIN	1	1	1	1	1	1	2	1	1	103.66	62.06
KAW61	1	1	1	2	1	1	2	1	1	105.50	53.78
LONGHORN	1	1	1	1	1	1	2	1	1	106.83	62.04
MT06103	1	1	1	1	1	1	1	1	2	110.66	59.48
AGATE	1	1	1	1	1	1	1	1	1	111.00	72.98

**Coleoptile length (mm), ‡Cumulative substitution effect. '1' represents positive allele and '2' represents the negative allele.*

TABLE 3 | Annotation of candidate genes in the demarcated QTL regions identified through GWAS in hard winter wheat association mapping panel (HWWAMP).

Chr	QTL	Gene ID[a]	Start position of the gene (bp)[a]	Gene Annotation
2AS	*QCL.sdsu-2AS*	*TraesCS2A02G025800*	12,129,444	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like
		TraesCS2A02G025900	12,139,588	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like
		TraesCS2A02G026500	12,247,082	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like
		TraesCS2A02G036900	15,756,318	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like
		TraesCS2A02G037900	15,959,789	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like
		TraesCS2A02G033900	15,011,079	mannose/glucose-specific jacalin-like lectin
2BS	*QCL.sdsu-2BS*	*TraesCS2B02G009100*	5,041,094	cytochrome P450 87A3-like
		TraesCS2B02G010100	5,628,213	probable indole-3-pyruvate monooxygenase YUCCA5
2DS	*QCL.sdsu-2DS*	*TraesCS2D02G012100*	5,747,458	probable indole-3-pyruvate monooxygenase YUCCA5
		TraesCS2D02G012800	6,204,775	cytochrome P450 87A3
		TraesCS2D02G014400	7,062,903	cytochrome P450 85A1
		TraesCS2D02G014500	7,072,238	cytochrome P450 85A1
		TraesCS2D02G014600	7,085,341	cytochrome P450 85A1
		TraesCS2D02G014700	7,089,687	cytochrome P450 85A1
3BS	*QCL.sdsu-3BS*	*TraesCS3B01G051000*	25,906,973	expansin
		TraesCS3B01G051100	25,921,029	expansin
		TraesCS3B01G051200	26,043,431	expansin
		TraesCS3B01G051300	26,057,175	expansin
		TraesCS3B01G051400	26,191,126	expansin
		TraesCS3B01G051500	26,246,150	expansin
		TraesCS3B01G051600	26,301,286	expansin
		TraesCS3B01G051800	26,385,625	expansin
		TraesCS3B01G051900	26,399,446	expansin
		TraesCS3B01G052000	26,430,002	expansin
4BS	*QCL.sdsu-4BS*	*TraesCS4B02G052000*	40,780,124	phytochrome A
		TraesCS4B02G049800	38,280,457	receptor protein kinase TMK1-like
4BL	*QCL.sdsu-4BL*	*TraesCS4B02G389500*	665,956,360	putative 2-oxoglutarate-dependent dioxygenase
5BL	*QCL.sdsu-5BL*	*TraesCS5B02G356700*	536,321,998	auxin Efflux Carrier family protein isoform X1

[a]*Gene ID and physical positions are based on CS RefSeq v1.1 (IWGSC, 2018).*

insensitive dwarfing genes, which are strongly associated with shorter coleoptiles (Rebetzke et al., 1999; Li et al., 2017; Grover et al., 2018). One of the easier ways to increase coleoptile length is pyramiding of larger effect QTLs in modern-day wheat cultivars. A number of studies have shown that coleoptile length is under strong additive gene control (Rebetzke et al., 2007a; Spielmeyer et al., 2007; Yu and Bai, 2010; Li et al., 2011; Rebetzke et al., 2014; Singh et al., 2015; Li et al., 2017), thus identification of novel QTLs for increased coleoptile length would be desirable. Moreover, limited information is available in winter wheat, compelling winter wheat breeders to rely on spring wheat resources. Accordingly, we employed GWAS using 298 hard winter wheat lines in this study to develop resources for longer coleoptile length in winter wheat.

Phenotypic Evaluation for Coleoptile Length

Our results for phenotypic evaluation show that sufficient variation for coleoptile length exists in the hard winter wheat association panel, with coleoptile length ranging from 49.4 to 111 mm which overlaps with previous studies; 25 to 170 mm (Rebetzke et al., 2014) and 57 to 202 mm (Li et al., 2017). Variations among the ranges in different studies can be attributed to the diversity among the lines used and the temperature at which seedlings were grown. HWWAMP constitutes of released winter wheat cultivars and breeding lines from US winter wheat breeding programs; however, more diverse germplasm was evaluated in other studies (Rebetzke et al., 2014; Li et al., 2017). The average coleoptile length of lines from the South Dakota breeding program was highest, whereas, lines from the Michigan breeding program had the shortest coleoptile, but we did not see any significant differences among any of the breeding programs. This suggests that there is no specific focus or indirect selection for coleoptile length in any of the hard winter wheat breeding programs in the US.

Plant height has been known to be correlated with be the coleoptile length (Allan et al., 1962; Allan, 1980; Yu and Bai, 2010; Li et al., 2011; Rebetzke et al., 2016). Although we did not collect the plant height data on 298 accessions for this experiment, the HWWAMP has been evaluated for agronomic traits including plant height under the USDA-NIFA TCAP grant at several locations and the data is available in the wheat T3 database. We compared plant height at four locations to the coleoptile length of 298 accessions in this study. As expected, plant height and coleoptile length showed correlation (0.28, 0.30, 0.26, and 0.37 for four locations, respectively), but these correlations were not very high. This suggests that other factors (genomic regions) in addition to plant height QTLs identified in this study affect the coleoptile length.

QTLs for Coleoptile Length

In the present study, MLM based genome wide associations identified eight QTLs associated with coleoptile length on seven different chromosomes. The identified QTLs were validated using five-fold cross-validation (Ramakrishnan et al., 2019). This approach validated six of the eight identified QTLs,

namely *QCL.sdsu-2AS, QCL.sdsu-2DS, QCL.sdsu-3BS, QCL.sdsu-4BS, QCL.sdsu-4BL, and QCL.sdsu-5BL* (**Table 1**). Another QTL, *QCL.sdsu-2BS* and *QCL.sdsu-6BL* were not validated using the five-fold approach. These could be potential associations affecting coleoptile length and need further validation.

We compared the findings of this study by fetching the physical location of previously reported QTLs from several coleoptile length mapping studies (Rebetzke et al., 2007a; Rebetzke et al., 2014; Singh et al., 2015; Li et al., 2017) (**Figure 4**). As a result, we identified three novel QTLs, namely, *QCL.sdsu-2AS*, *QCL.sdsu-4BL*, and *QCL.sdsu-5BL* and four QTLs that are in the proximity to previously mapped QTLs (**Figure 4**). Among the novel QTLs, *QCL.sdsu-5BL* explains largest variation (R^2 = 5.26%) followed by *QCL.sdsu-2AS* (R^2 = 5.00%). Furthermore, the pairwise comparison among the alleles of the significant SNPs revealed that *QCL.sdsu-5BL* has the largest substitution effect after *Rht-B1*. Therefore, *QCL.sdsu-5BL* is a valuable novel QTL which could be used to compensate for negative effect of *Rht-B1* locus on coleoptile length.

Two QTLs namely *QCL.sdsu-2DS* and *QCL.sdsu-3BS*, previously mapped using Simple sequence repeats (SSR) markers (Rebetzke et al., 2007b; Singh et al., 2015) were also validated using SNPs in this study. The newer positions of these two QTLs are likely more accurate as highly saturated SNP markers were used in the current study compared to less dense SSR markers used in the previous studies. Different studies (Rebetzke et al., 2014; Li et al., 2017) have reported a QTL for coleoptile length on chromosome 4BS. In this study, we identified a QTL (*QCL.sdsu-4BS*) in the same region, which is around 10 Mb apart from the *Rht-B1* gene (IWGSC, 2018). Based on the estimated LD (r^2 = 0.54) between the *Rht-B1* and *QCL.sdsu-4BS*, these two could be different regions or *QCL.sdsu-4BS* could likely represent *Rht-B1*. Further investigation is needed to validate the independence of these regions.

Out of the nine significant associations (including *Rht-B1*) found in the current study, seven are mapped to the B genome. Furthermore, among the total unique QTLs mapped for coleoptile length so far (including this study), 57% QTLs are mapped on the B genome, 26% QTLs are mapped on the D genome and 17% QTLs are mapped on the A genome. Thus, it seems that B genome comparatively may have more genes controlling the coleoptile length. It would be interesting to study the variation among the diploid progenitors of wheat for coleoptile length.

Pyramiding of favorable QTLs can be successfully used for developing varieties with longer coleoptile (Li et al., 2017). In agreement with the previous studies (Rebetzke et al., 2014; Li et al., 2017), we observed an additive effect for coleoptile length among the identified QTLs in the current study. The stacking of positive alleles at different loci increased coleoptile length in additive fashion (**Figure 5**). A cultivar 'AGATE' has all the positive alleles for associated SNPs and has the longest coleoptile length. We also compared the allelic composition of three cultivars having shortest coleoptile length. These three

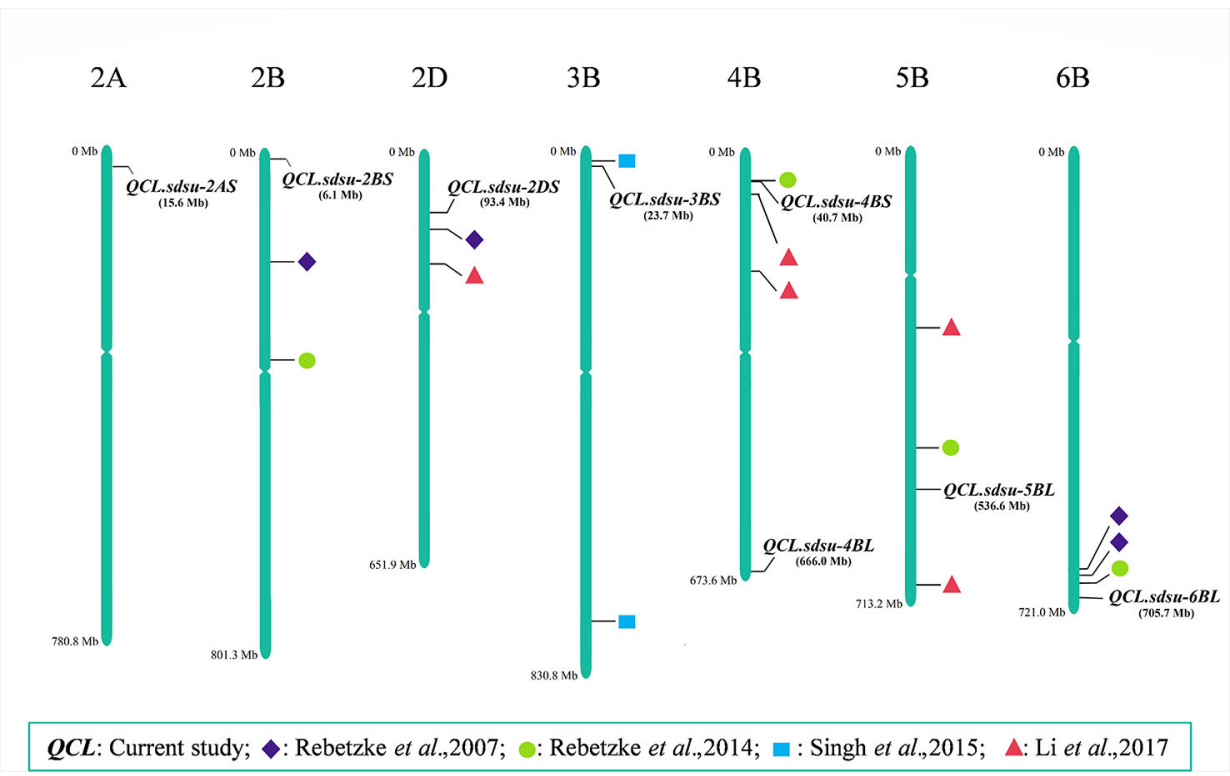


FIGURE 4 | Chromosomal positions of QTLs associated with coleoptile length identified in the current study and their comparative analysis with the previous studies. The scale represents physical distance based on Chinese Spring RefSeq 1.1 (IWGSC, 2018).

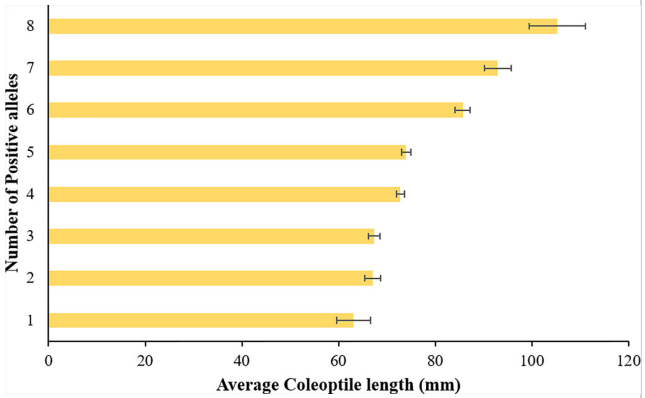


FIGURE 5 | Average coleoptile length corresponding to each stack of positive alleles in hard winter wheat association mapping panel (HWWAMP).

cultivars namely 'GARRISON', 'OK5723W', and 'OK04505' have negative alleles (allele 2) at six, five, and four associated SNPs, respectively. In addition, all three cultivars have the dwarfing allele for *Rht-B1*. Though, it will be desirable to keep the negative allele of *Rht-B1* so that the stature/height of cultivars remains semi-dwarf. We identified a breeding line 'MT06103' which has the positive alleles at all loci except for the *Rht-B1*. MT06103 has coleoptile length very close to 'AGATE' (**Table 2**). While studying the seedling emergence in fall sown wheat, Allan et al., 1962 also found a selection (14 X 50-3 B-4), which was moderately short in plant height but was ranked towards top with respect to coleoptile length. Thus, it is evident that coleoptile length can be improved while maintaining short stature of plant. Thus, such genotypes which already have all the favorable alleles can directly be exploited in winter wheat breeding programs to improve the coleoptile length of the new cultivars.

In silico Gene Annotation of the Candidate Regions

After a thorough examination of the available literature and proximity to the most significant SNPs, we identified 27 genes predicted to have a role that could likely affect coleoptile elongation (**Table 3**). We found genes with diverse functions, including phytohormone biosynthesis-related, cytochrome P450 family genes, expansins, etc. that are probable candidates. Further, it is expected that the genes common to many QTL regions are more likely to play a role in determining the length of coleoptile.

Phytohormones are the signaling molecules, which play a crucial role in the development and physiological processes in plants (Rudnicka et al., 2019). Specifically, auxins are a major group of phytohormones, which affect coleoptile length in grass species by inducing cell elongation either directly (Vanneste and Friml, 2009; Paque and Weijers, 2016), or by interacting with other plant hormones such as ethylene (Woodward and Bartel, 2005). Two genes from different candidate regions on chromosomes 2BS and 2DS were predicted as indole-3-pyruvate monooxygenase *YUCCA5* protein, which catalyzes the biosynthesis of indole-acetic acid (IAA), the most commonly occurring natural auxin, from tryptophan (Won et al., 2011). We also found a *PIN* protein (a component of auxin-efflux carrier family) in the *QCL.sdsu-5BL* region. The *PIN* proteins are known to play role in auxin transport and expressed in several plant tissues, affecting plant growth (Zhou and Luo, 2018). Whereas, another putative *ACO1*-like protein was found in the 2AS candidate region. *ACO1*-like protein is a part of the ethylene biosynthetic pathway and is speculated to affect rice coleoptile elongation in stress conditions (Hsu and Tung, 2017).

Brassinosteroids (BRs) play an important role in cell elongation and proliferation (Nakaya et al., 2002), and thus in determining plant height. A BR-deficient (*brd*) mutant was used to characterize *OsDWARF* gene in rice, an orthologue of the tomato *DWARF* gene and *CYP85A1* or *BR6OX1* in Arabidopsis (Shimada et al., 2001; Shimada et al., 2003) and found to affect polar elongation of stem cells (Hong et al., 2002). Another cytochrome P450 superfamily protein *CYP87A3* has been characterized in rice as an auxin-induced gene specifically expressed in coleoptiles (Chaban et al., 2003). In our study, we found putative cytochrome P450 85A1-like and cytochrome P450 87A3 proteins spanning the QTLs, *QCL.sdsu-2BS*, and *QCL.sdsu-2DS* which may affect coleoptile length in wheat. Additionally we found 10 genes all encoding putative expansin proteins in the genomic region spanning *QCL.sdsu-3BS*. Our finding corroborates with Singh et al. (2015) who also reported the presence of expansin like genes in this region while mapping coleoptile length in a biparental mapping population. Expansins have been reported to affect cell growth and elongation (Marowa et al., 2016); and express in wheat coleoptiles and correlate with the coleoptile growth (Gao et al., 2007; Gao et al., 2008). The cytochrome P450 superfamily genes and expansins are thus strong candidates for coleoptile length and need further investigation in wheat.

Further, phytochrome A (*PHY A*) protein identified in the *QCL.sdsu-4BS* candidate region is of specific importance with respect to coleoptile length. In rice, phytochrome A gene is well known to affect coleoptile elongation, plant height, and internode elongation either directly or by affecting jasmonate signaling genes (Garg et al., 2006; Riemann et al., 2008). Apart from these genes, we also found jacalin-like lectin, related to Horcolin protein specifically expressed in barley coleoptiles (Grunwald et al., 2007) and putative 2-oxoglutarate-dependent dioxygenase (**Table 3**), related to a versatile enzyme family catalyzing biosynthesis and catabolism of auxins and gibberellins (Farrow and Facchini, 2014).

CONCLUSION

Coleoptile length is regularly evaluated in advanced breeding lines in several breeding programs. However, due to limited knowledge about the underlying QTLs and linked molecular markers, breeding for coleoptile length becomes challenging. Characterization of eight QTLs associated with coleoptile length in winter wheat and identification of tightly linked SNPs could be

a valuable resource for wheat breeders. The critical SNPs identified in our study could be used to develop breeder friendly kompetitive allele-specific PCR (KASP) assays (**Supplementary Table S7**) for marker-assisted selection (Rasheed et al., 2016; Gill et al., 2019). Marker-assisted stacking of these QTLs would result in the development of wheat varieties with longer coleoptile. Also, these QTLs can be effectively combined with previously reported QTLs to breed for desired coleoptile length in wheat. In addition, these markers could be weighted and incorporated into the genomic selection strategy. Further functional genomic studies are crucial to validate the effect of the identified candidate genes on coleoptile length.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**. Additional physiological and agronomic data about the HWWAMP accessions is available in the T3/Wheat database (https://triticeaetoolbox.org/wheat/pedigree/pedigree_info.php).

AUTHOR CONTRIBUTIONS

JS and SS conceptualized the experiment and designed the methodology. JS, DS, YQ, JH, RT performed the investigation. JS, DS, HG, and NB performed the data analysis. HG, JS, DS, NB, and SS wrote the original manuscript. JH, YQ, RT, and BT contributed to the interpretation of results and revision of the manuscript. All authors approved the manuscript.

FUNDING

This project was collectively funded by the USDA hatch projects SD00H538-15 and SD00H695-20 and the Agriculture and Food Research Initiative Competitive Grants 2011-68002-30029 (Triticeae-CAP), 2017-67007-25939 (Wheat-CAP), and 2019-67013-29015 from the USDA National Institute of Food and Agriculture and South Dakota Wheat Commission grant 3X9267. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the South Dakota Agriculture Experimental Station (Brookings, SD, USA) for providing the resources to conduct the experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01345/full#supplementary-material

REFERENCES

Allan, R. E., Vogel, O. A., and Peterson, C. J. (1962). Seedling Emergence Rate of Fall-sown Wheat and Its Association with Plant Height and Coleoptile Length 1. *Agron. J.* 54 (4), 347–350. doi: 10.2134/agronj1962.00021962005400040022x

Allan, R. E. (1980). Influence of Semi dwarfism and Genetic Background on Stand Establishment of Wheat 1. *Crop Sci.* 20 (5), 634–638. doi: 10.2135/cropsci1980.0011183X002000050022x

Ayana, G. T., Ali, S., Sidhu, J. S., Gonzalez Hernandez, J. L., Turnipseed, B., and Sehgal, S. K. (2018). Genome-wide association study for spot blotch resistance in hard winter wheat. *Front. Plant Sci.* 9, 1–15. doi: 10.3389/fpls.2018.00926

Bai, G., Das, M. K., Carver, B. F., Xu, X., and Krenzer, E. G. (2004). Covariation for Microsatellite Marker Alleles Associated with 8 and Coleoptile Length in Winter Wheat. *Crop Sci.* 44, 1187. doi: 10.2135/cropsci2004.1187

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brown, P. R., Singleton, G. R., Tann, C. R., and Mock, I. (2003). Increasing sowing depth to reduce mouse damage to winter crops. *Crop Prot.* 22, 653–660. doi: 10.1016/S0261-2194(03)00006-1

Budak, N., Baenziger, P. S., Eskridge, K. M., Baltensperger, D., and Moreno-Sevilla, B. (1995). Plant Height Response of Semidwarf and Nonsemidwarf Wheats to the Environment. *Crop Sci.* 35, 447. doi: 10.2135/cropsci1995.0011183X003500020028x

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci.* 110 (20), 8057–8062. doi: 10.1073/pnas.1217133110

Chaban, C., Waller, F., Furuya, M., and Nick, P. (2003). Auxin responsiveness of a novel cytochrome p450 in rice coleoptiles. *Plant Physiol.* 133, 2000–2009. doi: 10.1104/pp.103.022202

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Endelman, J. B., and Jannink, J.-L. (2012). Shrinkage Estimation of the Realized Relationship Matrix. *Genes|Genomes|Genetics* 2, 1405–1413. doi: 10.1534/g3.112.004259

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Farrow, S. C., and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* 5, 524. doi: 10.3389/fpls.2014.00524

Gao, Q., Guo, Q. F., Xing, S. C., Zhao, M. R., Li, F., and Wang, W. (2007). The characteristics of expansins in wheat coleoptiles and their responses to water stress. *J. Plant Physiol. Mol. Biol.* 33, 402–410.

Gao, Q., Zhao, M., Li, F., Guo, Q., Xing, S., and Wang, W. (2008). Expansins and coleoptile elongation in wheat. *Protoplasma* 233, 73–81. doi: 10.1007/s00709-008-0303-1

Garg, A. K., Sawers, R. J. H., Wang, H., Kim, J.-K., Walker, J. M., Brutnell, T. P., et al. (2006). Light-regulated overexpression of an Arabidopsis phytochrome A gene in rice alters plant architecture and increases grain yield. *Planta* 223, 627–636. doi: 10.1007/s00425-005-0101-3

Gill, H. S., Li, C., Sidhu, J. S., Liu, W., Wilson, D., Bai, G., et al. (2019). Fine Mapping of the Wheat Leaf Rust Resistance Gene Lr42. *Int. J. Mol. Sci.* 20, 1–12. doi: 10.3390/ijms20102445

Grover, G., Sharma, A., Gill, H. S., Srivastava, P., and Bains, N. S. (2018). Rht8 gene as an alternate dwarfing gene in elite Indian spring wheat cultivars. *PloS One* 13, e0199330. doi: 10.1371/journal.pone.0199330

Grunwald, I., Heinig, I., Thole, H. H., Neumann, D., Kahmann, U., Kloppstech, K., et al. (2007). Purification and characterisation of a jacalin-related, coleoptile specific lectin from Hordeum vulgare. *Planta* 226, 225–234. doi: 10.1007/s00425-006-0467-x

Guttieri, M. J., Baenziger, P. S., Frels, K., Carver, B., Arnall, B., and Waters, B. M. (2015). Variation for grain mineral concentration in a diversity panel of current and historical Great Plains hard winter wheat germplasm. *Crop Sci.* 55, 1035–1052. doi: 10.2135/cropsci2014.07.0506

Hong, Z., Ueguchi-Tanaka, M., Shimizu-Sato, S., Inukai, Y., Fujioka, S., Shimada, Y., et al. (2002). Loss-of-function of a rice brassinosteroid biosynthetic enzyme, C-6 oxidase, prevents the organized arrangement and polar elongation of cells in the leaves and stem. *Plant J.* 32, 495–508. doi: 10.1046/j.1365-313X.2002.01438.x

Hsu, S.-K., and Tung, C.-W. (2017). RNA-Seq Analysis of Diverse Rice Genotypes to Identify the Genes Controlling Coleoptile Growth during Submerged Germination. *Front. Plant Sci.* 8, 762. doi: 10.3389/fpls.2017.00762

Huang, X., and Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Hunt, J. R., Hayman, P. T., Richards, R. A., and Passioura, J. B. (2018). Opportunities to reduce heat damage in rain-fed wheat crops based on plant breeding and agronomic management. *Field Crop Res.* 224, 126–138. doi: 10.1016/J.FCR.2018.05.012

International Wheat Genome Sequencing Consortium (IWGSC), T. I. W. G. S. C., IWGSC RefSeq principal investigators: I. R. principal, Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191. doi: 10.1126/science.aar7191

Jame, Y. W., and Cutforth, H. W. (2004). Simulating the effects of temperature and seeding depth on germination and emergence of spring wheat. *Agric. For. Meteorol.* 124, 207–218. doi: 10.1016/J.AGRFORMET.2004.01.012

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9, 29. doi: 10.1186/1746-4811-9-29

Li, P., Chen, J., Wu, P., Zhang, J., Chu, C., See, D., et al. (2011). Quantitative trait loci analysis for the effect of Rht-B1 dwarfing gene on coleoptile length and seedling root length and number of bread wheat. *Crop Sci.* 51, 2561–2568. doi: 10.2135/cropsci2011.03.0116

Li, G., Bai, G., Carver, B. F., Elliott, N. C., Bennett, R. S., Wu, Y., et al. (2017). Genome-wide association study reveals genetic architecture of coleoptile length in wheat. *Theor. Appl. Genet.* 130, 391–401. doi: 10.1007/s00122-016-2820-1

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Mahdi, L., Bell, C., and Ryan, J. (1998). Establishment and yield of wheat (Triticum turgidum L.) after early sowing at various depths in a semi-arid Mediterranean environment. *F. Crop Res.* 58, 187–196. doi: 10.1016/S0378-4290(98)00094-X

Marowa, P., Ding, A., and Kong, Y. (2016). Expansins: roles in plant growth and potential applications in crop improvement. *Plant Cell Rep.* 35 (5), 949–965. doi: 10.1007/s00299-016-1948-4

Nakaya, M., Tsukaya, H., Murakami, N., and Kato, M. (2002). Brassinosteroids Control the Proliferation of Leaf Cells of Arabidopsis thaliana. *Plant Cell Physiol.* 43, 239–244. doi: 10.1093/pcp/pcf024

Paque, S., and Weijers, D. (2016). Q&A: Auxin: the plant molecule that influences almost anything. *BMC Biol.* 14 (1), 67. doi: 10.1186/s12915-016-0291-0

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS One* 7, 1–8. doi: 10.1371/journal.pone.0032253

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959.

Ramakrishnan, S. M., Sidhu, J. S., and Ali, S. (2019). Molecular characterization of bacterial leaf streak resistance in hard winter wheat. *PeerJ* 7:e7276, 1–24. doi: 10.7717/peerj.7276

Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., et al. (2016). Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. *Theor. Appl. Genet.* 129 (10), 1843–1860. doi: 10.1007/s00122-016-2743-x

Rebetzke, G. J., Richards, R. A., Fischer, V. M., and Mickelson, B. J. (1999). Breeding long coleoptile, reduced height wheats. *Euphytica* 106, 159–168. doi: 10.1023/A:1003518920119

Rebetzke, G. J., Bruce, S. E., and Kirkegaard, J. A. (2005). Longer coleoptiles improve emergence through crop residues to increase seedling number and biomass in wheat (Triticum aestivum L.). *Plant Soil* 272, 87–100. doi: 10.1007/s11104-004-4040-8

Rebetzke, G. J., Ellis, M. H., Bonnett, D. G., and Richards, R. A. (2007a). Molecular mapping of genes for Coleoptile growth in bread wheat (Triticum aestivum L.). *Theor. Appl. Genet.* 114, 1173–1183. doi: 10.1007/s00122-007-0509-1

Rebetzke, G. J., Richards, R. A., Fettell, N. A., Long, M., Condon, A. G., Forrester, R. I., et al. (2007b). Genotypic increases in coleoptile length improves stand establishment, vigour and grain yield of deep-sown wheat. *F. Crop Res.* 100, 10–23. doi: 10.1016/j.fcr.2006.05.001

Rebetzke, G. J., Verbyla, A. P., Verbyla, K. L., Morell, M. K., and Cavanagh, C. R. (2014). Use of a large multiparent wheat mapping population in genomic dissection of coleoptile and seedling growth. *Plant Biotechnol. J.* 12, 219–230. doi: 10.1111/pbi.12130

Rebetzke, G. J., Zheng, B., and Chapman, S. C. (2016). Do wheat breeders have suitable genetic variation to overcome short coleoptiles and poor establishment in the warmer soils of future climates? *Funct. Plant Biol.* 43, 961. doi: 10.1071/FP15362

Riemann, M., Riemann, M., and Takano, M. (2008). Rice JASMONATE RESISTANT 1 is involved in phytochrome and jasmonate signalling. *Plant Cell Environ.* 31, 783–792. doi: 10.1111/j.1365-3040.2008.01790.x

Rudnicka, M., Ludynia, M., and Karcz, W. (2019). Effects of Naphthazarin (DHNQ) Combined with Lawsone (NQ-2-OH) or 1,4-Naphthoquinone (NQ) on the Auxin-Induced Growth of Zea mays L. Coleoptile Segments. *Int. J. Mol. Sci.* 20, 1788. doi: 10.3390/ijms20071788

Schillinger, W. F., Donaldson, E., Allan, R. E., and Jones, S. S. (1998). Winter Wheat Seedling Emergence from Deep Sowing Depths. *Agron. J.* 90, 582. doi: 10.2134/agronj1998.00021962009000050002x

Shimada, Y., Fujioka, S., Miyauchi, N., Kushiro, M., Takatsuto, S., Nomura, T., et al. (2001). Brassinosteroid-6-oxidases from Arabidopsis and tomato catalyze multiple C-6 oxidations in brassinosteroid biosynthesis. *Plant Physiol.* 126 (2), 770–779. doi: 10.1104/pp.126.2.770

Shimada, Y., Goda, H., Nakamura, A., Takatsuto, S., Fujioka, S., and Yoshida, S. (2003). Organ-Specific Expression of Brassinosteroid-Biosynthetic Genes and Distribution of Endogenous Brassinosteroids in Arabidopsis. *Plant Physiol.* 131, 287–297. doi: 10.1104/pp.013029

Sidhu, J. S., Ramakrishnan, S. M., Ali, S., Bernardo, A., Bai, G., Abdullah, S., et al. (2019). Assessing the genetic diversity and characterizing genomic regions conferring Tan Spot resistance in cultivated rye. *PloS One* 14, e0214519. doi: 10.1371/JOURNAL.PONE.0214519

Singh, K., Shukla, S., Kadam, S., Semwal, V. K., Singh, N. K., and Khanna-Chopra, R. (2015). Genomic regions and underlying candidate genes associated with coleoptile length under deep sowing conditions in a wheat RIL population. *J. Plant Biochem. Biotechnol.* 24, 324–330. doi: 10.1007/s13562-014-0277-3

Spielmeyer, W., Hyles, J., Joaquim, P., Azanza, F., Bonnett, D., Ellis, M. E., et al. (2007). A QTL on chromosome 6A in bread wheat (Triticum aestivum) is associated with longer coleoptiles, greater seedling vigour and final plant height. *Theor. Appl. Genet.* 115, 59–66. doi: 10.1007/s00122-007-0540-2

Stockton, R. D., Krenzer, E. G.Jr., Solie, J., and Payton, M. E. (1996). Stand establishment of winter wheat in Oklahoma: a survey. *J. Prod. Agric.* 9 (4), 571–575.

Team, R. C. (2016). *R: A language and environment for statistical computing [Computer software manual]* (Austria: Vienna).

Tuberosa, R., Salvi, S., Sanguineti, M. C., Landi, P., Maccaferri, M., and Conti, S. (2002). Mapping QTLS regulating morpho-physiological traits and yield: Case studies, shortcomings and perspectives in drought-stressed maize. *Ann. Bot.* 89, 941–963. doi: 10.1093/aob/mcf134

Vanneste, S., and Friml, J. (2009). Auxin: a trigger for change in plant development. *Cell* 136 (6), 1005–1016. doi: 10.1016/j.cell.2009.03.001

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Won, C., Shen, X., Mashiguchi, K., Zheng, Z., Dai, X., Cheng, Y., et al. (2011). Conversion of tryptophan to indole-3-acetic acid by TRYPTOPHAN AMINOTRANSFERASES OF ARABIDOPSIS and YUCCAs in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 108, 18518–18523. doi: 10.1073/pnas.1108436108

Woodward, A. W., and Bartel, B. (2005). A receptor for auxin. *Plant Cell* 17, 2425–2429. doi: 10.1105/tpc.105.036236

Wu, J. (2014). *Minique: An R package for linear mixed model analyses* (Vienna: R Found. Stat. Comput).

Yu, J., and Bai, G. H. (2010). Mapping quantitative trait loci for long coleoptile in Chinese wheat landrace Wangshuibai. *Crop Sci.* 50, 43–50. doi: 10.2135/cropsci2009.02.0065

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhou, J. J., and Luo, J. (2018). The PIN-FORMED auxin efflux carriers in plants. *Int. J. Mol. Sci.* 19 (9), 2759. doi: 10.3390/ijms19092759

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sidhu, Singh, Gill, Brar, Qiu, Halder, Al Tameemi, Turnipseed and Sehgal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 February 2020
doi: 10.3389/fgene.2020.00028

In Situ Genetic Evaluation of European Larch Across Climatic Regions Using Marker-Based Pedigree Reconstruction

Milan Lstibůrek[1], Silvio Schueler[2], Yousry A. El-Kassaby[3], Gary R. Hodge[4], Jan Stejskal[1], Jiří Korecký[1], Petr Škorpík[5], Heino Konrad[5] and Thomas Geburek[5]*

[1] Faculty of Forestry and Wood Sciences, Czech University of Life Sciences, Praha, Czechia, [2] Department of Forest Growth and Silviculture, Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Wien, Austria, [3] Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, Vancouver, BC, Canada, [4] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, United States, [5] Department of Forest Genetics, Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Wien, Austria

OPEN ACCESS

Edited by:
Charles Chen,
Oklahoma State University,
United States

Reviewed by:
Joana Isabel Robalo,
University Institute of Psychological, Social and Life Sciences, Portugal
Richard Buggs,
Queen Mary University of London, United Kingdom

***Correspondence:**
Milan Lstibůrek
lstiburek@fld.czu.cz

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 18 September 2019
Accepted: 08 January 2020
Published: 13 February 2020

Citation:
Lstibůrek M, Schueler S, El-Kassaby YA, Hodge GR, Stejskal J, Korecký J, Škorpík P, Konrad H and Geburek T (2020) In Situ Genetic Evaluation of European Larch Across Climatic Regions Using Marker-Based Pedigree Reconstruction. Front. Genet. 11:28. doi: 10.3389/fgene.2020.00028

Sustainable and efficient forestry in a rapidly changing climate is a daunting task. The sessile nature of trees makes adaptation to climate change challenging; thereby, ecological services and economic potential are under risk. Current long-term and costly gene resources management practices have been primarily directed at a few economically important species and are confined to defined ecological boundaries. Here, we present a novel *in situ* gene-resource management approach that conserves forest biodiversity and improves productivity and adaptation through utilizing basic forest regeneration installations located across a wide range of environments without reliance on structured tree breeding/conservation methods. We utilized 4,267 25- to 35-year-old European larch trees growing in 21 reforestation installations across four distinct climatic regions in Austria. With the aid of marker-based pedigree reconstruction, we applied multi-trait, multi-site quantitative genetic analyses that enabled the identification of broadly adapted and productive individuals. Height and wood density, proxies to fitness and productivity, yielded *in situ* heritability estimates of 0.23 ± 0.07 and 0.30 ± 0.07, values similar to those from traditional "structured" pedigrees methods. In addition, individual trees selected with this approach are expected to yield genetic response of 1.1 and 0.7 standard deviations for fitness and productivity attributes, respectively, and be broadly adapted to a range of climatic conditions. Genetic evaluation across broad climatic gradients permitted the delineation of suitable reforestation areas under current and future climates. This simple and resource-efficient management of gene resources is applicable to most tree species.

Keywords: genetic evaluation, pedigree reconstruction, sustainable forestry, European larch, genetic gain, forest tree breeding

INTRODUCTION

The composition, function, and service of terrestrial ecosystems are increasingly threatened by the steady global warming trend (Walther et al., 2002; Walther, 2010; Hanewinkel et al., 2013). Plants' immediate response to climate change is manifested in altered phenology (Wolkovich et al., 2012), increased growth (Pretzsch et al., 2014), and mortality (Allen et al., 2010). Assisted gene flow has been considered as a viable option for dealing with the mismatch between environmental alterations caused by climate change and the migration pace of plant populations (McLachlan et al., 2007; Kremer et al., 2012). However, assisted gene flow has not been thoroughly tested as genotypes are transferred to novel environments with altered thermal (Vitt et al., 2010), photoperiod (Saikkonen et al., 2012; Frascaria-Lacoste and Fernández-Manjarrés, 2012), and edaphic conditions (Kranabetter et al., 2012). Furthermore, epigenetic after-effects associated with plants transfer (Holeski et al., 2012; Bräutigam et al., 2013) and phenotypic plasticity (Alberto et al., 2013) have been discounted. These factors, collectively, provide sound reasons to explore alternative forest tree gene management approaches.

Forest tree gene resource management, with concurrent selective breeding and gene conservation, are long-term endeavors involving hundreds of parents and thousands of offspring tested at multiple locations, requiring substantial resources, elaborate logistics, and sustained organizational commitment, and more importantly, are predominantly encapsulated within specific ecological boundaries known as breeding zones (White et al., 2007). These extensive programs often follow the recurrent selection scheme with repeated rounds of breeding, testing, and selection, resulting in cumulative improvement (genetic response to selection) delivered through specialized seed production populations known as seed orchards. In conventional selective breeding programs, controlled pollinations following specific mating designs produce structured pedigrees (White et al., 2007), which are evaluated in replicated test sites within defined ecological boundaries (Hanewinkel et al., 2013), a prerequisite for effective genetic evaluation and selection. These considerable efforts are restricted to few economically important species, thus facilitating widespread cultivation of few species, with potential adverse effects on tree species diversity and ecological services provision (Isbell et al., 2015; Hua et al., 2016). Moreover, reforestation with orchard-produced seedlings is restricted to their respective ecological boundaries; thus, these programs can be spatially static and slow in responding to environmental contingencies or market demands.

European larch, an economically important deciduous conifer, is distributed in Central Europe. It is native to the Alps and the Carpathian Mountains, with smaller disjunct populations in northeastern Europe. This shade-intolerant species is primarily planted within mixed forests due to its high ecological value and excellent wood characteristics. Despite its wide planting outside of its native range, the gene resource management effort for the species is mainly focused on seed provision and gene conservation (Pâques et al., 2013). The species occurs naturally across a discontinuous range in the Alps, Sudetes, and Carpathians, as well as in Polish lowlands. This shade-intolerant species shows a subcontinental climate preference and high site tolerance. Due to its high resistance and durability, larch wood is a traditional material for building and roof construction in the Alpine area, with increasing importance in modern architecture and furniture design. As *L. decidua* has the finest wood characteristics among temperate European conifers, it has been widely planted throughout the continent in artificial plantations, thus facilitating translocations of genetic materials for more than three centuries (Jansen and Geburek, 2016). Under climate change conditions, the species appeared to be highly vulnerable to drought events (Allen et al., 2010). *L. decidua* shows high levels of genetic variation including drought sensitivity across the species range (George et al., 2017). Thus, the species regional improvement activities are focused on conserving its genetic diversity and utilizing local sources for increasing the species adaptation to climate change.

At the northeastern fringe of the Alps, European larch improvement activities are conducted within a spatially and climatically heterogeneous landscape. This region reaches from lowland areas around the river Danube (∼200 m a.s.l.) through the hilly landscape of the alpine foreland up to mountains of 900 m in the northern calcareous Alps. Present climate conditions are represented by four climatic zones: 1) pannonical continental climate with hot summers and frequent droughts in low elevations of the East and Northeast, 2) temperate Atlantic climate with warm temperatures and frequent precipitations in the western Alpine foreland, 3) temperate climate with continental influence at low elevations of the Eastern Alps, and 4) harsh mountain conditions at higher elevation with lower temperatures and low winter temperatures. Global warming in the Alpine region has already resulted in a significant 2)∘C temperature increase since 1880, about twice as high as the global average (Allen et al., 2010).

Utilizing the European larch breeding program in Austria, we investigated a feasible alternative that would efficiently address the global climate change issues and overcome the major limitations of assisted gene flow, and deliver substantial production and social benefits to the human society. We analyzed 25- to 35-year-old larch progenies originating from open pollination in a common parental source (a seed orchard) and growing in 21 reforestation installations (typical forest stands). Utilizing the "Breeding-without-Breeding" methodology with phenotypic preselection (El-Kassaby and Lstibůrek, 2009; Lstibůrek et al., 2015), we reconstructed the parentage of individual progenies and estimated heritabilities for height and wood density yielding similar values to those from typical full-sib forest genetic trials. Genetic evaluation across broad climatic gradients permitted the delineation of suitable reforestation areas under current and future climates. Following the evaluation, the second-generation seed orchard was established from the top-ranking selections.

MATERIALS AND METHODS

Source Population and Climate Data

The seed orchard [Nat. Reg. No. Lä P3 (4.2/sm-tm)] for which we aimed to conduct accelerated gene-resource management is located at an altitude of 520 m a.s.l., and its seed material is

considered to be the most valuable larch seeds for the mountainous areas of the northern alpine foreland. The orchard was established in 1954 over 3.15 ha, with 1,666 vegetative propagules of 53 phenotypically selected parent trees. Since its establishment, the main objective of the Austrian larch seed orchard program was to secure seed supply with minimal genetic testing; thus, controlled pollinations and progeny tests were not conducted.

Next, we identified 21 reforestation installations (sites) within comparable tree ages (25 to 37 years), sufficient size (at least 200 remaining trees in more or less regular planting designs), low level of environmental variation within the site, and composition in which larch is the single or dominant tree species. These 21 reforestation installations all originated from mixed seedlots harvested from the seed orchard, are located at altitudes between 250 and 800 m, and span through a geographical space of about 170 km W-E and 110 km N-S.

Climate data for the 21 reforestation installations with putative seed deployment areas under the present and future climate conditions were obtained from the locally downscaled high-resolution WorldClim models (Hijmans et al., 2005); WorldClim has a spatial resolution of 30 arc-seconds. For an unbiased climate comparison, we obtained all monthly climate parameters (average monthly mean, minimum and maximum temperatures, average monthly precipitation) as well as various derived bioclimatic variables (Meier et al., 2012), and the impact of these climate descriptors, as well as their inter-correlations, was distilled through principal component analysis (PCA) in which the climate of the reforestation sites was used as active cases. Potential/future seed deployment areas in Austria were identified by including gridded climate data as inactive cases (Meier et al., 2012). As these deployment areas should cover the broad range of the four climatic groups, we used the maximum and minimum of the 21 plantations sites within the first two principal components to delimit the Austrian landscape. For prediction of the future climate we used the Max Planck Institute Earth System model (MPI-ESM-LR) under the Rcp45 scenario (Giorgetta et al., 2013), for the period 2041–2060 (**Figure 1**).

Phenotyping and Genotyping

Phenotyping was conducted on the 21 reforestation sites after excluding individuals with damage and/or bad form. In total, 4,267 trees were measured for height [m] and scored for wood density using the pilodyn penetration method (Cown, 1978). Individual tree position was determined by triangulation. First, a compact block of trees was identified within each reforestation site (denoted as random selections) (Lstibůrek et al., 2015). Second, within each site and based on height measurements,

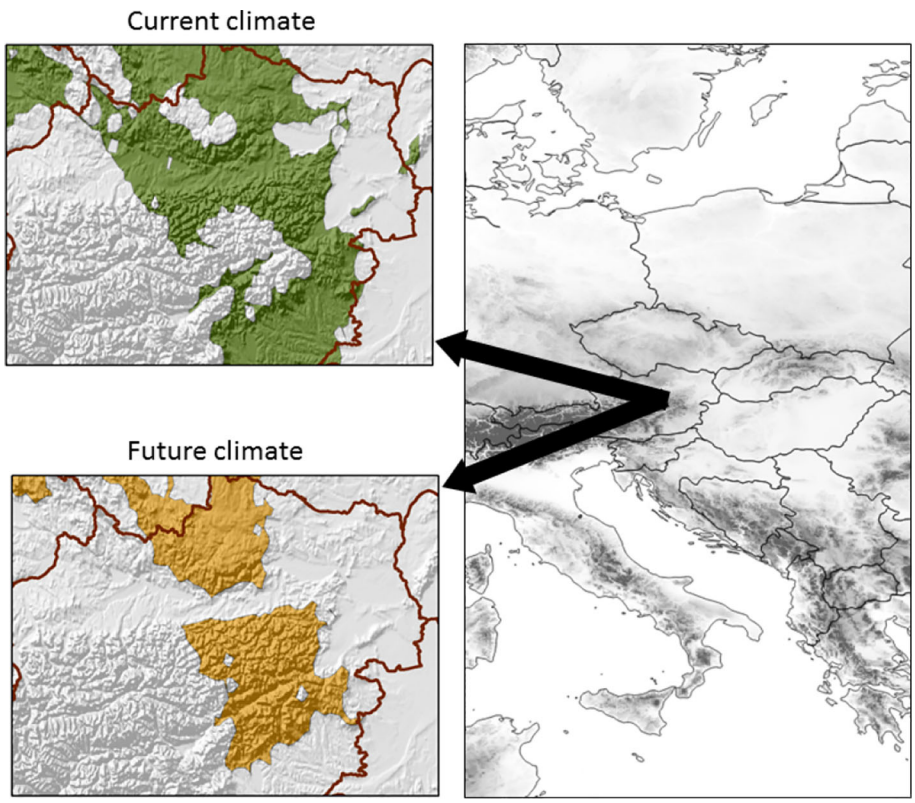


FIGURE 1 | Potential planting area of selected superior larch genotypes under current (top) and expected future (2050) climate conditions (bottom) at the northwestern fringe of the Alpine area.

the top-ranking 25% phenotypes were identified as selection candidates (pre-selection) (Lstibůrek et al., 2015). To account for common environment effects, the average of eight direct neighbors was subtracted from the phenotypic observation of a given selected individual (Zobel and Talbert, 2003), and the adjusted values were used in the genetic evaluation. The total number of pre-selection individuals across the 21 reforestation sites is 1,088 (representing 579 and 509 random and top-phenotype selections, respectively). This sample size was optimized to meet three important criteria: 1) achieving comparable genetic response to selection to that of traditional recurrent selection with structured control crosses (i.e., progeny testing) (White et al., 2007), 2) reconstructing pedigree with sufficient accuracy (Marshall et al., 1998; Kalinowski et al., 2007), and 3) satisfying the declared effective population size (i.e., genetic diversity) in the target seed production population (Lstibůrek et al., 2011). The above calculation of sample size also accounted for anticipated pollen contamination, i.e., paternal contributions originating from parents outside the seed orchard (Lstibůrek et al., 2012).

Tissue samples for DNA extraction were collected using a 15-mm hole-punch to obtain cambium cell layers. Samples were immediately dried and stored in silica gel. DNA extraction followed a modified CTAB protocol (Lefort and Douglas, 1999), using app. 100 mg of frozen cambium tissue after grinding in a Mixer Mill MM200 (Retsch). Extracted DNA was fingerprinted using three microsatellite multiplexes accommodating 5 (Ld30, bcLK189, bcLK228, bcLK263, and Ld56), 4 (Ld31, Ld50, bcLK211, bcLK253), and 4 (Ld58, Ld42, Ld101, 4 Ld45) (Wagner et al., 2012). In total, 53 parental and 1,088 offspring trees were genotyped.

Pedigree Reconstruction

We performed pedigree reconstruction, yielding 1,024 offspring in 491 full-sib families, representing the largest known forest tree pedigree assembly. The likelihood-based method Cervus (Marshall et al., 1998) was used to reconstruct family relationships (**Figure 2**). Pedigree analysis parameters were: unknown sexes, no assumption for putative maternal contributor, LOD score (natural logarithm of the overall likelihood ratio), and Delta (the difference in LOD between the two most likely candidate parents), reflecting inputted parameters of genotyping errors and incomplete sampling of the parental population. Initial simulation of parentage analysis was processed for 10,000 offspring based on 53 unique genotypes that were considered as candidate parents with six parameter scenarios, including input parameters such as proportion of sampled parental population (0.5, 0.6, 0.7, 0.8, 0.9, 1) and maximal genotyping error rate (0.01, 0.1, 0.01, 0.1, 0.01, 0.1) to assess the parentage assignment robustness. Additional parameters include: minimum number of typed loci of 6, monoecious species with polygamous mating, consideration of selfing, and parentage assignment 99% confidence were kept equal. Only consistent outcomes of family assignment across all scenarios were accepted and used in downstream analysis.

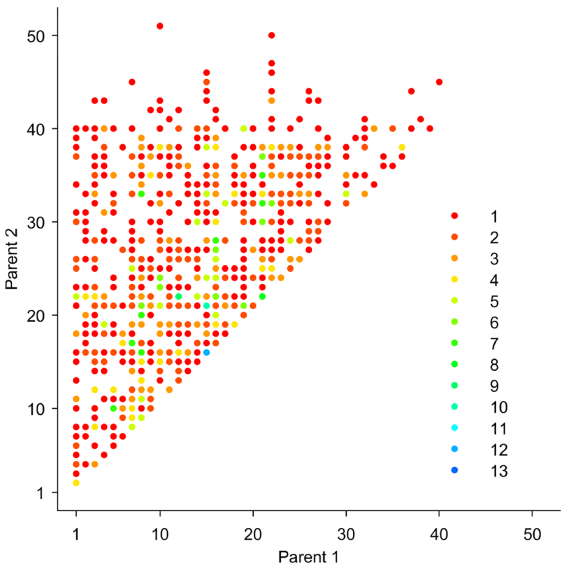


FIGURE 2 | Pedigree reconstruction results showing the formation of full-sib families and their parental combinations and respective family sizes (1-13, reciprocal crosses were grouped) (N = 1,024 offspring).

Statistical Analysis

Pedigree-based genetic analyses were used and variance components, heritabilities, genetic correlations, and individual tree breeding values were estimated/predicted using the bivariate animal model (Henderson, 1984), combining genotyped parental trees and offspring records trees after excluding those with external male parents as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Za} + \mathbf{Yd} + \mathbf{Wu} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the vector of observations for the two traits; $\mathbf{X}$ is the incidence matrix for the fixed effect $\boldsymbol{\beta}$ (trait means); $\mathbf{Z}$ is the corresponding incidence matrix related to random additive genetic effects (breeding values,$\mathbf{a} \sim N(0, \sigma_a^2)$); $\mathbf{Y}$ is the incidence matrix related to random dominance genetic effects $\boldsymbol{d} \sim N(0, \sigma_d^2)$), while $\mathbf{W}$ is an incidence matrix for random genotype by environment (or climatic region) interaction $\boldsymbol{u} \sim N(0, \sigma_u^2)$), and the random residual error effects are distributed as $\boldsymbol{e} \sim N(0, \sigma_e^2)$).

The covariance matrix for the random additive genetic effects was modelled using the heterogeneous covariance structure as

$$\sigma_a^2 = \begin{bmatrix} \sigma_{a1}^2 & \sigma_{a1a2} \\ \sigma_{a2a1} & \sigma_{a2}^2 \end{bmatrix} \otimes \mathbf{A} \quad (2)$$

where $\mathbf{A}$ is the average numerator relationship matrix, σ_{a1a2} is the additive covariance between traits 1 and 2, and $\otimes$ is the Kronecker product operator. A corresponding structure was used for the dominance effects with σ_a^2 being replaced by σ_d^2 and $\mathbf{A}$ by $\mathbf{D}$, i.e., the dominance genetic relationship matrix. The covariance matrix for the random site effects (genotype by

environment interaction) was modelled using a heterogeneous general correlation matrix (equivalent of an unstructured covariance matrix, but with different parametrization) suitable for such a complex correlation structure as

$$\sigma_u^2 = \begin{bmatrix} \sigma_{a1e}^2 & r_{12} \\ r_{21} & \sigma_{a2e}^2 \end{bmatrix} \otimes \mathbf{I} \quad (3)$$

where **I** is the identity matrix.

The random residual error effect was modelled using an unstructured covariance matrix structure as

$$\sigma_e^2 = \begin{bmatrix} \sigma_{e1}^2 & \sigma_{e1e2} \\ \sigma_{e2e1} & \sigma_{e2}^2 \end{bmatrix} \otimes \mathbf{I} \quad (4)$$

where σ_{e1e2} is the residual covariance between the two traits. Random effects were assumed to be independent. The above genetic evaluation was conducted in ASReml software (Gilmour et al., 2008).

Future Parental Population Selection

A selection index was calculated with equal economic weighting on both height and wood density traits. A linear optimum selection model was constructed to maximize the selection response, while meeting the prescribed effective population size in the target seed orchard (Lstibůrek et al., 2015). All selections were unrelated in order to minimize inbreeding depression in seed orchard crop, i.e., future forest plantations. The Optimum-Neighborhood Algorithm was implemented to promote panmixia within the new orchard (Chaloupková et al., 2016).

RESULTS

The wide climatic gradient was confirmed by PCA of the plantation site climate, which resulted in four distinct climatic groups (**Figure 3**), thus extending the testing beyond the confinement of a defined ecological testing target. Furthermore, the reforestation installations were grouped at the "warmer end" of the species distribution (**Figure 4**), thus offering stronger environmental testing conditions (i.e., additional "ecological tension").

Pedigree reconstruction assembled 491 full-sib families, representing 35% of the possible 53-parent half diallel (**Figure 2**). This represents the largest known forest tree pedigree assembly. There was gametic contribution from the entire orchard's parental population, resulting in 1,088 offspring available for the quantitative genetics analyses. It is interesting to note that the gene flow from outside the orchard accounted for 8.4% and 3.4% of the observed matings in the random and pre-selection samples, respectively, meeting theoretical expectations (Lstibůrek et al., 2012). Additionally, offspring resulting from self-pollination was not detected.

Variance components were estimated from the multi-site bivariate mixed linear model. The final model with fixed site and random pedigree effects resulted in heritability (h^2) estimates of 0.25 (SE = 0.065) and 0.30 (SE = 0.072) for height and wood density, respectively, corresponding to reported estimates from Larix species "structured" testing trials (Pâques, 2004; Ratcliffe et al., 2014). Both traits produced non-significant dominance interaction, thus simplifying the model, and the expected family performance was estimated by the mean additive genetic value of the two respective parents. Negative but negligible genetic correlation between height and wood density was observed (-0.04 ± 0.20), corresponding to the known general trend in most conifers (Zobel and Jett, 2012).

No significant genetic variation by environment (site) interaction was observed across the 21 studied sites, leading us to conclude that individuals' additive genetic values are indicative of their general performance across the range of studied sites and that the studied population consists of generalists that performed well under wide site and climatic conditions and further demonstrating that the selected individuals would form appropriate seed production population for planting over a wide climate regime range including potential future conditions (**Figure 1**). These conclusions are supported by the fact that respective parents/families have been tested across a broad range of environments, and selections have been identified across all sites.

Following the genetic evaluation, we selected 25 unrelated offspring individuals with which to establish the new seed production population with improved climate change adaptability and productivity; thus, the phenotyping and genotyping effort were sufficient in capturing the target effective population size. Selected individuals yielded genetic responses of 1.1 and 0.7 standard deviations for fitness and productivity attributes, respectively. Scions collected from selected offspring were grafted onto rootstocks, and the second-generation seed orchard was established (the advanced generation seed production population).

DISCUSSION

In this manuscript, we developed and validated a novel large-scale *in situ* forest gene resource management scheme to identify productive and climatically adapted individuals originating from an Austrian European larch seed production population (seed orchard), utilizing traditional reforestation installations planted widely across the landscape and spanning three decades. Our thesis is based on the expectation that thriving individuals within these installations have been spatially and temporary challenged and thus have effectively dealt with the negative impacts of climate change. In conventional selective breeding programs, a structured pedigree is produced from controlled pollinations following specific mating designs (White et al., 2007), and evaluated in replicated test sites within defined ecological boundaries (Hanewinkel et al., 2013), a prerequisite for effective genetic evaluation and selection. Our approach is anchored on meeting two conditions, namely, the successful assembly of a "structured pedigree" from seed orchard offspring produced under natural pollination (El-Kassaby and Lstibůrek, 2009), and whether progeny evaluation can be conducted within

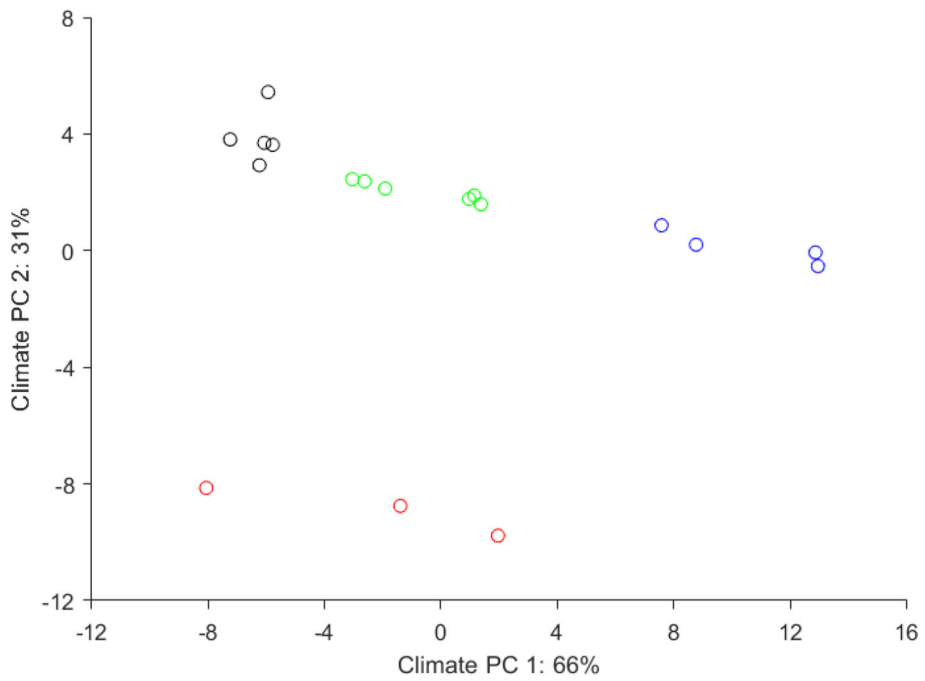


FIGURE 3 | Climatic distribution of the 21 European larch reforestation installations grouped following principal component analysis of climatic parameters. The four climatic groups represented: 1) continental Pannonian climate (black), 2) temperate climate with Atlantic influence (red), 3) a temperate climate with continental influence (green) and, 4) a mountainous climate (blue). These were used as climatic categories to test the presence or absence of genotype-climate interactions.

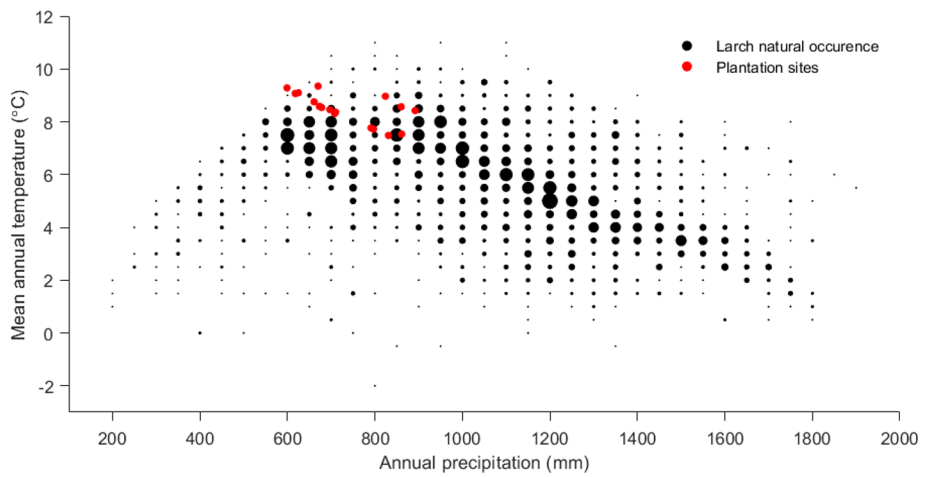


FIGURE 4 | Climate (precipitation and temperature) of the European larch reforestation installations in comparison to species natural occurrence demonstrating that our test sites are close to the warmer border of the species range.

reforestation installations (Lstibůrek et al., 2015) rather than progeny test sites. Meeting these two conditions effectively bypasses the conventional breeding and testing phases used in recurrent selection strategies, thus accelerating and simplifying the selection process. Given that reforestation sites are usually employed across more and a wider range of site and climate conditions than traditional genetic trials, our approach also allows predicting potential application regions for the improved forest seeds. However, for the full Alpine range of larch and for a climate warmer than observed today, the

application might be restricted, as we could expect GxE interaction outside of the current conditions (Koralewski et al., 2015).

This is the first proof of the "Breeding without Breeding, BwB" concept (El-Kassaby and Lstibůrek, 2009) in a large operational forestry program. Our results are in agreement with the respective theoretical expectations published earlier. First, phenotypic preselection provided sufficient distribution of candidates meeting the prescribed effective population size of the new seed orchard, which is in agreement to Lstibůrek et al. (2011). Second, phenotypic preselection was efficient at reducing the contamination rate among the parents of the genotyped subset of offspring, which corresponds to both theoretical expectations (Lstibůrek et al., 2012) as well as the actual findings in Scots pine (Korecký et al., 2014). As noted earlier, actual heritabilities, genetic correlations, and respective standard errors are within the range of conventional breeding programs, facilitating genetic gains that were also in agreement to computer simulations and deterministic expectations (Lstibůrek et al., 2015). Further, we observed the beneficial effect of the increased size of the candidate population (i.e., an increase in selection intensity), which further boosted the genetic response of selection (Lstibůrek et al., 2017) beyond the levels of conventional breeding programs. We can therefore conclude that BwB strategies provide an effective and economically feasible method to breed outcrossing forest tree species. We therefore forecast the utility of BwB methods in operational forestry as they facilitate full-scale landscape gene-resource management of forest trees.

In line with the discussion in Lstibůrek et al. (2017), we advocate the utility of BwB approaches (such as the current study) for the following reasons. (1) Absence of full-sib crosses, as the method relies on natural pollination in breeding arboretums (seed orchards). (2) Absence of progeny trials. Genetic testing can be performed within commercial forest stands. (3) Genetic evaluation thus takes place on a landscape level, emphasizing adaptive traits and their respective interaction with environmental conditions. (4) Strategies are open to NGS platforms. One can replace the BLUP based evaluation (as implemented here) with the genomic alternative (e.g., GBLUP) with all the added benefits of extracting additional genetic parameters (El-Dien et al., 2016). When considering these alternatives, breeders may compare theoretical gain efficiencies between BLUP and GBLUP approaches (Stejskal et al., 2018). At the same time, operational implementation could still remain identical to the current study.

In summary, the approach presented here is a flexible and dynamic gene-resource management scheme that is not encumbered by the predetermined fixed ecological zonation, commonly implemented for forest trees. The reforestation installations permitted effective and rigorous genetic evaluation over numerous sites with varying ecological diversity (geographic distribution) and extended the testing timeframe, thus speeding the selection of adapted individuals and matching them with the most appropriate planting location. The approach is simple and cost-efficient, enabling improvement and conservation of commercial and non-commercial species under rapidly changing environmental conditions.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in https://github.com/mlstiburek/European-larch.

AUTHOR CONTRIBUTIONS

ML, SS, YE-K, and TG conceived the project and designed the study. ML estimated sample sizes. SS, HK, and TG carried out the genotyping. PŠ coordinated the fieldwork. JK conducted the pedigree reconstruction. GH, JS, and ML ran statistical analyses. ML, SS, TG, YE-K, JK, and JS contributed to writing the manuscript.

FUNDING

This research was funded by OP RDE grant Extemit-K, No. CZ.02.1.01/0.0/0.0/15003/0000433 (ML), the Austrian Research Promotion Agency (FFG) and the Cooperation Platform Forst Holz Papier (FHP) and LIECO nurseries and the Austrian Federal Forests (ÖBf) (SS, PŠ, HK, and TG), the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant and the Johnson's Family Forest Biotechnology Endowment (YE-K), Camcore, Department of Forestry and Environmental Resources, NC State University (GH).

ACKNOWLEDGMENTS

We thank Bill Hansson for reviewing our work and for his valuable advice.

REFERENCES

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., et al. (2013). Potential for evolutionary responses to climate change–evidence from tree populations. *Global Change Biol.* 19, 1645–1661. doi: 10.1111/gcb.12181

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., et al. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *For. Ecol. Manag.* 259, 660–684. doi: 10.1016/j.foreco.2009.09.001

Bräutigam, K., Vining, K. J., Lafon-Placette, C., Fossdal, C. G., Mirouze, M., Marcos, J. G., et al. (2013). Epigenetic regulation of adaptive responses of forest tree species to the environment. *Ecol. Evol.* 3, 399–415. doi: 10.1002/ece3.461

Chaloupková, K., Stejskal, J., El-Kassaby, Y. A., and Lstibůrek, M. (2016). Optimum neighborhood seed orchard design. *Tree Genet. Genomes* 12, 105. doi: 10.1007/s11295-016-1067-y

Cown, D. (1978). Comparison of the pilodyn and torsiometer methods for the rapid assessment of wood density in living trees. *New Zeal J. For. Sci.* 8, 384–391.

El-Dien, O. G., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., and El-Kassaby, Y. A. (2016). Implementation of the realized genomic relationship matrix to open-

pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3: Genes Genomes Genet.* 6, 743–753. doi: 10.1534/g3.115.025957

El-Kassaby, Y. A., and Lstibůrek, M. (2009). Breeding without breeding. *Genet. Res.* 91, 111–120. doi: 10.1017/S001667230900007X

Frascaria-Lacoste, N., and Fernández-Manjarrés, J. (2012). Assisted colonization of foundation species: lack of consideration of the extended phenotype concept - Response to Kreyling et al., (2011). *Restor. Ecol.* 20, 296–298. doi: 10.1111/j.1526-100X.2012.00875.x

George, J.-P., Grabner, M., Karanitsch-Ackerl, S., Mayer, K., Weißenbacher, L., Schueler, S., et al. (2017). Genetic variation, phenotypic stability, and repeatability of drought response in European larch throughout 50 years in a common garden experiment. *Tree Physiol.* 37, 33–46. doi: 10.1093/treephys/tpw085

Gilmour, A., Gogel, B., Cullis, B., Thompson, R., Butler, D., Cherry, M., et al. (2008). ASReml user guide release 3.0. *VSN Int. Ltd*.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model Earth Syst.* 5, 572–597. doi: 10.1002/jame.20038

Hanewinkel, M., Cullmann, D. A., Schelhaas, M.-J., Nabuurs, G.-J., and Zimmermann, N. E. (2013). Climate change may cause severe loss in the economic value of European forest land. *Nat. Clim. Change* 3, 203–207. doi: 10.1038/nclimate1687

Henderson, C. R. (1984). *Applications of linear models in animal breeding* Vol. 462 (Guelph: University of Guelph).

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Holeski, L. M., Jander, G., and Agrawal, A. A. (2012). Transgenerational defense induction and epigenetic inheritance in plants. *Trends Ecol. Evol.* 27, 618–626. doi: 10.1016/j.tree.2012.07.011

Hua, F., Wang, X., Zheng, X., Fisher, B., Wang, L., Zhu, J., et al. (2016). Opportunities for biodiversity gains under the world's largest reforestation programme. *Nat. Commun.* 7, 12717. doi: 10.1038/ncomms12717

Isbell, F., Craven, D., Connolly, J., Loreau, M., Schmid, B., Beierkuhnlein, C., et al. (2015). Biodiversity increases the resistance of ecosystem productivity to climate extremes. *Nature* 526, 574–577. doi: 10.1038/nature15374

Jansen, S., and Geburek, T. (2016). Historic translocations of European larch (Larix decidua Mill.) genetic resources across Europe–A review from the 17th until the mid-20th century. *For. Ecol. Manag.* 379, 114–123. doi: 10.1016/j.foreco.2016.08.007

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Koralewski, T. E., Wang, H.-H., Grant, W. E., and Byram, T. D. (2015). Plants on the move: assisted migration of forest trees in the face of climate change. *For. Ecol. Manag.* 344, 30–37. doi: 10.1016/j.foreco.2015.02.014

Korecký, J., Lstibůrek, M., and El-Kassaby, Y. A. (2014). Congruence between theory and practice: reduced contamination rate following phenotypic pre-selection within the breeding without breeding framework. *Scand. J. For. Res.* 29, 552–554. doi: 10.1080/02827581.2014.945616

Kranabetter, J., Stoehr, M., and O'Neill, G. (2012). Divergence in ectomycorrhizal communities with foreign Douglas-fir populations and implications for assisted migration. *Ecol. Appl.* 22, 550–560. doi: 10.1890/11-1514.1

Kremer, A., Ronce, O., Robledo-Arnuncio, J. J., Guillaume, F., Bohrer, G., Nathan, R., et al. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecol. Lett.* 15, 378–392. doi: 10.1111/j.1461-0248.2012.01746.x

Lefort, F., and Douglas, G. C. (1999). An efficient micro-method of dna isolation from mature leaves of four hardwood tree species Acer, Fraxinus, Prunus and Quercus. *Ann. For. Sci.* 56, 259–263. doi: 10.1051/forest:19990308

Lstibůrek, M., Ivanková, K., Kadlec, J., Kobliha, J., Klápště, J., and El-Kassaby, Y. A. (2011). Breeding without breeding: minimum fingerprinting effort with respect to the effective population size. *Tree Genet. Genomes* 7, 1069–1078. doi: 10.1007/s11295-011-0395-1

Lstibůrek, M., Klápště, J., Kobliha, J., and El-Kassaby, Y. A. (2012). Breeding without breeding: effect of gene flow on fingerprinting effort. *Tree Genet. Genomes* 8, 873–877. doi: 10.1007/s11295-012-0472-0

Lstibůrek, M., Hodge, G. R., and Lachout, P. (2015). Uncovering genetic information from commercial forest plantations—making up for lost time using "breeding without breeding". *Tree Genet. Genomes* 11, 55. doi: 10.1007/s11295-015-0881-y

Lstibůrek, M., El-Kassaby, Y. A., Skrøppa, T., Hodge, G. R., Sønstebø, J. H., and Steffenrem, A. (2017). Dynamic gene-resource landscape management of norway spruce: combining utilization and conservation. *Front. Plant Sci.* 8, 1810. doi: 10.3389/fpls.2017.01810

Marshall, T., Slate, J., Kruuk, L., and Pemberton, J. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* 7, 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

McLachlan, J. S., Hellmann, J. J., and Schwartz, M. W. (2007). A framework for debate of assisted migration in an era of climate change. *Conserv. Biol.* 21, 297–302. doi: 10.1111/j.1523-1739.2007.00676.x

Meier, E. S., Lischke, H., Schmatz, D. R., and Zimmermann, N. E. (2012). Climate, competition and connectivity affect future migration and ranges of European trees. *Global Ecol. Biogeogr.* 21, 164–178. doi: 10.1111/j.1466-8238.2011.00669.x

Pâques, L. E. (ed.). (2013). *Forest tree breeding in Europe. Current state-of-the-art and perspectives.* (Dordrecht: Springer). doi: 10.1007/978-94-007-6146-9

Pâques, L. E. (2004). Roles of European and Japanese larch in the genetic control of growth, architecture and wood quality traits in interspecific hybrids (Larix× eurolepis Henry). *Ann. For. Sci.* 61, 25–33. doi: 10.1051/forest:2003081

Pretzsch, H., Biber, P., Schütze, G., Uhl, E., and Rötzer, T. (2014). Forest stand growth dynamics in Central Europe have accelerated since 1870. *Nat. Commun.* 5, 4967. doi: 10.1038/ncomms5967

Ratcliffe, B., Hart, F. J., Klápště, J., Jaquish, B., Mansfield, S. D., and El-Kassaby, Y. A. (2014). Genetics of wood quality attributes in western larch. *Ann. For. Sci.* 71, 415–424. doi: 10.1007/s13595-013-0349-x

Saikkonen, K., Taulavuori, K., Hyvönen, T., Gundel, P. E., Hamilton, C. E., Vänninen, I., et al. (2012). Climate change-driven species' range shifts filtered by photoperiodism. *Nat. Clim. Change* 2, 239–242. doi: 10.1038/nclimate1430

Stejskal, J., Lstibůrek, M., Klápště, J., Čepl, J., and El-Kassaby, Y. (2018). Effect of genomic prediction on response to selection in forest tree breeding. *Tree Genet. Genomes* 14, 74. doi: 10.1007/s11295-018-1283-8

Vitt, P., Havens, K., Kramer, A. T., Sollenberger, D., and Yates, E. (2010). Assisted migration of plants: changes in latitudes, changes in attitudes. *Biol. Conserv.* 143, 18–27. doi: 10.1016/j.biocon.2009.08.015

Wagner, S., Gerber, S., and Petit, R. J. (2012). Two highly informative dinucleotide SSR multiplexes for the conifer Larix decidua (European larch). *Mol. Ecol. Resour.* 12, 717–725. doi: 10.1111/j.1755-0998.2012.03139.x

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J., et al. (2002). Ecological responses to recent climate change. *Nature* 416, 389–395. doi: 10.1038/416389a

Walther, G.-R. (2010). Community and ecosystem responses to recent climate change. *Phil. Tran. R. Soc. B.* 365, 2019–2024. doi: 10.1098/rstb.2010.0021

White, T. L., Adams, W. T., and Neale, D. B. (2007). *Forest Genetics* (Wallingford, Oxfordshire, UK: CABI Publishing, CAB International). doi: 10.1079/9781845932855.0000

Wolkovich, E. M., Cook, B. I., Allen, J. M., Crimmins, T. M., Betancourt, J. L., Travers, S. E., et al. (2012). Warming experiments underpredict plant phenological responses to climate change. *Nature* 485, 494–497. doi: 10.1038/nature11014

Zobel, B. J., and Jett, J. B. (2012). *Genetics of Wood Production* (Springer-Verlag Berlin Heidelberg: Springer Science & Business Media).

Zobel, B. J., and Talbert, J. (2003). *Applied forest improvement* (Caldwell, NJ: The Blackburn Press).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Lstibůrek, Schueler, El-Kassaby, Hodge, Stejskal, Korecký, Škorpík, Konrad and Geburek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fgene.2020.00022

Genome-Wide Association Mapping to Identify Genetic Loci for Cold Tolerance and Cold Recovery During Germination in Rice

Ranjita Thapa[1†], *Rodante E. Tabien*[2], *Michael J. Thomson*[1] *and Endang M. Septiningsih*[1*]

[1] Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, United States, [2] Texas A&M AgriLife Research Center, Beaumont, TX, United States

OPEN ACCESS

Edited by:
Charles Chen,
Oklahoma State University,
United States

Reviewed by:
Sabrina Moriom Elias,
University of Dhaka,
Bangladesh
Prasanta Kumar Subudhi,
Louisiana State University,
United States

***Correspondence:**
Endang M. Septiningsih
eseptiningsih@tamu.edu

[†]**Present address:**
Ranjita Thapa,
Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, United States

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 20 September 2019
Accepted: 07 January 2020
Published: 21 February 2020

Citation:
Thapa R, Tabien RE, Thomson MJ and Septiningsih EM (2020) Genome-Wide Association Mapping to Identify Genetic Loci for Cold Tolerance and Cold Recovery During Germination in Rice. Front. Genet. 11:22. doi: 10.3389/fgene.2020.00022

Low temperature significantly affects rice growth and yield. Temperatures lower than 15°C are generally detrimental for germination and uniform seedling stand. To investigate the genetic architecture underlying cold tolerance during germination in rice, we conducted a genome-wide association study using a novel diversity panel of 257 rice accessions from around the world and the 7K SNP marker array. Phenotyping was conducted in controlled growth chambers under dark conditions at 13°C. The rice accessions were measured for low-temperature germinability, germination index, coleoptile length under cold stress, plumule length at 4-day recovery, and plumule length recovery rate. A total of 51 QTLs were identified at $p < 0.001$ and 17 QTLs were identified using an FDR < 0.05 across the different chilling indices with the whole panel of accessions. At the threshold of $p < 0.001$, a total of 20 QTLs were identified in the subset of *japonica* accessions, while 9 QTLs were identified in the subset of *indica* accessions. Considering the recurring SNPs and linked SNPs across different chilling indices, we identified 31 distinct QTL regions in the whole panel, 13 QTL regions in the *japonica* subset, and 7 distinct QTL regions in the *indica* subset. Among these QTL regions, three regions were common between the whole panel and *japonica*, three regions were common between the whole panel and *indica*, and one region was common between *indica* and *japonica*. A subset of QTL regions was potentially colocalized with previously identified genes and QTLs, including 10 from the *japonica* subset, 4 from the *indica* subset, and 6 from the whole panel. On the other hand, a total of 21 potentially novel QTL regions from the whole panel, 10 from the *japonica* subset, and 1 from the *indica* subset were identified. The results of our study provide useful information on the genetic architecture underlying cold tolerance during germination in rice, which in turn can be used for further molecular study and crop improvement for low-temperature stressed environments.

Keywords: low temperature stress, cold tolerance, cold recovery, germination, genome-wide association study, SNP, QTL, rice

INTRODUCTION

Rice is more susceptible to cold stress than other cereal crops due to its origin in the tropical and subtropical regions (Zhao et al., 2017). Low temperature causes major stress for rice growing in 25 countries (Cruz et al., 2013) and to more than 15 million ha of rice grown worldwide (Bai et al., 2016). One of the major challenges for rice production under direct-seeded cultivation, especially in high altitude regions in the tropics or regions with temperate climates, is low-temperature sensitivity during the germination stage (Schläppi et al., 2017). Cold stress during germination causes poor germination and retarded plant growth. Vigorous germinated seedling is a necessity for good plant establishment. Breeding of rice cultivars with tolerance of low temperature, however, has been challenging due to various factors: response of rice plants to cold varies with growth stages (Liu et al., 2015); low-temperature tolerance is controlled by quantitative loci where many genes with small effects contributing to the phenotype (Ji et al., 2009); and epistatic interaction among alleles at unlinked loci (Zhang et al., 2014a). A wide range of variations to cold tolerance among *Oryza sativa* has been reported where accessions of the *japonica* subspecies were generally being more tolerant than *indica* (Baruah et al., 2009). A few studies have been performed to improve cold tolerance of the *indica* cultivars using *japonica* cultivars; however, due to lack of genetic diversity in *japonica* germplasm, further improvement of *japonica* cultivars has been quite challenging (Zhang et al., 2014a).

Thus far, only very few genes controlling chilling tolerance have been identified in different stages of rice growth (Cruz et al., 2013; Zhang et al., 2014b; Zhang et al., 2014c). The first gene reported for low-temperature germinability was *qLTG3-1*, where the gene encodes for a secreted hybrid glycine-rich protein and a single nucleotide was the causal polymorphism (Fujino et al., 2008). It is highly imperative to identify additional chilling tolerance QTLs and genes to better understand the mechanisms of chilling tolerance in rice and to assist in developing high-yielding rice with higher tolerance of cold during germination. QTL mapping and genome-wide association study (GWAS) are two widely used tools to discover the genetic control of complex traits. Most of the published data on genetic loci controlling chilling tolerance in rice were obtained by bi-parental mapping populations from *O. sativa* ssp *indica* X *O sativa* ssp *japonica* crosses where *japonica* subspecies usually used as the donors for cold tolerance (Mackill and Lei, 1997; Cruz and Milach, 2004; Mao and Chen, 2012; Ma et al., 2015). The major drawback of bi-parental mapping is the limitation of genetic background to parental lines. More recently, GWAS has also been utilized to study cold tolerance in rice, with the advantage of scanning a large number of accessions for genetic loci controlling this trait. These studies have led to the discovery of QTLs associated with low-temperature germination during seedling stage and plumule growth recovery after chilling stress in rice. Pan et al. (2015) identified 22 QTLs for cold tolerance during germination stage using SSR markers in 174 Chinese accessions. Sales et al. (2017) detected 24 SNPs associated with low-temperature germination and growth rate at low temperature; while Shakiba et al. (2017) reported 42 QTLs controlling cold tolerance at seedling stage. Fujino et al. (2015) conducted GWAS mapping with 117 markers using a Hokkaido rice core panel, comprising 63 Japanese landraces and breeding lines and discovered 6 QTLs for cold tolerance at heading stage and 17 QTLs for low-temperature germinability. Lv et al. (2016) reported 132 loci associated with 16 traits evaluated under natural chilling and cold shock stress using a large collection of 529 rice accessions with more than 4.35 million SNP markers. Schläppi et al. (2017) identified a total of 48 QTLs for chilling tolerance in 202 *O. sativa* accessions from the USDA mini-core collection.

Various methods, traits measured, and temperatures have been used to identify underlying genes of cold tolerance in rice during germination stage. Shakiba et al. (2017) evaluated cold tolerance at germination stage using the "ragdoll method" exposed at 12°C for 35 days. Sales et al. (2017) germinated rice seeds for 21 days at 15°C to evaluate the cold tolerance during germination stage. Schläppi et al. (2017) conducted GWAS during germination, seedling and recovery stage in 202 *O. sativa* accessions. For germination cold tolerance, growth rate of plumule after 30 days of cold exposure at 10°C was measured 4 days after recovery at 28 ± 1°C. The mean length of 2-week old seedlings at V2 stage was recorded before cold exposure and again after 1 week of recovery, the length of the recovered seedlings was measured. The growth at 28 ± 1°C following a 1-week chilling stress treatment at 10 ± 1°C was recorded to estimate leaf recovery growth rate after cold exposure. These different stress treatments and different indicators used to study cold tolerance have resulted in variation in the number and location of the identified QTLs (Zhang et al., 2014c). Cruz and Milach (2004), however, suggested that variation in coleoptile growth and percentage of seeds superior to 5-mm coleoptile length at cold temperatures were sufficient to identify cold-tolerant genotypes. The differences in seedling vigor among genotypes may cause difficulty in the identification of cold-tolerant lines. Because of this, several researchers have emphasized the evaluation of test entries in both control (ambient) and cold temperature to enable the separation of seedling vigor from cold tolerance (Sales et al., 2017).

In this study, we performed GWAS on a novel rice diversity panel of 257 accessions using a 7K rice SNP array (S. McCouch, M. Thomson, and K. Morales, personal communication). The objectives of this study were to evaluate the diversity rice panel for cold tolerance and cold recovery during the germination stage and to identify QTLs and the underlying candidate genes.

MATERIALS AND METHODS

Rice Accessions

The 257 rice accessions/lines used in this study were obtained from the USDA-ARS National Small Grains Collection (Aberdeen, Idaho), the Genetic Stocks-Oryza (GSOR) collection located at the USDA-ARS Dale Bumpers National Rice Research Center (USDA-ARS DBNRRC; Stuttgart, AR), and

the inbred rice breeding program at the Texas A&M AgriLife Research Center in Beaumont, Texas (**Supplementary Table 1**). This panel represented accessions or breeding lines belonging to the *indica* subspecies (*indica* and *aus)*, the *japonica* subspecies (*aromatic*, *tropical japonica*, and *temperate japonica*), *O. glaberrima*, and several Nerica lines (derivatives of *O. sativa*/*O. glaberrima* interspecific crosses). This novel diversity panel was selected to represent geographic diversity, including 62 accessions from South Asia, 50 from Central and Western Asia, 27 from Southeast Asia, 8 from East Asia, 34 from Africa, 15 from Europe and Russia, 6 from Latin America, and 55 from North America. Seed multiplication was performed in the Texas A&M AgriLife Research Center in Beaumont, TX (summer 2016). Seeds from one panicle of each accession were direct-seeded in a single row for seed multiplication in summer 2017. All plants were maintained following the Texas production guidelines. After maturity, per plant harvest was performed and the seeds were dried in a heated air dryer at 37°C for 5 days and then stored at 4°C. To break the dormancy, seeds were incubated at 50°C for 5 days. The germination test of each accession was performed using the roll paper method (http://www.knowledgebank.irri.org/step-by-step-production/pre-planting/seed-quality).

Indices for Evaluating Cold Tolerance

To screen for the cold tolerance variability in the collected germplasm, different parameters were used, including low-temperature germinability (LTG), germination index (GI), coleoptile length under cold stress (CLC), plumule length at 4-day recovery (PLR), and plumule length recovery rate (PLRR) that are described in detail below. The experiment was conducted in a growth chamber in a controlled-dark condition following a completely randomized design with three replications, and 30–40 seeds per replication were used. Seeds of all accessions were rinsed with 5% Tween-20 for 5 min followed by thorough rinsing with 10% bleach (sodium hypochlorite) for 10 min and washed with autoclaved distilled water 3 times to prevent contamination.

For control samples, 30–40 sterilized seeds were placed on water-soaked filter paper placed in the petri dishes. The petri dishes were then wrapped in aluminum foil and placed for germination in a growth chamber maintained at 30°C. The experiment was conducted in a completely randomized design and the dark condition was provided to mimic the natural dark condition under the soil during the germination stage. After 7 days of germination, the average germination percentage per accession was taken from all the three replicates.

Low-Temperature Germinability (LTG) and Germination Index (GI)

Surface sterilized seeds were incubated in water-soaked filter paper in petri dishes, 30–40 seeds were placed in each petri dish and these were then wrapped with aluminum foil. For each entry, three plates were randomly distributed in the growth chamber set at 13°C temperature. Another set was grown in the growth chamber at 30°C temperature as controls. Germinated seeds were counted in each petri dish obtained after 7 days in the 30°C growth chamber (control) and after 28 days in the 13°C growth chamber (cold treatment). Germination was defined as visible coleoptile emergence (>5 mm) through the hull. The low-temperature germinability (LTG) was calculated as the percent of seeds germination at 13°C after 28 days. The mean LTG scores were recorded from three petri dishes and normalized with the mean percent germinability of seeds at 30°C (NTG) which was used to calculate the germination index (GI). GI was determined as LTG divided by NTG times 100.

Coleoptile Length Under Cold Stress (CLC)

After counting the germinated seeds, all the germinating seedlings were placed on a sterile black background paper along with the ruler for photographs. The images of all the germinating seeds from each replication were then taken with a Pentax camera. Later, the images were imported to ImageJ software and the coleoptile length of all the germinated seeds was measured and averaged to represent the mean of coleoptile length of each accession after cold exposure. The arithmetic means of the measurement were used for GWAS mapping.

Plumule Length At 4-Day Recovery (PLR) and Plumule Length Recovery Rate (PLRR)

After photographing all germinating seeds, the seedlings were returned to their corresponding petri dish, covered with foil and then were moved to a growth chamber maintained at 30°C and were kept for 4 days. After the recovery period of 4 days, pictures were taken, and plumule lengths were measured using ImageJ. The average from three replication was taken as plumule length after recovery (PLR) for each accession. The mean plumule growth rate after cold germination was estimated by subtracting the mean coleoptile length after 28 days at 13°C from the mean plumule length on day 4 at 30°C after recovery and dividing the obtained value by 4 to represent plumule length recovery rate (PLRR). The PLRR value, therefore, indicates the growth rate of the plumule over a period of 4 days under normal conditions (30°C).

Genotyping

The young leaves were collected from the field in Beaumont, Texas in 2017 and sent to a genotyping service lab for DNA extraction and genotyping (Eurofins BioDiagnostics, Inc., River Falls, WI). Genotyping of all the accessions was performed using a 7K Illumina iSelect custom-designed array by following the Infinium HD Array Ultra Protocol. The 7K array, called the C7AIR, was designed by the McCouch Lab at Cornell University and consists of 7,098 SNPs. The Cornell_7K_Array_Infinium_Rice (C7AIR) design represents an improved version of Cornell_6K_Array_Infinium_Rice (C6AIR) (Thomson et al., 2017). Genotype data used in this study were called using Genome Studio software (Illumina, USA). SNPs of call rate <90% and minor allele frequency <5% were removed from the dataset. The quality of each SNP was confirmed manually by re-clustering. For our study, a subset of 5,185 high-quality SNP markers obtained after removal of rare allele markers at 5% or less and heterozygosity of more than 20% were used to perform the genome-wide association analysis.

General Statistics, Population Structure, and Association Mapping

The basic statistics for all traits were analyzed, including heritability (Singh et al., 1993). Spearman's correlation coefficients between the chilling indices were also calculated using R software version 3.5.1 (Lenth, 2016). Additionally, the mean and standard error for the five selected traits were calculated for each sub-population generated from the STRUCTURE program (Pritchard et al., 2000); comparisons were then made between these distributions to the generally more tolerant *temperate japonica* population using a Student's *t*-test.

The Bayesian model of the Markov Chain Monte Carlo (MCMC) implemented in the STRUCTURE program (Pritchard et al., 2000) was used to estimate the population structure. The burn-in length and number of replications were both set to be 100,000. For each number of populations (Q), five iterations were performed for the number of populations 2 to 10. The Structure Harvester program (Earl, 2012) was used to perform the analysis. The coefficient of ancestry (Q) threshold was defined at 70% to refer an individual with its inferred ancestry from one single group; while the accessions which were unable to be assigned to only one group were determined as mixed ancestry. We also used the Bayesian clustering program fastStructure (Raj et al., 2014) to estimate the different levels of Q (Q = 1–10).

GWAS of all *japonica* (*temperate japonica*, *tropical japonic*a, *aromatic*), *indica* (*aus* and *indica*), and the whole panel were conducted using their corresponding data sets. The Genome Association and Prediction Integrated Tool (GAPIT) package (Lipka et al., 2012) with a genotype matrix of 5185 SNPs and a phenotype matrix of 257 accessions were used to perform the GWAS analysis. To predict the genomic regions associated with the traits, we used mixed linear model (MLM) of GAPIT (Zhang et al., 2010). For MLM, we used both kinship (K) matrix as the variance-covariance matrix between the individuals and population structure (Q) matrix to control false positive. The structure data was obtained from the STRUCTURE software (Pritchard et al., 2000) and the kinship relationship matrix (K) was obtained from the TASSEL 4.0 software (Bradbury et al., 2007). For association mapping in *japonica* and *indica* sub-populations, considering the low sample size, the MLM model of GAPIT using principal components (PCs) was used to avoid overcorrection (Hsu and Tung, 2015).

The MLM model used is: $Y = \beta X + \gamma P + Zu + \epsilon$; where Y is the vector of the phenotypic data, X is the vector of genotypic data, β represents the SNP effect, P is the vector of the Q matrix representing population structure, γ is the effect of population structure, u refers to the random effect from kinship, Z is the Kinship matrix, and ϵ corresponds to random error. The expected p-values versus the observed p-values test statistics for the SNP markers were plotted (QQ plot) to assess the control of type I errors under multiple run parameters. The markers were defined to be significantly associated to chilling indices based on $p < 0.001$. The extent of LD in rice on average ranges from 100 to 500 kb (Garris et al., 2005; Myles et al., 2009; Tung et al., 2010). Hence, we defined two or more SNPs positioned within ~250 kb as a single QTL. The Manhattan plot distribution chart was obtained by the R software. The percent variance explained by all significant SNPs discovered for each trait was estimated by subtracting the R^2 of the model without SNP from R^2 of the model with SNP (Zhang et al., 2010). Candidate genes at or near the QTLs identified in this study (within ~250 kb) were from the QTARO database (http://qtaro.abr.affrc.go.jp/; Yonemaru et al., 2010) and other previously published literature.

RESULTS

Phenotypic Performance and Correlation Among Traits

Most of the rice accessions used in this study have more than 90% germination rate at 30°C. However, we observed a wide variation in coleoptile length (**Table 1**). In most cases, germination and coleoptile length were significantly decreased when the rice seeds were germinated at a low temperature of 13°C. Under cold exposure, LTG ranged from 0% to 100%. Cold temperature delayed the germination rate of rice seeds and many of the lines started germinating after 7 days of sowing. The range of the coleoptile length was found to be 0 cm to 1.69 cm; while the mean was 0.69 cm. The PLR and the PLRR ranged from 0 cm to 5.33 cm and 0 cm/d to 1.08 cm/d, with the mean values of 2.96 cm and 0.57 cm/day, respectively. Overall, the broad sense heritability estimations for all traits were high, ranging from 86.8% to 94.0% (**Table 1**).

Based on the population structure, the whole panel was categorized into nine sub-populations, including the admixtures (**Table 2**). We observed that among the highest LTGs were the group of Texas breeding lines and the US released varieties, followed by NERICA lines and *temperate japonica*, with the means of 86.2%, 80.81%, and 80.3%, respectively; while the lowest was *O. glaberrima* with the mean of 41.12%. Among the *japonica*, *aromatic* has the lowest LTG (64.17%), a similar rate to the *aus* group (53.76%). Interestingly, the *indica* lines used in our study (73.67%) had comparable germination rates under cold stress with several of the *japonica* lines. The three groups having the highest LTG are almost the same as CLC, with *temperate japonica* having the highest mean

TABLE 1 | Descriptive statistics of each trait measured in the whole diversity panel.

Trait	Description	Measurement unit	Mean	Range	SE	H^2
LTG	Low-temperature germinability	%	69.21	0–100	1.69	94.0
GI	Germination index	NA	72.47	0–107.60	1.76	87.7
CLC	Coleoptile length under cold conditions	cm	0.69	0–1.69	0.02	89.2
PLR	Plumule length after recovery	cm	2.96	0–5.33	0.07	90.8
PLRR	Plumule length recovery rate	cm/d	0.57	0–1.08	0.01	86.8

TABLE 2 | Phenotypic performance of nine sub-populations generated within the whole diversity panel.

Sub-pop[a]	# samples	LTG[b] (%)	NTG[c] (%)	GI[d]	CLC[e] (cm)	PLR[f] (cm)	PLRR[g] (cm/d)
Admixture	11	60.73 ± 8.85	94.08 ± 4.60	63.78 ± 9.18*	0.67 ± 0.10*	2.63 ± 0.27**	0.49 ± 0.04**
Aromatic	20	64.17 ± 5.26*	93.59 ± 1	68.43 ± 5.50**	0.61 ± 0.05***	3.13 ± 0.16*	0.63 ± 0.03
Aus	53	53.76 ± 3.86***	96.74 ± 0.48***	55.25 ± 3.93***	0.48 ± 0.03***	2.46 ± 0.11***	0.5 ± 0.02***
Oryza glaberrima	20	41.12 ± 7.85***	94.74 ± 0.92	42.47 ± 8.07***	0.54 ± 0.09***	2.07 ± 0.27***	0.38 ± 0.05***
Nerica	9	80.81 ± 4.98	92.48 ± 1.45	87.74 ± 5.85	0.8 ± 0.09	2.65 ± 0.15***	0.46 ± 0.02***
Indica	48	73.67 ± 3.56	96.42 ± 0.64***	76.34 ± 3.65*	0.63 ± 0.05***	2.67 ± 0.15***	0.51 ± 0.03***
Temperate japonica	30	80.3 ± 2.59	92.17 ± 0.94	87.04 ± 2.56	0.94 ± 0.06	3.68 ± 0.14	0.69 ± 0.03
Texas	44	86.2 ± 5.43	95.51 ± 0.67**	90.21 ± 5.60	0.91 ± 0.06	3.96 ± 0.17	0.76 ± 0.04*
Tropical japonica	22	77.16 ± 5.43	97.4 ± 0.67***	79.24 ± 5.60	0.74 ± 0.06*	2.89 ± 0.16***	0.54 ± 0.04**

[a]*These sub-populations generated by the STRUCTURE software.*
[b]*Low-temperature germinability.*
[c]*Normal-temperature germinability.*
[d]*Germination index.*
[e]*Coleoptile length after cold exposure.*
[f]*Plumule length after recovery.*
[g]*Plumule length recovery rate.*
**p-value < 0.05; **p-value < 0.01; ***p-value < 0.001.*

for CLC (0.94 cm), followed by Texas and US lines (0.91 cm) and then the NERICA lines (0.80 cm). Similarly, the smallest length for CLC was also observed in *aus* followed by *O. glaberrima* and *aromatic*. CLC of the *indica* (0.63 cm) was generally shorter compared to the *japonica* groups except for the *aromatic* (0.61 cm). The recovery process from cold stress was also evaluated using the PLR and PLRR parameters. For PLR, the Texas lines and *temperate japonica* had the longest plumule growth with the mean values of 3.96 cm and 3.68 cm, respectively; whereas the shortest growth was seen in *O. glaberrima* with a mean value of 2.07 cm. A similar trend was observed for PLRR (**Table 2**). The values of LTG and PLR were significantly lower for *O. glaberrima*, *aus*, and *admixtures* (**Table 2**; **Figure 1**). For CLC and PLRR, the values were significantly lower for *O. glaberrima*, *aus*, *aromatic*, *indica*, and admixtures. We also observed significantly higher PLRR of Texas lines compared to *temperate japonica* ($p < 0.05$); whereas no significant difference of CLC was observed between Texas lines and *temperate japonica*. Very high significant positive correlations between all the chilling indices were detected (**Table 3**), albeit with different levels of significance ranging from 0.44 (between CLC and PLRR) to 0.97 (between LTG and GI). The results showed that

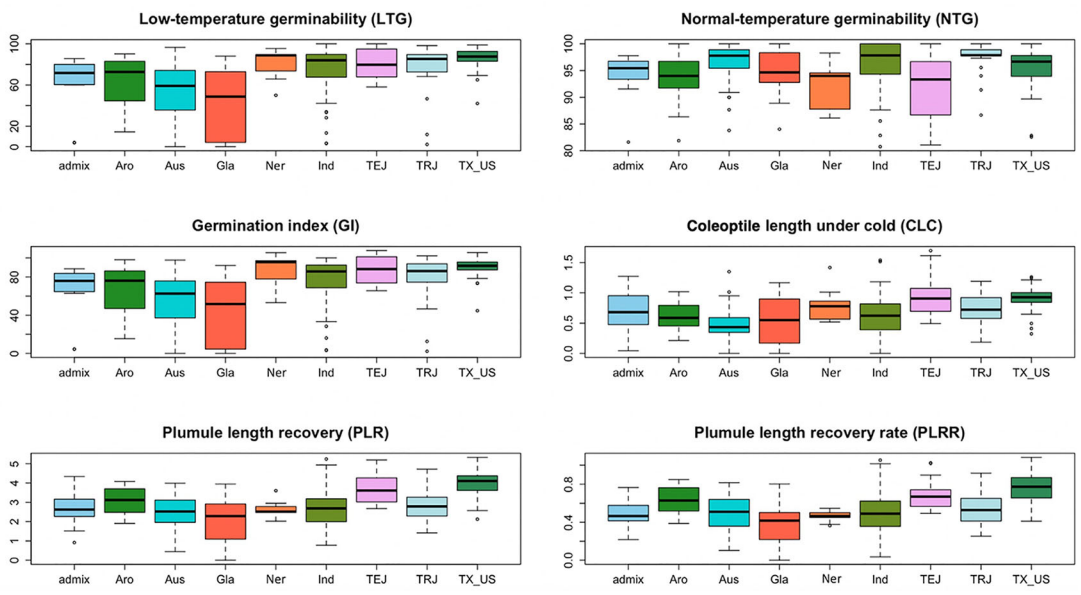


FIGURE 1 | The box plots from all sub-populations identified by the STRUCTURE software presented for all the traits measured: low-temperature germinability (LTG), percentage germination under normal condition (NTG), germination index (GI), coleoptile length under cold (CLC), plumule length recovery (PLR), and plumule length recovery rate (PLRR).

TABLE 3 | Correlation analysis of different cold tolerance traits among all accessions.

Trait	LTG[a]	GI[b]	CLC[c]	PLR[d]	PLRR[e]
LTG	1	0.97***	0.64***	0.58***	0.47***
GI	0.97***	1	0.65***	0.60***	0.50***
CLC	0.64***	0.65***	1	0.67***	0.44***
PLR	0.58***	0.60***	0.67***	1	0.96***
PLRR	0.47***	0.50***	0.44***	0.96***	1

[a]*Low-temperature germinability.*
[b]*Germination index.*
[c]*Coleoptile length after cold exposure.*
[d]*Plumule length after recovery.*
[e]*Plumule length recovery rate.*
****p-value < 0.0001.*

rice accessions having good germination under cold stress in general also having higher coleoptile length under cold stress and high recovery rate as well.

GWAS for Identification of QTLs

The population structure analysis for the whole accessions identified nine sub-populations. The results for the *japonica* and *indica* group-specific GWAS, however, showed an overcorrection for the population structure when both population structure (Q) and kinship matrix (K) were considered in the mixed model (*japonica* MLM and *indica* MLM). To avoid this overcorrection and to control the false-negative results, a GAPIT model considering the principal components (PCs) were used to individually analyze the *indica* and *japonica* varietal groups. Only two main sub-populations were observed in the *indica* group and three sub-populations were observed in the *japonica* group as depicted by the PCA plot results from GAPIT output. In *indica* group, the first PC and second PC explained 28% and 5% of the total variance, respectively, whereas in *japonica* group, the first PC, second PC, and third PC explained 32%, 5%, and 4% of the total variance, respectively.

GWAS of Chilling Tolerance Indices for All Accessions

A total of 51 QTLs were identified at $p < 0.001$, with 11, 15, 9, 9, and 7 QTLs were discovered to be significantly associated with LTG, GI, CLC, PLR, and PLRR, respectively (**Table 4**; **Supplementary Table 2**; **Figure 2**; **Supplementary Figure 1**). Out of the 51 QTLs, 17 of them were detected at FDR < 0.05, with 4, 4, 2, 6, and 1 QTLs were found to be associated with LTG, GI, CS, PLR, and PLRR, respectively (**Table 4**). The amount of phenotypic variance explained (R^2) ranged from 0.5% to 20.6% for LTG, 1.3% to 4.8% for GI, 1.8% to 12.9% for CLC, 0.6% to 8.6% for PLR, and 0.8% to 10.0% for PLRR.

Considering the reoccurring SNPs and very closely linked SNPs in multiple chilling indices of the 51 QTLs, 31 unique QTL regions were identified to be significantly associated with cold tolerance indices within well-fitted QQ plots. Out of these 31 regions, 17 of them harbored at least 2 QTLs (QTL clusters) from the various chilling indices. For example, five QTLs identified

TABLE 4 | QTLs with FDR < 0.05 detected in the whole panel, *japonica* and *indica* subsets, and colocalized genes and QTLs.

QTL ID	Trait[a]	Colocated QTL in this study	Group	Chr.	Position (bp)	*p*-value	FDR	R^2	Potentially colocated QTL/gene	Reference
qPLRR-1	PLRR	*qLTG-1-1; qGI-1-1; qCLC-1-2; qPLR-1*	Full set	1	2,994,2776	4.39E–29	2.28E–25	10.0		
qLTG-1-1	LTG	*qGI-1-1; qCLC-1-2; qPLR-1; qPLRR-1*	Full set	1	2,9942,776	2.70E–14	1.40E–10	20.1		
qCLC-1-2	CLC	*qLTG-1-1; qGI-1-1; qPLR-1; qPLRR-1*	Full set	1	29,942,776	1.29E–09	6.69E–06	12.9		
qPLR-1	PLR	*qLTG-1-1; qGI-1-1; qCLC-1-2; qPLRR-1*	Full set	1	29,942,776	2.74E–09	1.42E–05	8.4		
qGI-1-1	GI	*qLTG-1-1; qCLC-1-2; qPLR-1; qPLRR-1*	Full set	1	29,942,776	8.02E–06	0.0191873	1.5		
qPLR-2	PLR	*qPLRR-2*	Full set	2	26,231,409	5.31E–05	0.0458891	1.4	*qSWTPNCT2-2; qnob-5*	Shakiba et al., 2017
qGI-5-3	GI		Full set	5	28,831,954	1.48E–05	0.0191873	3.1	*Os05g0574500*	Chen et al., 2011
qLTG-5-1	LTG	*qGI-5-1*	Full set	5	805,425	1.83E–05	0.0237779	3.7		
qPLR-5	PLR	*qLTG-5-2; qGI-5-2; qCLC-5*	Full set	5	7,195,992	2.35E–05	0.0405948	2.2		
qGI-9-2	GI		Full set	9	15,414,541	1.18E–05	0.0191873	4.8	*Os09g0417600*	Yokotani et al., 2013
qPLR-9-2	PLR	*qGI-9-1; qlnPLR-9; qPLRR-9*	Full set	9	14,648,157	7.42E–06	0.0192423	3.2	*qCTS-9; OsWRKY76; Os09g0410300*	Peng et al., 2010
qCLC-9	CLC	*qJaCLC-9*	Full set	9	9,230,514	1.26E–05	0.0325983	5.1	*qLTSS9-1*	Schläppi et al., 2017
qPLR-9-3	PLR	*qGI-9-2*	Full set	9	15,399,656	4.05E–05	0.0420114	8.6		
qPLR-9-4	PLR	*qGI-9-3*	Full set	9	16,325,535	4.05E–05	0.0420114	8.6	*qSWTCT9; qCTS9-8*	Wang et al., 2016; Shakiba et al., 2017
qLTG-10	LTG	*qGI-10; qPLR-10*	Full set	10	13,897,640	6.19E–09	1.60E-05	4.2		
qGI-10	GI	*qLTG-10; qPLR-10*	Full set	10	13,897,640	1.21E–06	0.0062886	1.3		
qLTG-11-2	LTG		Full set	11	87,88,201	4.87E–06	0.0084191	10.6		

[a]*LTG, low-temperature germinability; NTG, normal-temperature germinability; GI, germination index; CLC, coleoptile length after cold exposure; PLR, plumule growth after cold exposure; PLRR, plumule growth rate after cold exposure.*

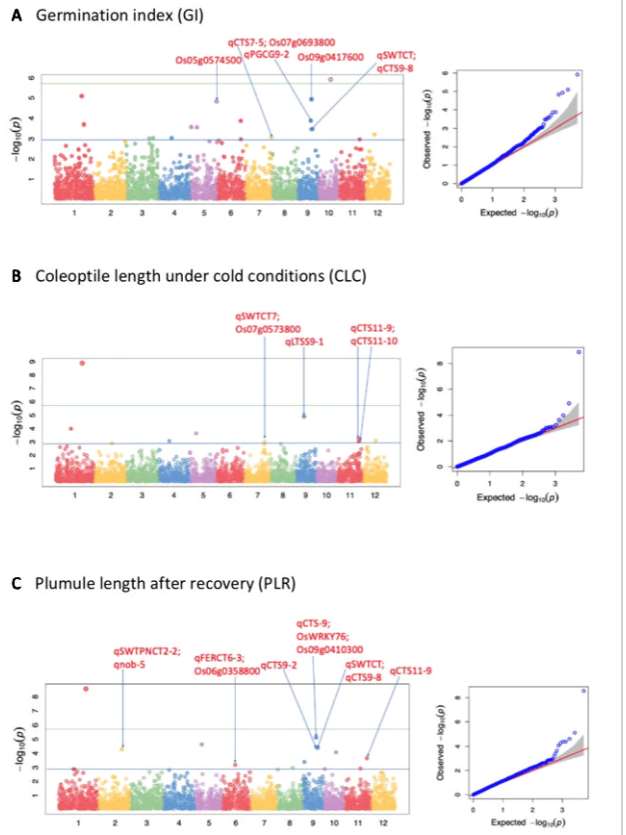


FIGURE 2 | Selected Manhattan and QQ plots of the whole panel for germination index (GI), shoot length under cold (CLC), and plumule length recovery (PLR). The solid blue line shows the p-value 0.001 significant threshold; while the solid green line shows the Bonferroni correction.

from the five cold indices (*qLTG-1-1*, *qGI-1-1*, *qCLC-1-2*, *qPLR-1*, and *qPLRR-1*) shared the same SNP peak marker at position 29.9 Mb on chromosome 1; four QTLs from the four indices (*qLTG-5-2*, *qCLC5*, *qGI5-2*, and *qPLR5*) share the same SNP peak at 7.2 Mb on chromosome 5; four other QTLs also shared with the same SNP peak at 14.6 Mb, where 3 QTLs were from all accessions (*qGI-9-1*, *qPLR-9-2*, *qPLRR-9*) and the other one was from the *indica* subspecies (*qInPLR-9*); another four QTLs shared a very closely linked region between 17.7–117.8 Mb on chromosome 3, where two of them were from all panel (*qLTG-3-1* and *qGI-3-1*) and the other two were from the *japonica* group (*qJaLTG-3* and *qJaGI-3*); a QTL from the whole set (*qPLRR-6-1*) shared a similar region with two other QTLs from *japonica* (*qJaLTG-6-2* and *qJaGI-6-3*) at position 17.1–17.8 Mb, three QTLs (*qLTG-10*, *qGI-10*, and *qPLR-10*) shared the same SNP peak at 13.9 Mb on chromosome 10; while another three QTLs also shared the same SNP peak at 24.8 Mb where 2 of them were from the whole set (*qCLC-11-1* and *qPLR-11*) and the other one was from the *indica* group (*qInPLR11*); the rest of the QTL clusters consisted of two QTLs located on chromosomes 1, 2, 3, 4, 5, 9, 11, and 12.

GWAS of Chilling Tolerance Indices for *japonica*

At cut-off p-value of < 0.001, we identified 20 QTLs associated with the chilling tolerance indices in *japonica* subspecies, 2 of the SNPs were detected at FDR < 0.1 (**Supplementary Table 2**; **Supplementary Figure 2**). Among the 20 QTLs, we identified 9, 8, and 3 QTLs associated with LTG, GI, and CLC, respectively. The phenotypic variance explained by the QTLs were in the range of 12.91% to 21.07% for LTG, 12.61% to 20.32% for GI, and 15.79% to 20.54% for CLC. Considering the reoccurring SNPs in multiple chilling indices and linked SNPs, we identified 12 unique QTL regions. Among these 12 QTL regions, only 3 of them contained a single QTL; while the rest harbored at least 2 QTLs. For example, three of the QTL regions were shared by QTLs from the whole panel, two regions were on chromosomes 3 and 6 as mentioned above and the other region was at position 9.2 Mb on chromosome 9, which shared by 2 QTLs (*qCLC-9* and *qJaCLC-9*). A QTL from the *japonica* group (*qJaCLC-2*) also shared a similar region at 0.6 Mb to 0.9 Mb on chromosome 2 with a QTL from the *indica* (*qInPLR-2*); the rest of the QTL regions contained two QTLs each identified on chromosomes 1, 5, 6, 8, and 12.

GWAS of Chilling Tolerance Indices for *indica*

At cut-off p-value of < 0.001, we identified 9 QTLs associated with the chilling tolerance indices in *indica* subspecies. Among the nine QTLs, two QTLs each were found to be associated with LTG, GI, CLC, and three QTLs were with PLR (**Supplementary Table 2**; **Supplementary Figure 3**). Considering the reoccurring SNPs or closely linked SNPs, we identified seven unique QTL regions in *indica* subspecies. Among these regions, 4 of them harbored more than one QTL: a QTL on chromosome 9 (*qInPLR-9*) shared the same SNP peak at position of 14.6 Mb with three other QTLs detected from all accessions as mentioned above; another QTL on chromosome 11 (*qInPLR-11*) shared a SNP peak at position 24.8 Mb with two other QTLs from the whole set (*qCLC-11-1* and *qPLR-11*); q*InCLC-11* shared a SNP peak at 25.6 Mb with *qCLC-11-2* on chromosome 11; and *qIn-PLR-2* shared a closely linked SNP peaks on chromosome 2 as mentioned above; a SNP peak at 1.4 Mb on chromosome 6 was shared by *qInLTG-6* and *qInGI-6*; another peak SNP at position 20.8 Mb on chromosome 7 was shared by *qInLTG-7* and *qInGI-7*. The phenotypic variance explained by the significant SNPs were in the range of 9.96% to 11.38% for LTG, 9.78% to 12.03% for GI and 12.51% to 16.2% for CLC, and 8.85% to 10.75% for PLR, respectively.

Candidate Gene and QTL Comparisons

Among the 31 unique QTL regions ($p < 0.001$) associated with chilling indices of the whole set of accessions, we identify 10 loci potentially co-localized with the previously identified genes/QTLs related to cold stress in rice, including cold tolerance during germination, seedling and reproductive stage, and cold recovery (**Table 4**; **Supplementary Table 2**).

A QTL associated with GI located on chromosome 5, *qGI-5-3*, was identified to be positioned at 127.9 kb away from the *OsRAN2* gene (Os05g0574500) previously reported to be responsible for cell

division in cold condition (Chen et al., 2011). Another QTL for GI, *qGI-7* was found to be 141.6 kb away of *Omega-3 fatty acid desaturase* (Os07g0693800) and *qCTS7-5* which were reported to be responsible for cold tolerance at seedling stage (Wang et al., 2016). A few QTLs on chromosome 9 at around significant peak at 14.6 Mb which were significantly associated with GI, PLR (whole panel and *indica*), and PLRR, located in the vicinity of *WRKY transcription factor* (Os09g0417600) previously reported to cause increasing tolerance to cold stress in rice (Yokotani et al., 2013) and *qPGCG9-2* which was previously reported as a QTL controlling plumule growth recovery rate under cold stress during seedling stage (Schläppi et al., 2017), and 117 kb away from *OsWRKY76*, a gene similar to BRI1-KD interacting protein 120 (Os09g0410300) related to cold tolerance and *qCTS-9* previously reported related to tolerance during seedling stage (Peng et al., 2010). A SNP peak at 9.2 Mb on chromosome 9 for CLC detected by the whole set of accessions and the *japonica* panel, *qCLC-9* and *qJaCLC-9* were positioned at a distance of 169.5 kb away from *qLTSS9-1*, a QTL responsible for cold tolerance at seedling stage (Schläppi et al., 2017). A QTL associated with PLR, *qPLR-9-1* found to be 21.39 kb away from *qCTS9-2* discovered to be associated with seedling growth under cold stress (Wang et al., 2016). A SNP peak at position 24.9 on chromosome 11 associated with CLC and PLR of the full set and PLR of the *indica* was potentially colocalized with *qCTS11-9* previously reported to be responsible for cold tolerance during seedling growth (Wang et al., 2016). Similarly, another SNP peak at position 25.1 Mb on chromosome 11 associated with CLC of the full set and *indica* was 159 kb away from *qCTS11-10*, a QTL responsible for cold tolerance in seedling stage (Wang et al., 2016).

Several of our reported QTLs are found to be located in close vicinity of previously reported QTLs controlling for cold tolerance at reproductive stage in rice. For examples, a SNP peak at position 26.2 Mb on chromosome 2 associated with both PLR and PLRR was found at a distance of 109.7 kb away from the previously reported QTLs *qSWTPNCT2-2* and *qnob-5* (Shakiba et al., 2017); *qCLC-7* was identified at 135.7 kb away from *qSWTCT7* (Shakiba et al., 2017); a SNP peak at 16.3 Mb on chromosome 9 associated with GI and PLR was 214.46 kb away from a QTL for cold tolerance at reproductive stage *qSWTCT9* (Shakiba et al., 2017) and 143.79 kb away from *qCTS9-8*, a previously identified QTL for cold tolerance at seedling stage (Wang et al., 2016).

Among the 13 unique QTL regions in *japonica*, we found four GWAS sites potentially colocalized with previously identified QTLs/genes. A QTL, *qJaCLC-1-1* associated with CLC in *japonica* subspecies was found to be 140 kb distance away from *qCTS1-5*, a QTL previously reported to be responsible for cold stress tolerance in the seedling stage in rice (Wang et al., 2016) and 242.2 kb away from *qCTGERM1-8*, a QTL controlling cold stress tolerance in germination stage (Shakiba et al., 2017). Another QTL, *qJaGI-6-1* was at distance of 196.93 kb away from a QTL, *qCTS6-2* previously reported to be responsible for cold tolerance at seedling stage (Wang et al., 2016). A QTL for CLC, *qJaCLC-9* was found to be potentially collocated with a QTL, *qLTSS9-1* previously reported to be controlling for cold tolerance during seedling stage in rice (Schläppi et al., 2017). A QTL on chromosome 9, *qJaLTG-11* was identified to be potentially collocated with a QTL, *qCTGERM11-1*, previously reported to be controlling for cold tolerance during germination stage (Shakiba et al., 2017).

Among the seven unique QTL regions identified in *indica* subspecies, 6 of them were found to be potentially colocalized with previously identified genes/QTLs. A significant SNP on chromosome 6 at position 1.4 Mb was associated with LTG and GI in our study was only 10.51 kb away from *OsDREB1C* (Os06g0127100), which was reported to be associated with cold, drought and stress tolerance in rice (Ito et al., 2006). The SNP peak at 20.8 Mb on chromosome 7 which shared by *qInLTG-7* and *qInGI-7* was potentially collocated with *OsFAD9*, *FAD8* (Os07g0693800), and *qCTS7*-5, which were previously reported to be controlling for cold tolerance in the seedling stage in rice (Wang et al., 2016). A QTL, *qInCLC-8* on chromosome 8 at position 10.36 Mb was found to be potentially colocalized with *qCTGERM8-1*, and *qCTS8-2*, which were responsible for cold tolerance during germination and seedling stage in rice (Wang et al., 2016; Shakiba et al., 2017). The QTL regions on chromosome 9 and 11 that were also shared with the cold indices of the whole set have been discussed above.

DISCUSSION

It has been a challenge to map loci associated with abiotic stress tolerance traits like cold tolerance due to the polygenic nature of the loci (Shakiba et al., 2017). The separate GWAS analysis of low-temperature germinability (LTG) and germination index (GI) helped us to discover whether the chilling tolerance was due to the inherent cold tolerance ability or due to high seedling vigor. Moreover, the plumule length traits (PLR and PLRR assays) helped us to determine if there is a quantitative effect on subsequent growth and development of seedlings after a recovery period at normal temperature. These assays are important to measure, as some accessions with good LTG indices did not grow well after a temperature shift to 30°C and *vice versa*. All of these assays may address a realistic scenario in direct-seeding method of rice cultivation where germinating rice seeds or young seedlings may get exposed to warm-growth promoting temperature after an extended period of cold exposure.

The inbred lines developed at the Beaumont Research Center that were used in our study generally had a good level of tolerance under cold stress during germination, including the recovery phase. On the other hand, the *aus* sub-population had the lowest value of CLC while *O. glaberrima* species had the lowest values of GI, PLR, and PLRR demonstrating that these groups are not good sources of cold-tolerant genes. However, our sample size representing *O. glaberrima* might be too small; therefore, research focusing on this species with a greater number of samples is needed to have more conclusive results. Accessions belonging to the NERICA lines are found to have good GI, CLC, PLR, and PLRR. The *aromatic* and *aus* groups were found to have low tolerance to cold stress indicated by the low values of different chilling indices which were similar to the findings of other studies (Schläppi et al., 2017; Shakiba et al., 2017).

Interestingly, we didn't find a significant difference in LTG between highly tolerant *temperate japonica* and *indica* groups. As the LTG values observed were relatively similar between different sub-populations, there is a chance that both *indica* and *japonica* subspecies may carry the alleles contributing to superior LTG abilities. This also shows that there are many accessions of *indica* species which have good germination under cold stress. This is in agreement with the recent findings of Shakiba et al. (2017) where they had identified *indica* specific LTG QTL and have reported that both *indica* and *japonica* sub-species are expected to have alleles contributing to superior LTG abilities. On the contrary, we observed significantly lower values of CLC, PLR, and PLRR of the *indica* group than the *temperate japonica*. These findings showed that although the *indica* group has good germination ability under cold stress conditions, their growth rapidly gets retarded under cold condition.

The results of our study showed that the Texas breeding lines, *temperate japonica* and *tropical japonica*, were more tolerant of cold stress whereas *aus*, *aromatic*, and *indica* lines were more susceptible to cold conditions. The phenotypic measurement of different chilling indices revealed that *japonica* subspecies were generally more tolerant than *indica*. This finding is consistent with previous findings (Cui et al., 2002; Morsy et al., 2005; Lv et al., 2016). This could be because in general *indica* accessions are more adapted to higher temperature regions of low latitude while *japonica* accessions are more adapted to lower temperature regions of a higher latitude and higher elevations. This history of adaptation between *indica* and *japonica* accessions is also reflected by genes having a ratio of nonsynonymous vs synonymous substitution rates (Ka/Ks ratio) greater than 1.0, which indicates positive selection, as shown by a study between the *indica* rice 9311 and Nipponbare (Sun et al., 2015). A comparison of the QTLs in our study having FDR < 0.05 with the list of 3,340 genes with Ks of zero and Ka above zero (**Table S1** from Sun et al., 2015) revealed 5 out of 13 cold-tolerance loci in our study contain genes under selection between *indica*/*japonica* within 250 kb, including a match within 7 kb of our QTL cluster at 29.9 Mb on chromosome 1 (data not shown). Although this may be suggestive that *indica*/*japonica* alleles at some cold tolerance loci may have been under selection, further analysis would be needed to validate these results. In any case, the presence of differences in the genetic architecture of cold tolerance among different subspecies and sub-populations analyzed in this study provides opportunities for enhancing cold tolerance through molecular breeding.

Spearman's correlation analysis showed that all the indices were highly correlated with each other. Likewise, the heritability values of all the traits were also high. Schlappi et al. (2017) had reported that low-temperature germination (LTG) and plumule growth recovery rate (PLRR) were not correlated or weakly correlated with other indices while PLR was highly correlated with other indices. In our study, however, we found a high correlation between LTG with all other measured indices, albeit with different levels. Partly, this could be attributed to the differences in tolerance ability of the different accessions used in both studies. We also observed some significant loci detected in either *japonica* or *indica* were also observed in the whole set. This indicates that the significant SNPs detected in the whole set might come from that particular subspecies.

We observed seven SNP peaks/QTL regions that were shared between LTG and GI (**Table 4**; **Supplementary Table 2**). This is in agreement with the correlation analysis where a highly positive correlation was observed between LTG and GI (0.97), since GI is largely derived from LTG, especially for lines with similar levels of germination under normal conditions. This result also shows that the significant associations discovered from GI are mostly due to the tolerance of the accessions to cold germination and not due to the seedling vigor. There were three SNP peaks/QTL regions associated with both CLC and PLR, this may indicate that there may be some similar genetic mechanism or overlapping mechanisms underlying coleoptile length growth at low temperature and plumule recovery after cold stress exposure. Three of the significant SNPs associated with PLR were found to be associated with PLRR. This is in agreement with the highly positive correlation analysis of chilling indices PLR and PLRR (0.96). This further suggests that LTG and GI, PLR and PLRR may share some common genetic mechanisms. Fine mapping and ultimately cloning of the responsible genes could be performed to confirm whether the overlapping QTLs associated with one or more genetic factors.

Some of the significant SNPs identified from our GWAS study were located within the LD regions of known cold tolerance genes or previously reported QTLs, including 10 in the whole panel, 4 in *japonica*, and 6 in *indica*. In addition to validating our GWAS results, many of the identified QTLs near the previously mapped chilling tolerance related genes in rice help us to narrow down the QTL region and provide further support of the location of the underlying genes. Among the most interesting regions identified were near those QTLs which were found to be located very close to the genes involved in cold stress tolerance, including *OsDREB1C* (10.51 kb) and *OSWRKY76* (117 kb).

In summary, our novel diversity panel has little overlap with previously studied rice diversity panels, including the RDP1/RDP2, USDA Rice Mini-Core, and the 3,000 Rice Genomes panel, which may lead to the discovery of additional novel genetic loci for cold tolerance in rice. The GWAS QTLs detected in our study may provide additional information on the genetic structure of cold tolerance and recovery during germination in rice. In the future, some selected QTLs can be targeted for further molecular studies to better understand the mechanisms underlying cold tolerance and recovery of germinating rice seeds. Some selected cold tolerance-associated SNP markers can also potentially be used for MAS in rice improvement efforts. Further, a set of new highly tolerant rice accessions can potentially be used as novel donors for further genetic studies and crop improvement programs.

DATA AVAILABILITY STATEMENT

The genotyping data has been deposited at Dryad (https://doi.org/10.5061/dryad.q83bk3jdt).

AUTHOR CONTRIBUTIONS

RT and ES designed the experiment. ES conceived the project. MT, ES, and RET developed the rice panel. RET and RT performed seed multiplication and post-harvest processes. RT performed the experiment. RT and ES analyzed data and wrote the manuscript. MT and RET edited the manuscript. All read and approved the manuscript.

FUNDING

This work was supported by the Texas A&M AgriLife Research and the USDA NIFA Hatch projects # 1009299 and 1009300.

ACKNOWLEDGMENTS

We thank Chersty Harper and the members of the Rice Genetics Lab of 2016/2017 for technical assistance on seed multiplication and post-harvest processes; Susan McCouch at Cornell University for developing the 7K SNP chip; Karina Morales for arranging the leaf tissue shipment for genotyping; and Eurofins for genotyping the rice panel.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00022/full#supplementary-material

REFERENCES

Bai, X., Zhao, H., Huang, Y., Xie, W., Han, Z., Zhang, B., et al. (2016). Genome-wide association analysis reveals different genetic control in panicle architecture between *indica* and *japonica* rice. *Plant Genome* 9, 1–10. doi: 10.3835/plantgenome2015.11.0115

Baruah, A. R., Ishigo-Oka, N., Adachi, M., Oguma, Y., Tokizono, Y., Onishi, K., et al. (2009). Cold tolerance at the early growth stage in wild and cultivated rice. *Euphytica* 165, 459–470. doi: 10.1007/s10681-008-9753-y

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Cruz, R. P. D., and Milach, S. C. K. (2004). Cold tolerance at the germination stage of rice: methods of evaluation and characterization of genotypes. *Sci. Agr.* 61, 1–8. doi: 10.1590/S0103-90162004000100001

Chen, N., Xu, Y., Wang, X., Du, C., Du, J., Yuan, M., et al. (2011). OsRAN2, essential for mitosis, enhances cold tolerance in rice by promoting export of intranuclear tubulin and maintaining cell division under cold stress. *Plant Cell Environ.* 34, 52–64. doi: 10.1111/j.1365-3040.2010.02225.x

Cruz, R. P. D., Sperotto, R. A., Cargnelutti, D., Adamski, J. M., De Freitasterra, T., and Fett, J. P. (2013). Avoiding damage and achieving cold tolerance in rice plants. *Food Energy Secur.* 2, 96–119. doi: 10.1002/fes3.25

Cui, K., Peng, S., Xing, Y., Xu, C., Yu, S., and Zhang, Q. (2002). Molecular dissection of seedling-vigor and associated physiological traits in rice. *Theor. Appl. Genet.* 105, 745–753. doi: 10.1007/s00122-002-0908-2

Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Res.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Fujino, K., Sekiguchi, H., Matsuda, Y., Sugimoto, K., Ono, K., and Yano, M. (2008). Molecular identification of a major quantitative trait locus, qLTG3–1, controlling low-temperature germinability in rice. *Proc. Natl. Acad. Sci. U. S. A.* 105, 12623–12628. doi: 10.1073/pnas.0805303105

Fujino, K., Obara, M., Shimizu, T., Koyanagi, K. O., and Ikegaya, T. (2015). Genome-wide association mapping focusing on a rice population derived from rice breeding programs in a region. *Breed. Sci.* 65, 403–410. doi: 10.1270/jsbbs.65.403

Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S., and McCouch, S. (2005). Genetic structure and diversity in *Oryza sativa* L. *Genetics* 169, 1631–1638. doi: 10.1534/genetics.104.035642

Hsu, S.-K., and Tung, C.-W. (2015). Genetic mapping of anaerobic germination-associated QTLs controlling coleoptile elongation in rice. *Rice* 8, 38. doi: 10.1186/s12284-015-0072-3

Ito, Y., Katsura, K., Maruyama, K., Taji, T., Kobayashi, M., Seki, M., et al. (2006). Functional analysis of rice DREB1/CBF-type transcription factors involved in cold-responsive gene expression in transgenic rice. *Plant Cell Physiol.* 47, 141–153. doi: 10.1093/pcp/pci230

Ji, S., Jiang, L., Wang, Y., Zhang, W., Liu, X., Liu, S., et al. (2009). Quantitative trait loci mapping and stability for low temperature germination ability of rice. *Plant Breed.* 128, 387–392. doi: 10.1111/j.1439-0523.2008.01533.x

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *J. Stat. Software* 69, 1–33. doi: 10.18637/jss.v069.i01

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, W., Lu, T., Li, Y., Pan, X., Duan, Y., Min, J., et al. (2015). Mapping of quantitative trait loci for cold tolerance at the early seedling stage in landrace rice Xiang 743. *Euphytica* 201, 401–409. doi: 10.1007/s10681-014-1227-9

Lv, Y., Guo, Z., Li, X., Ye, H., Li, X., and Xiong, L. (2016). New insights into the genetic basis of natural chilling and cold shock tolerance in rice by genome-wide association analysis. *Plant Cell Environ.* 39, 556–570. doi: 10.1111/pce.12635

Ma, Y., Dai, X., Xu, Y., Luo, W., Zheng, X., Zeng, D., et al. (2015). COLD1 confers chilling tolerance in rice. *Cell* 160, 1209–1221. doi: 10.1016/j.cell.2015.01.046

Mackill, D. J., and Lei, X. (1997). Genetic variation for traits related to temperate adaptation of rice cultivars. *Crop Sci.* 37, 1340–1346. doi: 10.2135/cropsci1997.0011183X003700040051x

Mao, D., and Chen, C. (2012). Colinearity and similar expression pattern of rice DREB1s reveal their functional conservation in the cold-responsive pathway. *PloS One* 7, e47275. doi: 10.1371/journal.pone.0047275

Morsy, M. R., Almutairi, A. M., Gibbons, J., Yun, S. J., and Benildo, G. (2005). The OsLti6 genes encoding low-molecular-weight membrane proteins are differentially expressed in rice cultivars with contrasting sensitivity to low temperature. *Gene* 344, 171–180. doi: 10.1016/j.gene.2004.09.033

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell.* 21, 2194–2202. doi: 10.1105/tpc.109.068437

Pan, Y., Zhang, H., Zhang, D., Li, J., Xiong, H., Yu, J., et al. (2015). Genetic analysis of cold tolerance at the germination and booting stages in rice by association mapping. *PloS One* 10, e0120590. doi: 10.1371/journal.pone.0120590

Peng, Y., Bartley, L. E., Canlas, P., and Ronald, P. C. (2010). OsWRKY IIa transcription factors modulate rice innate immunity. *Rice* 3, 36–42. doi: 10.1007/s12284-010-9039-6

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Raj, A., Stephens, M., and Pritchard, J. K. (2014). Variational inference of population structure in large SNP datasets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Sales, E., Viruel, J., Domingo, C., and Marqués, L. (2017). Genome wide association analysis of cold tolerance at germination in temperate *japonica* rice (Oryza sativa L.) varieties. *PloS One* 12, e0183416. doi: 10.1371/journal.pone.0183416

Schläppi, M. R., Jackson, A. K., Eizenga, G. C., Wang, A., Chu, C., Shi, Y., et al. (2017). Assessment of five chilling tolerance traits and GWAS mapping in rice using the USDA Mini-Core collection. *Front. Plant Sci.* 8, 957. doi: 10.3389/fpls.2017.00957

Shakiba, E., Edwards, J. D., Jodari, F., Duke, S. E., Baldo, A. M., Korniliev, P., et al. (2017). Genetic architecture of cold tolerance in rice (*Oryza sativa*) determined through high resolution genome-wide analysis. *PloS One* 12, e0172133. doi: 10.1371/journal.pone.0172133

Singh, M., Ceccarelli, S., and Hamblin, J. (1993). Estimation of heritability from varietal trials data. *Theor. Appl. Genet.* 86, 437–441. doi: 10.1007/BF00838558

Sun, X., Jia, Q., Guo, Y., Zheng, X., and Liang, K. (2015). Whole-genome analysis revealed the positively selected genes during the differentiation of *indica* and *temperate japonica* rice. *PloS One* 10 (3), e0119239. doi: 10.1371/journal.pone.0119239

Thomson, M. J., Singh, N., Dwiyanti, M. S., Wang, D. R., Wright, M. H., Perez, F. A., et al. (2017). Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. *Rice* 10, 40. doi: 10.1186/s12284-017-0181-2

Tung, C.-W., Zhao, K., Wright, M. H., Ali, M. L., Jung, J., Kimball, J., et al. (2010). Development of a research platform for dissecting phenotype–genotype associations in rice (*Oryza* spp.). *Rice* 3, 205–217. doi: 10.1007/s12284-010-9056-5

Wang, D., Liu, J., Li, C., Kang, H., Wang, Y., Tan, X., et al. (2016). Genome-wide association mapping of cold tolerance genes at the seedling stage in rice. *Rice* 9, 61. doi: 10.1186/s12284-016-0133-2

Yokotani, N., Sato, Y., Tanabe, S., Chujo, T., Shimizu, T., Okada, K., et al. (2013). WRKY76 is a rice transcriptional repressor playing opposite roles in blast disease resistance and cold stress tolerance. *J. Exp. Bot.* 6, 5085–5097. doi: 10.1093/jxb/ert298

Yonemaru, J.-I., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K., and Yano, M. (2010). Q-TARO: QTL annotation rice online database. *Rice* 3, 194–203. doi: 10.1007/s12284-010-9041-z

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355. doi: 10.1038/ng.546

Zhang, F., Ma, X.-F., Gao, Y.-M., Hao, X.-B., and Li, Z.-K. (2014a). Genome-wide response to selection and genetic basis of cold tolerance in rice (Oryza sativa L.). *BMC Genet.* 15, 55. doi: 10.1186/1471-2156-15-55

Zhang, Q., Chen, Q., Wang, S., Hong, Y., and Wang, Z. (2014b). Rice and cold stress: methods for its evaluation and summary of cold tolerance-related quantitative trait loci. *Rice* 7, 24. doi: 10.1186/s12284-014-0024-3

Zhang, S., Zheng, J., Liu, B., Peng, S., Leung, H., Zhao, J., et al. (2014c). Identification of QTLs for cold tolerance at seedling stage in rice (Oryza sativa L.) using two distinct methods of cold treatment. *Euphytica* 195, 95–104. doi: 10.1007/s10681-013-0977-0

Zhao, J., Zhang, S., Dong, J., Yang, T., Mao, X., Liu, Q., et al. (2017). A novel functional gene associated with cold tolerance at the seedling stage in rice. *Plant Biotechnol. J.* 15, 1141–1148. doi: 10.1111/pbi.12704

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Thapa, Tabien, Thomson and Septiningsih. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

MINI REVIEW
published: 28 February 2020
doi: 10.3389/fgene.2019.01389

Genome-Wide Association Studies and Genomic Selection in Pearl Millet: Advances and Prospects

Rakesh K. Srivastava[1], Ram B. Singh[1], Vijaya Lakshmi Pujarula[1], Srikanth Bollam[1], Madhu Pusuluri[1], Tara Satyavathi Chellapilla[2], Rattan S. Yadav[3] and Rajeev Gupta[1]**

[1] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India, [2] All India Coordinated Research Project on Pearl Millet (AICRP-PM), Indian Council of Agricultural Research (ICAR), Jodhpur, India, [3] Institute of Biological, Environmental & Rural Sciences (IBERS), Aberystwyth University, Gogerddan, United Kingdom

OPEN ACCESS

Edited by:
Nunzio D'Agostino,
Università degli Studi di Napoli Federico II, Italy

Reviewed by:
Cheng Sun,
Chinese Academy of Agricultural Sciences, China
Vandana Jaiswal,
Institute of Himalayan Bioresource Technology (CSIR), India
Zhenbin Hu,
Kansas State University, United States

***Correspondence:**
Rakesh K. Srivastava
r.k.srivastava@cgiar.org
Rajeev Gupta
g.rajeev@cgiar.org

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 20 September 2019
Accepted: 19 December 2019
Published: 28 February 2020

Citation:
Srivastava RK, Singh RB, Pujarula VL, Bollam S, Pusuluri M, Chellapilla TS, Yadav RS and Gupta R (2020) Genome-Wide Association Studies and Genomic Selection in Pearl Millet: Advances and Prospects. Front. Genet. 10:1389. doi: 10.3389/fgene.2019.01389

Pearl millet is a climate-resilient, drought-tolerant crop capable of growing in marginal environments of arid and semi-arid regions globally. Pearl millet is a staple food for more than 90 million people living in poverty and can address the triple burden of malnutrition substantially. It remained a neglected crop until the turn of the 21st century, and much emphasis has been placed since then on the development of various genetic and genomic resources for whole-genome scan studies, such as the genome-wide association studies (GWAS) and genomic selection (GS). This was facilitated by the advent of sequencing-based genotyping, such as genotyping-by-sequencing (GBS), RAD-sequencing, and whole-genome re-sequencing (WGRS) in pearl millet. To carry out GWAS and GS, a world association mapping panel called the Pearl Millet inbred Germplasm Association Panel (PMiGAP) was developed at ICRISAT in partnership with Aberystwyth University. This panel consisted of germplasm lines, landraces, and breeding lines from 27 countries and was re-sequenced using the WGRS approach. It has a repository of circa 29 million genome-wide SNPs. PMiGAP has been used to map traits related to drought tolerance, grain Fe and Zn content, nitrogen use efficiency, components of endosperm starch, grain yield, etc. Genomic selection in pearl millet was jump-started recently by WGRS, RAD, and tGBS (tunable genotyping-by-sequencing) approaches for the PMiGAP and hybrid parental lines. Using multi-environment phenotyping of various training populations, initial attempts have been made to develop genomic selection models. This mini review discusses advances and prospects in GWAS and GS for pearl millet.

Keywords: pearl millet, genetic resources, genomic resources, genomic selection, genome-wide association studies, molecular markers

INTRODUCTION

Pearl millet (*Pennisetum glaucum* (L) R. Br., syn. *Cenchrus americanus* (L.) Morrone) is an important C_4 small-grained field crop of traditional smallholder farming systems that belongs to the grass family *Poaceae* and subfamily Panicoideae. An archaeological survey indicates that pearl millet was initially domesticated at the southern edge of the Sahara Desert in West Africa about 2500 BC (Manning

et al., 2011). Pearl millet is a diploid ($2n = 2x = 14$), cross-pollinated warm-season crop with tremendous photosynthetic potential and high biomass production capacity. It is highly tillering, polymorphic, has a short life cycle, a large genome size (1.76 Gb), and an outbreeding nature (Bennett et al., 2000, Varshney et al., 2017). Climate-adaptive phenotypic, physiological, and reproductive attributes of pearl millet make this crop well-suited to grow in marginal conditions, such as poor soil fertility, limited soil water content, high salinity, extreme soil pH ranges, high soil Al^{3+} saturation, high temperatures, and scant rainfall. Pearl millet can thrive and produce a substantial amount of grain in drought-prone areas that receiving average annual precipitation <250 mm, whereas other cereal crops, such as maize, rice, sorghum, bread wheat, and barley, are likely to fail to give economic returns (Nambiar et al., 2011). Pearl millet is cultivated over ~27 million hectares in arid and semi-arid areas of Asia and Sub-Saharan Africa and is the primary food source for about 90 million resource-poor populations residing in marginal areas globally. Remarkably, the natural attribute of this crop to withstand ambient temperatures up to 42°C at the reproductive phase makes it suited for growth *via* irrigation in the extremely hot summers in north-western parts of India (Gupta et al., 2015).

Pearl millet has several nutritional properties compared to other staple cereal grains, and it is an excellent source of organic as well as inorganic nutrients and a cost-effective source of energy (Kumar et al., 2016). Pearl millet grains are rich in fibers (1.2 g/100 g), α-amylose, amino acids, proteins (8–19%), and low starch, mineral nutrients including phosphorus, magnesium, iron, and zinc. Owing to having such nutritional values, pearl millet ensures food and nutritional security for farmers living in poverty (Nambiar et al., 2011, Kanatti et al., 2014). Pearl millet is a rich source of several polyphenols, and other biologically important ingredients make it suited to play a role in reducing the rate of fat absorption, the lowering of glycemic indices, as well as in overcoming the risk of cardiac diseases, diabetes, and other medical problems. Overall, pearl millet has the capacity to combat micronutrient deficiency across developing countries (Rai et al., 2012) since it contributes 30–40% of inorganic nutrients and provides affordable staple food with an adequate level of iron and zinc in its cultivating areas (Rao et al., 2006).

An alternative approach to the QTL mapping is the genome-wide association study (GWAS) or association mapping (AM) approach (Gómez et al., 2011) based on the principle of a linkage disequilibrium (LD) to detect a substantial association between DNA marker and target trait (Gupta et al., 2005). Genetic linkage is found through extensive genotyping of a panel of germplasm or breeding populations showing contrasting phenotypes across variable environments. It has an immense power in identifying specific genes controlling the expression of the desired traits (Kraakman et al., 2004). The potential advantage of association mapping is the likelihood of a superior resolution mapping utilizing mass recombination events from numerous meiotic events throughout the germplasm evolutionary history. It has the power to evaluate and characterize several alleles concurrently in diploid (Zhao et al., 2007) as well as in polyploid crops (Breseghello and Sorrells, 2006). Association mapping offers many benefits over linkage mapping since it provides better mapping resolution due to historical mutations and recombinations in genetic lineages, which leads to the identification of markers in the vicinity of governing genes (Liu et al., 2016). Genetic polymorphisms having strong linkage with a genomic locus leading to phenotypic differences is expected to be substantially associated with a target trait across the panel of germplasm.

The analysis of QTL effects for minor QTLs using linkage mapping and genome-wide association mapping is often biased. Therefore, scientific groups have for years been trying to solve the issue of how to tackle these complex traits and outcomes in terms of genomic selection (GS). Genomic selection is a breeding approach exploiting high-density DNA markers distributed across the genome to facilitate the rapid selection of the best candidates and offers opportunities to enhance genetic gains (Meuwissen et al., 2001). GS uses different prediction models by combining the genotyping and phenotyping datasets of the training population (TP), which is subsequently used to determine genomic-estimated breeding values (GEBVs) for every genotype of breeding population (BP) from their genotyping scores. These GEBVs permit breeders to envisage superior genotypes that would be suitable either as a parent in hybridization or for next-generation advancement of the breeding program. The basic principle is that the information derived from several markers widely distributed over the genome, having the potential to reveal genetic variations in the genome, can evaluate breeding values without prior information of where the selected genes are located (Crossa et al., 2017).

In this paper, we review the advances made in the development of genetic and genomic resources for their use in genome-wide association studies (GWAS) and genomic selection (GS) in pearl millet.

DEVELOPMENT OF GENETIC RESOURCES

Genetic resources are the fundamental materials that play a pivotal role in plant genomic and phenomic studies to boost major scientific discoveries in advanced agriculture systems. Fortunately, genetic recourses have been collected and preserved by many national and international gene banks around the world. Pearl millet accessions have been collected and conserved by 97 gene banks (66,682 accessions) globally, in which ICRISAT has the largest collection (~21,594 pearl millet accessions from 51countries) (Singh and Upadhyaya, 2016). More importantly, core and mini core collections have been developed at ICRISAT and serve as essential resources for allele mining studies for the identification of agronomic studies, and they are also used for the development of tolerant lines for both abiotic and biotic stresses. Likewise, one more genotype-based reference set has been developed, and it comprises 300 pearl millet accessions (Upadhyaya et al., 2011). At ICRISAT, most of the accessions were evaluated for several agronomical traits, and these show the extent of genetic diversity and phenotypic variance for most of the qualitative and quantitative traits (Singh and Upadhyaya 2016). It is evident that vast genetic variability is the determining factor for the

identification of promising germplasm for the desired trait (Upadhyaya et al., 2007). In addition to ICRISAT, major germplasm are preserved at the Institute of Research for Development (IRD, France), in which 3,968 accessions are maintained from 16 countries, and 3,821 accessions of cultivated *P. glaucum* and related species maintained at the Canadian Genetic Resources (Saskatoon, Canada). Additionally, there are 1,283 active collections of pearl millet accessions collected and preserved at the US Germplasm Resource Information Network (GRIN) (Yadav et al., 2007). For conducting AM studies, diverse genetic resources are the essential inputs, and pearl millet genetic resources are found to have enormous genetic diversity. For this reason, performing AM studies for desired traits in pearl millet crops is imperative and will provide immense genomic resources for future studies. Over the last five years, significant work has been carried out on pearl millet related AM studies, and this gives information about genetic diversity and linkage disequilibrium (LD). To get over this problem, ICRISAT, in association with AU, developed a world association mapping panel called the Pearl Millet inbred Germplasm Association Panel (PMiGAP). This panel comprises 346 lines consisting of germplasm lines, landraces, and breeding lines representing global pearl millet diversity. These lines were generated by repeated rounds of selfing (S_0 through S_{11}) from 1,000 accessions representing diverse cultivars, landraces, and mapping population parents of 27 countries. Thus, PMiGAP may be considered an excellent genetic resource for GWAS studies into pearl millet crop. By the year 2015, out of 346 PMiGAP lines, Sehgal used 250 lines for AM studies and evaluated these for drought-related traits under field conditions. Similarly, during another study on AM, in which 500 pearl millet lines included 252 global accessions and 248 Senegalese landraces, they found extant genetic diversity between global and Senegalese accessions (Hu et al., 2015). In addition to the above studies, several RIL (recombinant inbred line) populations were also developed for biotic and abiotic stresses, quality, as well as yield and yield-related traits. Rajaram et al. (2013) constructed pearl millet consensus maps by using four RIL populations (ICMB 841-P3 × 863B-P2 (RIP A), H 77/833-2 × PRLT 2/89-33 (RIP B), 81B-P6 × ICMP 451-P8 (RIP C), and PT 732B-P2 × P1449-2-P1 (RIP D). In other studies, iron- and zinc-related QTLs were identified in ICMB 841-P3 × 863B-P2 (144 progenies) and ICMS 8511-S1-17-2-1-1-B-P03 × AIMP 92901-S1-183-2-2-B-08 (317 progenies) RIL populations, respectively (Kumar et al., 2016; Kumar et al., 2018). In a recent study, Chelpuri et al. (2019) identified QTLs with resistance to major pathotype isolates of the downy mildew pathogen in the pearl millet RIL population, ICMB 89111-P6 × ICMB 90111-P6 (187 progenies). Therefore, there is a good opportunity for pearl millet researchers who can access these useful genetic resources to meet their research needs.

DEVELOPMENT OF GENOMIC RESOURCES AND TRAIT MAPPING

Molecular or DNA-based markers, genetic linkage maps, and genomic sequence data are important genomic resources to perform a genetic evaluation and marker-assisted breeding in any plant species. Over the last decade, several types of molecular markers, genomic tools, and genetic linkage maps have been developed and deployed in millets (Serba and Yadav, 2016). Several DNA-based molecular markers, including restriction fragment length polymorphism [RFLP; (Liu et al., 1994)], amplified fragment length polymorphism [AFLP; (Devos et al., 1995)], random amplified polymorphic DNA (RAPD), expressed sequence tags-derived simple sequence repeats [EST-SSRs; (Senthilvel et al., 2008; Rajaram et al., 2013)] markers, sequence-tagged sites [STSs; (Allouis et al., 2001)], genomic simple sequence repeat [gSSRs; (Qi et al., 2004)], DArT array Technology [DArTs; (Senthilvel et al., 2010; Supriya et al., 2011)], conserved intron specific primers [CISP; (Sehgal et al., 2012)], single-stranded conformation polymorphism-SNP [SSCP-SNP; (Bertin et al., 2005)], and single nucleotide polymorphisms [SNPs; (Sehgal et al., 2012)] have been developed and exploited in genetic diversity, QTLs/genes identification, and marker-aided breeding for faster pearl millet breeding (**Table 1**). Molecular markers facilitate in analyzing genetic variations existed within the germplasm collections for precise selection of breeding parents in crossing programs, estimating population structure, and identification of QTLs for stress tolerance. Pearl millet has a wide range of DNA polymorphisms even in elite inbred parental lines of popular hybrids (Vadez et al., 2012).

Initially, RFLP-derived DNA markers were devised and used to map about 180 loci ranged approximately 350 cM under seven linkage groups in pearl millet (Liu et al., 1992; Liu et al., 1994). Later, these markers were exploited in QTL mapping for downy mildew resistance in pearl millet (Jones et al., 1995). A subset of 21 polymorphic EST-SSRs and 6 genomic SSR markers were developed using sequence information from 3,520 expressed sequence tags (ESTs) and used in genome mapping of different pearl millet mapping populations (Senthilvel et al., 2008). Subsequently, these potentially developed EST-SSRs were deployed in marker-aided breeding for yield and drought stress resistance in pearl millet at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). The development of a panel of 277 polymorphic DArT markers was reported from 6,900 DNA array-dart technology (DArT) clones using a PstI/BanII complexity reduction in a pearl millet RIL population (Senthilvel et al., 2010). Separately, 574 potential DArT markers were detected from 7,000 DArT clones obtained from 95 diverse genotypes using a PstI/BanII complexity reduction in genetically diverse inbred lines of pearl millet (Supriya et al., 2011). The mapping of 208 DArT markers along with 305 SSRs detected seven linkage groups covering 1,749 cM with an average intermarker distance of 5.73 cM and two co-localized QTLs for iron and zinc content on LG 3 were identified in pearl millet (Kumar et al., 2016). Using DArT markers, comparative mapping and genome organization analysis may easily be performed, and the price of marker-aided backcrossing (MABC) is also cheap relative to other markers systems.

Pearl millet EST resources were used to develop quality SNPs and CISP markers, and they were deployed to identify candidate

TABLE 1 | Details of mapped traits and genomic resources developed in pearl millet, related to grain quality, yield, fodder, biomass, and biotic and abiotic stresses.

Sl. No.	Mapped traits	Reference
1.	Reported large-effect Fe and Zn content QTLs using DArT and SSRs markers to construct a genetic linkage map with 317 RIL population developed from ICMS 8511-S1-17-2-1-1-B-P03 × AIMP 92901-S1-183-2-2-B-08 cross.	Kumar et al., 2018
2.	Pearl millet genome sequencing data was used to establish marker trait associations for genomic selection, to define heterotic pools, and to predict hybrid performance.	Varshney et al., 2017
4.	A set of 305 loci were used to construct a linkage map to map two QTLs for grain Fe content on LG3 and LG5 and two QTLs for grain Zn content on LG3 and LG7 using replicated samples of 106 pearl millet RILs (F6) derived from ICMB 841-P3 × 863B-P2 cross.	Kumar et al., 2016
5.	Identified 83,875 SNPs within 500 pearl millet accessions, consisting of 252 accessions and 248 Senegalese landraces, with genotyping by sequencing (GBS) of PstI-MspI reduced representation libraries.	Hu et al., 2015
6.	Thirty-seven SSRs and CSIP markers have been developed, spanning 7 LGs evaluated in irrigated and drought stress conditions,22 SNPs, and 3 InDels for abiotic stresses	Sehgal et al., 2015
7.	ISSR-based SCAR marker has been devised for downy mildew (DM) resistance in pearl millet and associated to DM resistance LG with genetic linkage distance of 0.72 cM	Jogaiah et al., 2014
8.	Seventy-five SNPs and CISP were developed from EST sequences using parents of two mapping populations for 18 genes	Sehgal et al., 2012
9.	Hundreds of polymorphic EST-derived SSRs were developed and deployed in mapping of RIL populations in pearl millet	Rajaram et al., 2010; Rajaram et al., 2013
10.	About 300 DArT markers have been used for the polymorphic in different pearl millet RIL populations	Senthilvel et al., 2010
11.	Cross-transferability of the 31-finger millet EST-SSRs were evaluated and found to be polymorphic in pearl millet	Arya et al., 2009
12.	Four EST-derived SSRs and 9 CISPs were used in linkage mapping using biparental mapping populations of pearl millet	Yadav et al., 2008
	A panel of 21 functionally informative EST-based SSRs and 6 gSSRs were developed in pearl millet	Senthilvel et al., 2008
13.	Nineteen EST-SSRs, among them 11 amplified and 4 were an appeared polymorphism on agarose gels	Yadav et al., 2007
14.	Sixteen EST-based polymorphic SSR markers	Mariac et al., 2006
14.	SSCP-SNP primes were developed through a comparison of rice and pearl millet EST collections	Bertin et al., 2005
15.	Thirty-six genomic SSRs were developed from genomic clones	Qi et al., 2004
16.	Genetic maps developed in four different crosses were integrated to generate a consensus map of 353 RFLP and 65 SSR markers.	Qi et al., 2004
17.	Eighteen potential SSR markers were developed from genomic sequences in pearl millet	Budak et al., 2003; Allouis et al., 2001
18.	RFLP probes were used to assess genetic diversity within and between 504 landraces of core collection using a subset comprising 10 accessions of Indian origin	Bhattacharjee et al., 2002

genes related to a major QTL for drought tolerance using diverse (H 77/833-2, PRLT 2/89-33, ICMR 01029, and ICMR 01004) genotypes that represented mapping populations parents (Sehgal et al. (2012). Later, 83,875 SNP markers were identified using genotyping-by-sequencing (GBS) of *Pst*I-*Msp*I reduced representation libraries in pearl millet lines, represented by 252 world germplasm accessions and 248 landraces from Senegal, which revealed wide genetic variability in comparison to other germplasm collection in Africa and Asia (Hu et al., 2015). Moreover, ISSR-based sequence characterized amplified region (SCAR) markers were devised to examine genetic variations between two (ICMR 01007 and ICMR 01004) genotypes of pearl millet and a contrast mapping population for downy mildew resistance. A polymorphic locus (1.4 kb size) was found in the ICMR 01004 genotype, and further PCR amplification of these polymorphic loci was produced to be closely associated with downy mildew resistant LG with a genetic distance of 0.72 cM. An identified SCAR marker was eventually validated using diverse pearl millet genotypes belonging to Asia and Africa, and the outcomes demonstrate that the marker was linked to downy mildew disease-resistant genotypes only (Jogaiah et al., 2014). The development of a linkage map was reported to integrate 256 DArT markers and 70 SSR markers and used to identify QTLs on LG1 with LOD score of 27 for rust resistance in 168 F_7 pearl millet RILs derived from cross 81B-P6 × ICMP 451-P8 (Ambawat et al., 2016). Using a total of 106 pearl millet RILs (F6) derived from ICMB 841-P3 × 863B-P2 cross and 305 (96 SSRs and 208 DArT) markers, a linkage map was generated to map QTLs for grain iron and zinc content (Kumar et al., 2016). Recently, Kumar et al. (2018) reported a large-effect Fe and Zn content quantitative trait loci (QTLs) linked with DArT and SSR markers to construct a genetic linkage map using 317 RIL population derived from the (ICMS 8511-S1-17-2-1-1-B-P03 × AIMP 92901-S1-183-2-2-B-08) cross (**Table 1**).

CASE STUDIES FOR GWAS IN PEARL MILLET

The advent of the recently decoded pearl millet genome has opened prodigious possibilities to discern several QTLs and the functions of its associated candidate genes governing diverse traits (Varshney et al., 2017). The genome size of pearl millet ~1.79 Gb, representing 38,579 genes, 88,256 SSRs, and 4,50,000 SNPs, will certainly be a valuable resource for constructing precision genetic maps (Varshney et al., 2017). Genetic mapping can be constructed in two different ways; one way is through QTL-mapping/interval mapping (IM) and the other is by using the association mapping (AM)/LD-mapping approach. The major difference in these two mapping strategies is based on the presumed idea over recombination events causative for the phenotypic variations (Myles et al., 2009). In general, QTL-mapping/IM can be done by developing various mapping populations *viz*., F_2, and recombinant inbred line (RIL), near-

isogenic line (NIL), back cross (BC), and doubled haploid (DH)-derived populations in which one can assume a clear cut degree of relatedness for the recombination events between the two contrasting parents for the trait of interest (Abdurakhmonov and Abdukarimov, 2008). Genetic mapping in this type of controlled population size results in the limited attainability of meiotic events and the products in the form of QTLs will be localized with lower resolution (10 to 20 cM intervals), and it is also an expensive approach to maintain a large number of populations (Jannink and Walsh, 2002; Flint-Garcia et al., 2003; Holland, 2007).

On the other hand, in AM there is no requirement for developing hybridization-based mapping populations; rather, it needs diverse germplasm accessions, including collections of different land-races, varieties, and a breeding material termed as a 'panel' where relatedness for the recombination events are not under control because of numerous meiotic recombinations across the diverse germplasm (Verdeprado et al., 2018). The principle of AM relies on the linkage disequilibrium (LD), a non-random association between two genes/markers/QTLs at different loci; however, a non-random association between these components in the same loci results in increased linkage disequilibrium levels (Flint-Garcia et al., 2003; Álvarez et al., 2014). Taking the advantage of multiple historic recombination events within the diverse accessions since their domestication, the AM approach can be best suited for the identification of genes or QTLs with high resolution (100–1000 Kb), and these are tightly linked to a broad range of phenotypic traits (Mackay et al., 2009). The potential of identifying promising QTLs, and also in detecting causal polymorphisms at the gene level, has made association mapping a powerful approach to develop marker-trait associations (MTAs) with great precision (Meuwissen and Goddard, 2000; Palaisa et al., 2003).

However, due to the high level of heterogeneity and heterozygosity in most of the germplasm accessions of pearl millet, very few association mapping strategies were delivered (Kannan et al., 2014); herein they are discussed and these detailed approaches may expand the scope of AM studies of pearl millet in future. A generalized workflow for the pearl millet genome-wide association studies (GWAS) pipeline is presented in **Figures 1A, B**. Pearl millet crop adaptation to various agro-climatic conditions is an important subject of study to explore the underlying genetics associated with this important nutri-cereal. Association studies made by Saïdou et al. (2009) on this aspect reveals the genetic factors responsible for the variations in flowering time at the phytochrome C (PHYC) (866 bp) locus, which is one of the key trait involved in crop adaptation. A total of 90 inbred and 598 pearl millet varieties from India, East, and West Africa were used for generating phenotypic data; followed by genotyping with 27 SSR and 6 AFLP markers. An LMM (linear mixed model) was used to identify a significant association between the phenotypic trait and genetic variations. With an aim to identify the best candidate gene loci associated with the flowering time, Saïdou et al. (2014) further explored an extra 100bp region surrounding the PHYC gene and performed an association study, MCMC method (Markov chain Monte Carlo method), to identify the tightly linked markers (75 SNPs and INDELS) surrounding the PHYC (6 Kb) genomic region and also to show the extent of LD to confer *PHYC* gene as the best candidate gene. By integrating the genome scan approach with association mapping, Mariac et al. (2011) identified the *PgMADS11* gene, a MADS-box gene family member which plays a key role during somatic and reproductive phase development respective of different climatic conditions. Phenotyping data for the targeted traits from the 90 inbred lines *viz*., flowering time (FT), stem diameter (SD), plant height (PH), spikelet length (SpL), and spikelet density (SpD) are used for the association analysis; and the significant identified association of *PgMADS11* alleles with a varied flowering time that deciphers the role of *PgMADS11* in the plant adaptation process towards climatic change. Association studies of the selective SSR markers with the flowering time, plant height, panicle length, stover and grain yield were deciphered by Kannan et al. (2014).

A set of 250 full-sib progenies and 34 SSR markers were used for GWAS analysis, and results revealed the strong association of the *Xpsmp*2248_162 marker allele at linkage group 6 (LG6) with earlier flowering time and reduced plant height. Marker allele, *Xpsmp*2224_157 on LG7 was strongly associated with the plant height. For panicle length, *Xpsmp*2077_136, *Xpsmp*2233_260, and *Xpsmp*2224_157 were strongly associated with LG2, LG5, and LG7, respectively, whereas the *Xpsmp*2237_230 marker allele showed strong positive association on LG7 with grain yield. For stover dry matter yield, the *Xicmp*3058 193 marker allele showed strong positive correlation on LG6. There is a pressing need for information on genes associated with low phosphorus tolerance, especially in the regions of West Africa. Gemenet et al. (2015) made the first-ever reported association analysis of the available 285 DArT markers with the phenotypic data generated from 151 PMiGAP lines from West Africa across six environs under high and low P conditions. Results showed that the *PgPb11603* DArT marker showed stable association with the flowering time, and the *PgPb12954* marker showed a significant association with the grain yield.

Association studies reveal that *Xibmsp11/AP6.1*, an SNP marker on an acetyl CoA carboxylase gene, is strongly associated with GY, GHI (grain harvest index), and PY (panicle yield) under both treatments; whereas InDel markers *viz*., *Xibmcp09/AP10.1* & *Xibmcp09/AP10.2* of a chlorophyll a/b binding protein gene are associated with GY and stay-green traits. Using association mapping, key alleles for grain iron and zinc were demonstrated by Anuradha et al. (2017). Developing MTAs (Marker Trait Associations) between 250 SSR and 17 genic markers with grain iron and zinc content for 130 diversified lines across different environs revealed that the Xicmp3092 marker had a strong association with grain iron content on LG 7, and markers Xpsmp2086 & Xpsmp2213 and Xipes0224 showed association with grain zinc content on LG 4 and LG 6, respectively; conserved association for grain iron and zinc, however, was exhibited by Xipes0180, Xpsmp2261, and Xipes0096 on LG 3, LG 5, and LG 7, respectively. Another association study by Varshney et al. (2017) delivered key findings

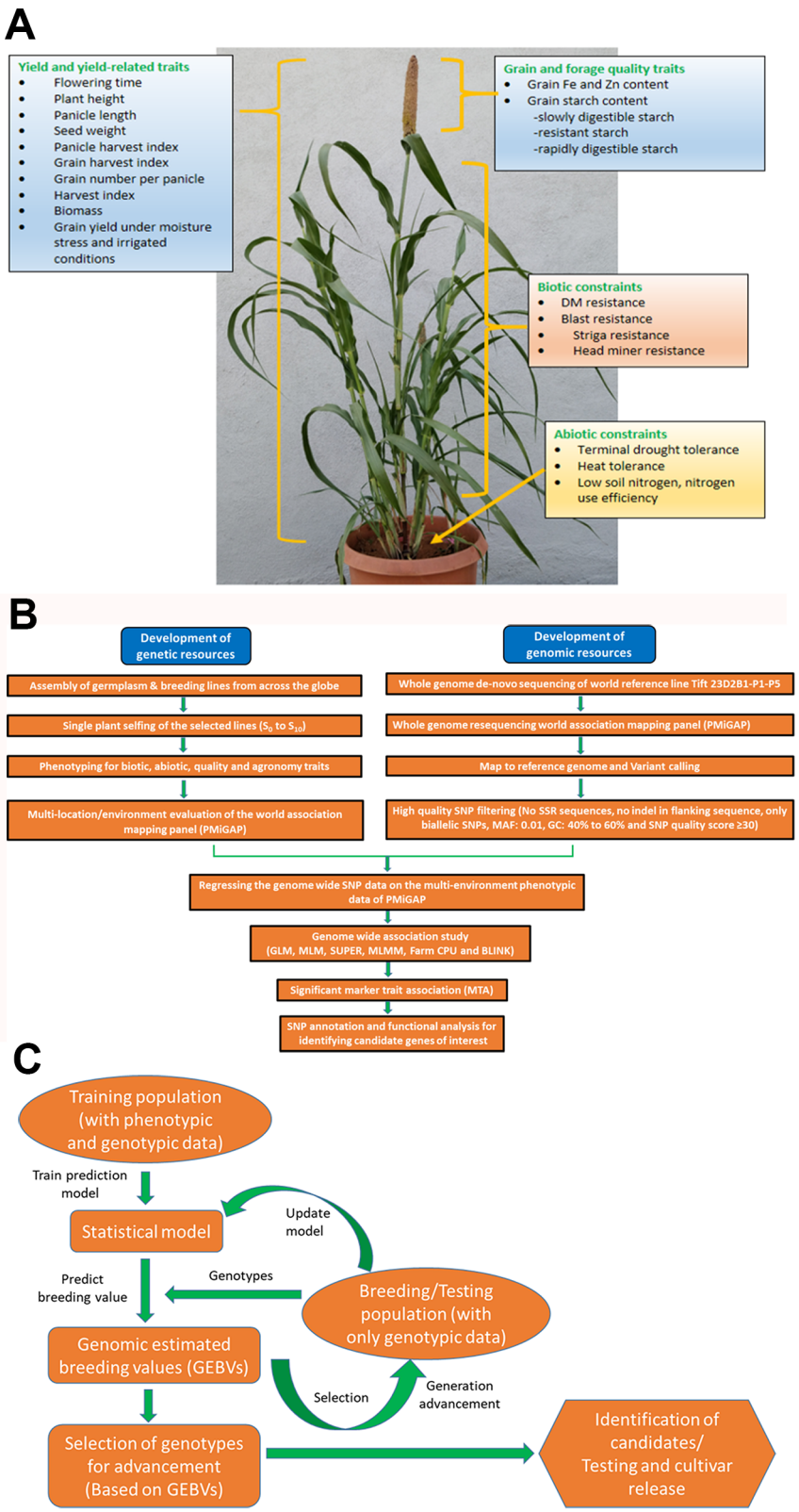


FIGURE 1 | (A) Depiction of traits for which genome-wide association studies (GWAS) and genomic selection (GS) is being attempted at ICRISAT, Patancheru. **(B)** Workflow for genome-wide association studies (GWAS) pipeline. **(C)** Workflow for genomic selection pipeline.

while establishing MTAs. Using whole-genome SNP data, a total of 3,117,056 SNPs were selected for GWAS analysis, and the phenotypic data for 20 agro-morphological traits was generated from 288 TCH (testcross hybrids) under two-stage (early & late) drought stress conditions along with respective controls. A significant association of the markers with the desired trait GNP (grain number per panicle) was exhibited on pseudomolecules Pg1 and Pg5. Genetic and genomic sequence information is now readily available for pearl millet. As AM will purvey a high-resolution power with the species exhibiting genotypic diversity across the germplasm (Álvarez et al., 2014), expanding AM studies in pearl millet will be increasingly fruitful for further crop improvement programs.

CASE STUDIES FOR GENOMIC SELECTION (GS) IN PEARL MILLET

Genomic (or genome-wide) selection (GS) is a promising strategy that has huge potential to explore and increase the genetic gain per selection in a breeding scheme per unit timeline and, thus, speed and efficacy in breeding programs (Spindel et al., 2015). GS has proven to be an economical and viable alternative to marker-assisted selection (MAS) and phenotypic selection (PS) for quantitative traits and accelerated crop improvement programs in cereals and several other crops (Heffner et al., 2009; Zhong et al., 2009; Crossa et al., 2010; Ornella et al., 2012; Poland et al., 2012; Spindel et al., 2015; Muleta et al., 2019). By developing efficient training population (having both genotypic and phenotypic data) designs, it predicts the genomic estimated breeding values (GEBV) of the testing population (having only genotypic data) by utilizing genome-wide high throughput DNA markers that are in linkage disequilibrium (LD) with QTL, and predicted GEBVs are used for selection (Meuwissen et al., 2001). One of the key advantages of GS is that decisions on selections can be taken during the off-season, leading to improvements in genetic gain on an annual basis (Heffner et al., 2009). Advancement and application of GS in pearl millet breeding programs facilitate precise prediction of hybrid performance along with ideal resource allocation. In ICRISAT, efforts are being made to exploit the available whole-genome resequencing (WGRS) data of PMiGAP lines along with phenotyping data for different traits for GWAS and GS. Building on the various target traits using GWAS (**Figures 1A, B**), various whole-genome prediction/genomic selection models are being developed and optimized in pearl millet. A generalized workflow for the pearl millet genomic selection pipeline is presented in **Figure 1C**.

Varshney and his group (Varshney et al., 2017) applied WGRS data for genomic selection by ridge regression best linear unbiased prediction (RR-BLUP) to predict grain yield for test crosses in four scenarios viz., the performance of grain yield in control, early stress, late stress, and across environments and observed high prediction accuracies for the performance of across environments. It was also reported that by using GS strategy (additive and dominance effects) the hybrid performance was also predicted by analyzing grain yield data with 302,110 SNPs, and 170 promising hybrid combinations were found, of which 11 hybrid combinations were already utilized for hybrid production with better performance and the remaining 159 hybrid combinations could be potential candidates for developing high yielding hybrids. A hierarchical clustering analysis of possible single cross combinations (167910) revealed two sets of lines with a higher hybrid performance by 8% by crossing each other. These hybrids could be a potential nucleus for establishing high-yielding heterotic gene pools for developing pearl millet hybrids with higher yield potential (Varshney et al., 2017). In a recent study, Liang et al. (2018) assessed two potential genotyping strategies viz., RAD-seq and tGBS, to characterize a set of ICRISAT-developed inbred pearl millet lines and evaluated the utility of genomic selection/prediction. By utilizing the projected hybrids from both (RADseq and tGBS) techniques and four genomic prediction schemes in pearl millet and assessed for each phenotype, 20 random rounds of five-fold cross-validation were performed for a tested SNP set. It was reported that, by utilizing hybrid data, the genomic prediction scheme (RR-BLUP) generated median prediction ranges (in parentheses) for different traits viz., 1,000 grain weight (0.73–0.74); days to flowering (0.87–0.89); grain yield (0.48–0.51); and plant height (0.72–0.73), respectively. Other traits with less/no heterosis, only hybrid, and hybrid/inbred schemes were also performed equivalently. It was also reported that hybrid GEBVs can be moderately improved by incorporating inbred phenotypic data sets, once inbred, and hybrid trait values relative to the mean trait values of that population. It was also well demonstrated that guileless integration of historical inbred phenotypic data into hybrid breeding programs could reduce the prediction accuracy of traits exhibiting heterosis. However, controlling the heterosis effects within the inbred genotype and trait data could improve the accuracy of GEBVs for hybrids, which, in turn, strengthens pearl millet hybrid breeding programs.

CHALLENGES IN USING GWAS AND GS FOR PEARL MILLET

Being a poor man's crop, pearl millet has attracted relatively less attention from various governments and policymakers in terms of support for the development of upstream science. This is particularly noted in areas such as GWAS and GS. The funding issues for carrying out this basic work in genomics has always remained an issue in pearl millet.

On the crop side, the high outcrossing rates, heterozygous nature, presence of inbreeding depression, and residual heterozygosity pose bottlenecks in inbred line development programs for the development of association mapping panels and for parental line/cultivar development were used in the training sets for GS. The presence of rapid linkage disequilibrium decay (LDD) warrants a relatively high number of markers for carrying out GWAS and GS. High rates of segregation distortions in specific populations may also pose serious challenges in GWAS and for getting high prediction

accuracies for robust GS model development. Single-cross hybrids occupy a major market share in India, while top-cross and three-way hybrids are important for Africa. The development of GS models for hybrid parental lines resulting in heterotic combinations is quite challenging. These warrant precise estimation of the general combining ability (GCA) and specific combining ability (SCA) for specific agro-ecologies and their precise genotype-by-environment (G × E) interactions.

CONCLUSIONS AND WAY FORWARD

Pearl millet is a nutritious, climate change ready crop capable of yielding economic return in marginal conditions where other cereals may fail. In recent years, pearl millet has seen an enormous increase in terms of various genetic and genomic tools at the disposal of pearl millet workers worldwide. Whole-genome sequencing of the pearl millet genome and resequencing efforts resulting in the generation of millions of genome-wide SNPs have facilitated efforts to map various yield and yield-related, key biotic and abiotic stress tolerance, and nutritionally important traits globally. These genomic resources have also facilitated taking up of the whole-genome prediction model development and validation efforts. There is a need to further validate the loci linked to various traits of interest and move from "loci" to "genes." There is an enormous opportunity to apply these learnings in the development of robust whole-genome prediction models with special emphasis on combining ability and heterotic gene pool studies for the development of heterotic hybrids.

AUTHOR CONTRIBUTIONS

RKS and RG planned and coordinated this study. RKS, RG, RY, TC, RBS, SB, MP, and VP contributed to this work and drafted the manuscript. RKS and RG edited the manuscript for publication.

ACKNOWLEDGMENTS

Funding support from the Department of Biotechnology (DBT), Government of India, the Biotechnology and Biological Sciences Research Council (BBSRC), United Kingdom, and the CGIAR Research Program on Grain Legumes and Dryland Cereals (CRP-GLDC) is gratefully acknowledged. This work has been published as part of the CRP-GLDC.

REFERENCES

Abdurakhmonov, I. Y., and Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *Int. J. Plant Genomics* 2008, 1–18. doi: 10.1155/2008/574927

Allouis, S., Qi, X., Lindup, S., Gale, M. D., and Devos, K. M. (2001). Construction of a BAC library of pearl millet. *Pennisetum Glaucum. Theor. Appl. Genet.* 102, 1200–1205. doi: 10.1007/s001220100559

Álvarez, M. F., Mosquera, T., and Blair, M. W. (2014). The use of association genetics approaches in plant breeding. *Plant Breed. Rev.* 38, 17–68. doi: 10.1002/9781118916865.ch02

Ambawat, S., Senthilvel, S., Hash, C. T., Nepolean, T., Rajaram, V., Eshwar, K., et al. (2016). QTL mapping of pearl millet rust resistance using an integrated DArT-and SSR-based linkage map. *Euphytica* 209, 461–476. doi: 10.1007/s10681-016-1671-9

Anuradha, N., Satyavathi, C. T., Bharadwaj, C., Nepolean, T., Sankar, S. M., Singh, S. P., et al. (2017). Deciphering genomic regions for high grain iron and zinc content using association mapping in pearl millet. *Front. Plant Sci.* 8, 412. doi: 10.3389/fpls.2017.00412

Arya, L., Verma, M., Gupta, V. K., and Karihaloo, J. L. (2009). Development of EST SSRs in finger millet (Eleusine coracana ssp coracana) and their transferability to pearl millet (Pennisetum glaucum). *J. Plant Biochem. Biotechnol.* 18, 97–100.

Bennett, M. D., Bhandol, P., and Leitch, I. J. (2000). Nuclear DNA amounts in angiosperms and their modern uses—807 new estimates. *Annals Bot.* 86, 859–909. doi.org/10.1006/anbo.2000.1253

Bertin, I., Zhu, J. H., and Gale, M. D. (2005). SSCP-SNP in pearl millet-a new marker system for comparative genetics. *Theor. Appl. Genet.* 110, 1467–1472. doi: 10.1007/s00122-005-1981-0

Bhattacharjee, R., Bramel, P., Hash, C., Kolesnikova-Allen, M., and Khairwal, I. (2002). Assessment of genetic diversity within and between pearl millet landraces. *Theor. Appl. Genet.* 105, 666–673. doi: 10.1007/s00122-002-0917-1

Breseghello, F., and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172, 1165–1177. doi: 10.1534/genetics.105.044586

Budak, H., Pedraza, F., Cregan, P. B., Baenziger, P. S., and Dweikat, I. (2003). Development and utilization of SSRs to estimate the degree of genetic relationships in a collection of pearl millet germplasm. *Crop Sci.* 43, 2284–2290.

Chelpuri, D., Sharma, R., Durga, K. K., Katiyar, P., Mahendrakar, M. D., Singh, R. B., et al. (2019). Mapping quantitative trait loci (QTLs) associated with resistance to major pathotype-isolates of pearl millet downy mildew pathogen. *Eur. J. Plant Pathol.*, 1–12. doi: 10.1007/s10658-019-01718-x

Crossa, J., Gustavo de, L. C., Paulino, P., Daniel, G., Juan, B., José, L. A., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Devos, K. M., Pittaway, T. S., Busso, C. S., Gale, M. D., Witcombe, J. R., and Hash, C. T. (1995). Molecular tools for the pearl millet nuclear genome. *Int. Sorghum Millets Newslett.* 36, 64–66.

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S.IV (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Boil.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Gemenet, D. C., Hash, C. T., Sanogo, M. D., Sy, O., Zangre, R. G., and Leiser, W. L. (2015). Phosphorus uptake and utilization efficiency in West African pearl millet inbred lines. *Field Crops Res.* 171, 54–66. doi: 10.1016/j.fcr.2014.11.001

Gómez, G., Álvarez, M. F., and Mosquera, T. (2011). Association mapping, a method to detect quantitative trait loci: statistical bases. *Agron. Colomb.* 29, 367–376.

Gupta, P. K., Rustgi, S., and Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* 57, 461–485. doi: 10.1007/s11103-005-0257-z

Gupta, S. K., Rai, K. N., Singh, P., Ameta, V. L., Gupta, S. K., Jayalekha, A. K., et al. (2015). Seed set variability under high temperatures during flowering period in pearl millet (*Pennisetum glaucum* L.(R.) Br.). *Field Crops Res.* 171, 41–53. doi: 10.1016/j.fcr.2014.11.005

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49 (1), 1–12. doi: 10.2135/cropsci2008.08.0512

Holland, J. B. (2007). Genetic architecture of complex traits in plants. *Curr. Opin.in Plant Boil.* 10, 156–161. doi: 10.1016/j.pbi.2007.01.003

Hu, Z., Mbacké, B., Perumal, R., Guèye, M. C., Sy, O., Bouchet, S., et al. (2015). Population genomics of pearl millet (Pennisetum glaucum (L.) R. Br.): comparative analysis of global accessions and Senegalese landraces. *BMC Genomics* 16, 1048. doi: 10.1186/s12864-015-2255-0

Jogaiah, S., Sharathchandra, R. G., Raj, N., Vedamurthy, A. B., and Shetty, H. S. (2014). Development of SCAR marker associated with downy mildew disease resistance in pearl millet (Pennisetum glaucum L.). *Mol. Biol. Rep.* 41, 7815–7824. doi: 10.1007/s11033-014-3675-7

Jannink, J. L., and Walsh, B. (2002). Association mapping in plant populations. *Quant. Genet. Genomics Plant Breed.*, 59–68. doi: 10.1079/9780851996011.0059

Jones, E. S., Liu, C. J., Gale, M. D., Hash, C. T., and Witcombe, J. R. (1995). Mapping quantitative trait loci for downy mildew resistance in pearl millet. *Theor. Appl. Genet.* 94, 448–456.

Kanatti, A., Rai, K. N., Radhika, K., Govindaraj, M., Sahrawat, K. L., Srinivasu, K., et al. (2014). Relationship of grain iron and zinc content with grain yield in pearl millet hybrids. *Crop Improv.* 41, 91–96.

Kannan, B., Senapathy, S., Raj, B., Gajaraj, A., Chandra, S., and Muthiah, A. (2014). Association analysis of SSR markers with phenology, grain, and stover-yield related traits in Pearl Millet (*Pennisetum glaucum* (L.) R. Br.). *Sci. World J.* 2014, 1–14. doi: 10.1155/2014/562327

Kraakman, A. T., Niks, R. E., Van den Berg, P. M., Stam, P., and Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168, 435–446. doi: 10.1534/genetics.104.026831

Kumar, S., Hash, C. T., Thirunavukkarasu, N., Singh, G., Rajaram, V., Rathore, A., et al. (2016). Mapping quantitave trait loci controlling high iron and zinc in self and open pollinated grains of pearl millet [*Pennisetum glaucum* (L) R. Br.]. *Front. Plant Sci.* 7, 1636. doi: 10.3389/fpls.2016.01636

Kumar, S., Hash, C., Nepolean, T., Mahendrakar, M., Satyavathi, C., Singh, G., et al. (2018). Mapping grain iron and zinc content quantitative trait loci in an iniadi-derived immortal population of pearl millet. *Genes* 9, 248.

Liang, Z., Gupta, S. K., Yeh, C. T., Zhang, Y., Ngu, D. W., Kumar, R., et al. (2018). Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. *G3: Genes Genomes Genet.* 8, 7, 2513–2522. doi: 10.3390/genes9050248

Liu, C. J., Witcombe, J. R., Pittaway, T. S., Nash, M., Hash, C. T., Busso, C. S., et al. (1994). An RFLP-based genetic map of pearl millet (*Pennisetum glaucum*). *Theor. Appl. Genet.* 89, 481–487. doi: 10.1534/g3.118.200242

Liu, C. J., Witcombe, J. R., Pittaway, T. S., Nash, M., Hash, C. T., and Gale, M. D. (1992). Restriction fragment length polymorphism in pearl millet, Pennisetum glaucum,". *Actes decolloque International*. Eds. J. C Mounolou, and A. Sarr (Paris: Bureau des Ressources Génétiques), 233–241.

Liu, N., Xue, Y., Guo, Z., Li, W., and Tang, J. (2016). Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. *Front. Plant Sci.* 7, 1046. doi: 10.1007/BF00225384

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10, 565–577. doi: 10.2135/cropsci2011.09.0503

Manning, K., Pelling, R., Higham, T., Schwenniger, J. L., and Fuller., D. Q. (2011). 4500-year old domesticated pearl millet (*Pennisetum glaucum*) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. *J. Archaeol. Sci.* 38, 312–322. doi: 10.1038/nrg2612

Mariac, C., Jehin, L., Saïdou, A. A., Thuillet, A. C., Couderc, M., and Sire, P. (2011). Genetic basis of pearl millet adaptation along an environmental gradient investigated by a combination of genome scan and association mapping. *Mol. Ecol.* 20, 80–91. doi: 10.1016/j.jas.2010.09.007

Mariac, C., Luong, V., Kapran, I., Mamadou, A., Sagnard, F., Deu, M., et al (2006). Diversity of wild and cultivated pearl millet accessions (Pennisetum glaucum [L.] R. Br.) in Niger assessed by microsatellite markers. *Theoret. Appl. Genet.* 11, 49–58. doi: 10.1007/s00122-006-0409-9

Meuwissen, T. H. E., and Goddard, M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155, 421–430. doi: 10.1111/j.1365-294X.2010.04893.x

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

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

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 212194-, 2202. doi: 10.1105/tpc.109.068437

Nambiar, V. S., Dhaduk, J. J., Sareen, N., Shahu, T., and Desai, R. (2011). Potential functional implications of pearl millet (*Pennisetum glaucum*) in health and disease. *J. Appl. Pharm. Sci.* 01, 62–67. doi: 10.1105/tpc.109.068437

Ornella, L., Singh, S., Perez, P., Burgueño, J., Singh, R., Tapia, E., et al. (2012). Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome* 5, 136–148. doi: 10.3835/plantgenome2012.07.0017

Palaisa, K. A., Morgante, M., Williams, M., and Rafalski, A. (2003). Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plan. Cell.* 15, 1795–1806. doi: 10.1105/tpc.012526

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.1105/tpc.012526

Qi, X., Pittaway, T. S., Lindup, S., Liu, H., Waterman, E., Padi, F. K., et al. (2004). An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, Pennisetum glaucum. *Theor. Appl. Genet.* 109, 1485–1493. doi: 10.3835/plantgenome2012.06.0006

Rai, K. N., Govindaraj, M., and Rao, A. S. (2012). Genetic enhancement of grain iron and zinc content in pearl millet. *Crops Foods*. 4, 119–125. doi: 10.1111/j.1757-837X.2012.00135.x.

Rajaram, V., Nepolean, T., Senthilvel, S., Varshney, R. K., Vadez, V., Srivastava, R. K., et al. (2013). Pearl millet [*Pennisetum glaucum* (L.) R. Br.] consensus linkage map constructed using four RIL mapping populations and newly developed EST-SSRs. *BMC Genomics* 14, 159. doi: 10.1007/s00122-004-1765-y

Rajaram, V., Varshney, R. K., Vadez, V., Nepolean, T., Senthilvel, S., and Kholova, J. (2010). Development of EST resources in pearl millet and their use in development and mapping of EST-SSRs in four RIL populations. *Plant Anim. Gen.* 18, 9–13.

Rao, P. P., Birthal, P. S., Reddy, B. V., Rai, K. N., and Ramesh, S. (2006). Diagnostics of sorghum and pearl millet grains-based nutrition in India. *International Sorghum and Millets News letter*. 47, 93–96.

Saïdou, A. A., Mariac, C., Luong, V., Pham, J. L., Bezançon, G., and Vigouroux, Y. (2009). Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics* 182, 899–910. doi: 10.1186/1471-2164-14-159

Saïdou, A. A., Clotault, J., Couderc, M., Mariac, C., Devos, K. M., and Thuillet, A. C. (2014). Association mapping, patterns of linkage disequilibrium and selection in the vicinity of the PHYTOCHROME C gene in pearl millet. *Theor. Appl. Genet.* 127, 19–32. doi: 10.1534/genetics.109.102756

Sehgal, D., Rajaram, V., Armstead, I. P., Vadez, V., Yadav, Y. P., Hash, C. T., et al. (2012). Integration of gene-based markers in a pearl millet genetic map for identification of candidate genes underlying drought tolerance quantitative trait loci. *BMC Plant Biol.* 12 (1), 9. doi: 10.1007/s00122-013-2197-3

Sehgal, D., Skot, L., Singh, R., Srivastava, R. K., Das, S. P., and Taunk, J. (2015). Exploring potential of pearl millet germplasm association panel for association mapping of drought tolerance traits. *PloS One* 10, 1–28. doi: 10.1186/1471-2229-12-9

Senthilvel, S., Jayashree, B., Mahalakshmi, V., Kumar, P. S., Nakka, S., Nepolean, T., et al. (2008). Development and mapping of simple sequence repeat markers for pearl millet from data mining of expressed sequence tags. *BMC Plant Biol.* 8, 119. doi: 10.1371/journal.pone.0122165

Senthilvel, S., Nepolean, T., Supriya, A., Rajaram, V., Kumar, S., Hash, C. T., et al. (2010).Development of a molecular linkage map of pearl millet integrating DArT and SSR markers, in: Proc. Plant Animal Genome 18 Conference, San Diego, CA. pp. 9–13. doi: 10.1186/1471-2229-8-119

Serba, D. D., and Yadav, R. S. (2016). Genomic tools in pearl millet breeding for drought tolerance: status and prospects. *Front. Plant Sci.* 7, 1724. doi: 10.3389/fpls.2016.01724

Singh, M., and Upadhyaya, H. D. (2016). Genetic and genomic resources for grain cereals improvement. Academic Press. pp: 253–289. doi: 10.1016/B978-0-12-802000-5.00006-X

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PloS Genet.* 11, p.e1004982. doi: 10.1371/journal.pgen.1005350

Supriya, A., Senthilvel, S., Nepolean, T., Eshwar, K., Rajaram, V., Shaw, R., et al. (2011). Development of a molecular linkage map of pearl millet integrating DArT and SSR markers. *Theor. Appl. Genet.* 123, 239–250. doi: 10.1371/journal.pgen.1004982

Upadhyaya, H. D., Reddy, K. N., and Gowda, C. L. L. (2007). Pearl millet germplasm at ICRISAT genebank-status and impact. *J. SAT Agricul. Res.* 3, 1–5.

Upadhyaya, H. D., Yadav, D., Reddy, K. N., Gowda, C. L. L., and Singh, S. (2011). Development of pearl millet minicore collection for enhanced utilization of germplasm. *Crop Sci.* 5, 217–223.

Vadez, V., Hash, T., Bidinger, F. R., and Kholova, J. (2012). Phenotyping pearl millet for adaptation to drought. *Front. Physiol.* 3, 1–386. doi: 10.3389/fphys.2012.00386

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* 35, 969. doi: 10.1007/s00122-011-1580-1

Verdeprado, H., Kretzschmar, T., Begum, H., Raghavan, C., Joyce, P., and Lakshmanan, P. (2018). Association mapping in rice: basic concepts and perspectives for molecular breeding. *Plan. Produc. Sci.* 21, 159–176. doi: 10.1038/nbt.3943

Yadav, O. P., Mitchell, S. E., Fulton, T. M., and Kresovich, S. (2008). Transferring molecular markers from sorghum, rice and other cereals to pearl millet and identifying polymorphic markers. An Open Access Journal published by ICRISAT. 6, 1–4.

Yadav, O. P., Mitchell, S. E., Zamora, A., Fulton, T. M., and Kresovich, S. (2007). Development of new simple sequence repeat markers for pearl millet. *J. SAT Agricul. Res.* 3, 1–34.

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PloS Genet.* 3, e4. doi: 10.1080/1343943X.2018.1483205

Zhong, S., Dekkers, J. C., Fernando, R. L., and Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182, 355–364. doi: 10.1371/journal.pgen.0030004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Srivastava, Singh, Pujarula, Bollam, Pusuluri, Chellapilla, Yadav and Gupta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 17 March 2020
doi: 10.3389/fevo.2020.00032

Predictive Characterization for Seed Morphometric Traits for Genebank Accessions Using Genomic Selection

Zakaria Kehel[1], Miguel Sanchez-Garcia[1], Adil El Baouchi[1], Hafid Aberkane[1], Athanasios Tsivelikas[1], Chen Charles[2] and Ahmed Amri[1]*

[1] Biodiversity and Crop Improvement Program, International Center for Agricultural Research in the Dry Areas, Rabat, Morocco, [2] Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, United States

Seed traits of bread wheat, including the seed size that is considered to be associated with early vigor of the crop and end-use quality, are valuable to farmers and breeders. In this study, a collection of 789 bread wheat landraces, held in-trust at the genebank of the International Center for Agricultural Research in the Dry Areas (ICARDA) were scanned for seed morphometric traits using GrainScan. Diversity analysis using the 12k DartSeq SNP markers revealed that these accessions can be grouped into five distinct clusters. To evaluate the performance for early selection from genebank accessions, we examined the accuracy of genomic selection models with genomic relationship that these landraces accounted for. Based on cross-validations, prediction accuracies for seed traits ranged from 0.64 for seed perimeter to 0.74 for seed width. The variability of prediction accuracies across random validations averaged at 0.14, with a range from 0.12 to 0.18, suggesting stable predictability and reproducible results even with a collection of much greater genetic diversity from genebank accessions. Adding the climatic relationship matrix between accessions based on passport information improved the predictive ability by 8%. Our results on seed traits demonstrated the capacity for estimating important agronomic phenotypes for genebank accessions directly based on genomic information, further advocating the advance in genomic technologies for identifying parental germplasm as potential donors of beneficial alleles for introgression.

Keywords: wheat, genomic selection, seed characteristics, landraces, genebank

OPEN ACCESS

Edited by:
Genlou Sun,
Saint Mary's University, Canada

Reviewed by:
Raj K. Pasam,
AgriBio, La Trobe University, Australia
Vetriventhan Mani,
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India

****Correspondence:***
Zakaria Kehel
z.kehel@cgiar.org

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *11 October 2019*
Accepted: *05 February 2020*
Published: *17 March 2020*

Citation:
Kehel Z, Sanchez-Garcia M, El Baouchi A, Aberkane H, Tsivelikas A, Charles C and Amri A (2020) Predictive Characterization for Seed Morphometric Traits for Genebank Accessions Using Genomic Selection. Front. Ecol. Evol. 8:32. doi: 10.3389/fevo.2020.00032

INTRODUCTION

Wheat is one of the most important cultivated food crops, and its cultivation goes back some 11,000–10,000 years ago (Nesbitt, 2002; Zohary et al., 2012). Wheat has been the fundamental staple food for the majority of human civilizations in Europe, West Asia, and North Africa (Curtis et al., 2002) because of its crucial nutritional value and its significant contribution to daily energy intake. Wheat is very diverse and widely adaptable (Levandi et al., 2014), and its gene pool is rich in genes that can be used to improve resistance/tolerance to biotic and abiotic stresses and micronutrient availability. However, to secure an efficient continuum between the conservation and the use of genetic resources, wheat accessions need to be well-characterized and evaluated for a range of traits. The study of this phenotypic diversity will result in better use in breeding programs.

The major obstacle to enhance the use of genebank material is the lack of adequate characterization and evaluation data, and thus, the inability to adequately respond to inquiries for these accessions that directly meet the needs of the users. Several methods of linking traits to a genebank accession have been reviewed (Anglin et al., 2018) including phenotyping of large or random samples, core and mini core collections, the focused identification of the germplasm strategy (FIGS) and Generation Challenge Program subsets and use of molecular techniques and genome wide association studies (GWAS). FIGS is a useful approach developed at the International Center for Agricultural Research in the Dry Area (ICARDA) to identify subset of accessions with a high probability of containing specific target traits based on the ecogeographical information of the sites where the populations were collected (Mackay and Street, 2004). Success in FIGS has been seen in the identification for sources of resistance to Sunn pest in wheat in Syria (El Bouhssini et al., 2009), and for Russian wheat aphid in bread wheat (El Bouhssini et al., 2011) and further in the identification of the traits related to abiotic stresses, such as drought adaptation in *Vicia faba* L. (Khazaei et al., 2013). FIGS, however, has not been used to study quantitative traits such as phenology and morphology.

Grain weight is one of the main wheat yield components, and grain size and shape have a direct impact on wheat market value. While flour is extracted from the endosperm, the inner part of the grain, and therefore spherical grains tend to produce more flour per kilogram of grain milled due to the lower surface/volume ratio (Evers et al., 1990). Also, the grain size was found to be associated with various characteristics of flour, such as protein quality and hydrolytic enzyme activity, which in turn determine baking quality and end-use suitability (Evers, 2000). Grain size in wheat is associated with seedling emergence and development, primarily through the influence on the rates of expansion of the first two leaves (Aparicio et al., 2002). Furthermore, grain weight has been associated with grain yield in a number of diverse environments of contemporaneous varietal panels (Lopes et al., 2012). In addition, research has found higher grain weight plays an important role in the robust establishment of bread wheat seedlings subjected to salinity stress (Grieve and Francois, 1992).

Previous studies have found larger grain size and shape variation in bread wheat landraces and old hexaploid species as compared to the tetraploid *Triticum* species (Gegas et al., 2010). This large variation has, however, decreased in modern germplasm, suggesting a breeding-related bottleneck on grain shape variability (Gegas et al., 2010). This bottleneck can be one of the reasons of the low (Austin et al., 1989; Brancourt-Hulmel et al., 2003; Sanchez-Garcia et al., 2013) or even negative (Siddique et al., 1989; Royo et al., 2007) contribution of grain size to wheat genetic gains in several countries. There is an urgent need to overcome this bottleneck by bringing novel diversity from genebanks to breeding programs.

Recent and rapid advancements in high throughput genotyping have greatly aided plant science through characterizing genetic diversity, genome-wide association studies, and genomic selection (GS). GS, as predictive analytics, uses genome-wide markers to predict genomic breeding values. GS has been widely applied to elite wheat germplasm (de los Campos et al., 2009, 2010; Crossa et al., 2010; González-Camacho et al., 2012; Heslot et al., 2012; Pérez-Rodríguez et al., 2012; López-Cruz et al., 2015; Hu et al., 2019). However, very few studies that evaluated the performance of GS with the inclusion of new diversity from genebanks can be found in the literature including *Thinopyrum intermedium* (Zhang et al., 2016), wheat landraces for rust resistance (Daetwyler et al., 2014; Pasam et al., 2017), mineral contents (Manickavelu et al., 2017), and heat and drought stress adaptation (Crossa et al., 2016).

Following the above, the objectives of this study were: (1) to examine the genomic prediction accuracy within a set of ICARDA bread wheat genebank collection for seed morphometric traits, (2) to study the effect of including a non-additive similarity matrix based on passport data, and (3) to study the effect of accounting for population structure in genomic prediction models.

MATERIALS AND METHODS

Landraces, Grain Color, and Morphology

Seven hundred eighty-nine (789) bread wheat landraces were randomly selected from the 4000 landraces grown at the ICARDA Marchouch station (33°36′ N 6°42′ W, 390 m a.s.l.) located in central Morocco during the cropping season 2016–2017 for the purpose of regeneration and characterization of accessions of genebank. Landraces are planted in a non-designed trial with two rows plot of 2 m long each. Best practices for the regeneration of wheat genetic resources were applied including supplemental irrigations and applying pesticides to control major diseases and pests to allow for good growing conditions and full expression of seed traits. Most accessions originated from North Africa, Middle East, and southeast Asia with a majority from Pakistan, Turkey and Morocco (Complete list can be found in **Supplementary Data Sheet 1**).

Random samples of 250–400 grains were obtained from the harvest of every plot and were scanned using a flatbed scanner (CanoScan LiDE 220; Canon). The images collected were analyzed using Grainscan software (Whan et al., 2014) and the morphological characteristics of every grain in every image obtained. Grain characteristics include the grain area (mm^2), perimeter (mm), grain length (mm), and width (mm). Additionally, Grainscan software produce for every grain analyzed an output of color channel intensity analogous to the standardized CIELAB colourspace (Whan et al., 2014). The color channels from GrainScan (ColCha1, ColCha2, and ColCha3) are considered therefore to be proxies for L*, which represents the lightness of the color; a*, which represent green or magenta; and b*, representing blue or yellow, respectively.

Genotypic Characterization and Diversity of Wheat Landraces

A high-throughput genotyping method using DArTseqTM technology was employed to generate genomic profile of the germplasm at the Genetic Analysis Service for Agriculture (SAGA) facility at the International Center for Maize and

Wheat Improvement (CIMMYT) in Mexico. DArTseq raw data were filtered according to markers criterion; minor allele frequency > 5% and missing data ≤ 20%. This resulted in a total of 12,472 Dartseq markers that were used in this study. Diversity analysis was performed using a discriminate analysis of principal component (DAPC) as described by Jombart et al. (2010) and principal component analysis (PCA) using R Core Team (2016).

Environmental Similarity Between Wheat Landraces

To characterize the environmental diversity and make an environmental similarity matrix based on passport data between bread wheat landraces, we collected datasets for a total of 36 potential drivers of crop diversity, including 35 climate variables and altitude (**Supplementary Table 1**). The climatic variables include the 19 bioclimatic variables from the WorldClim version 2 database (Fick and Hijmans, 2017), freely available at http://www.worldclim.org, and downloadable at 2.5 arc-min spatial resolution. Additional 16 climate variables were downloaded at the same spatial resolution from the Environmental Rasters for Ecological Modeling (ENVIREM) database (Title and Bemmels, 2018). These 35 variables (19 from WorldClim and 16 from ENVIREM) allow for a robust characterization of the climate signature of landraces and wild relatives (Braunisch et al., 2013; Title and Bemmels, 2018). The variables were scaled, and an Euclidian distance was computed, resulting into an environmental similarity matrix between landraces based on passport information.

Statistical Analysis

Genomic best linear unbiased prediction (G-BLUP) was used to perform genome wide predictions. We used a genomic relationship matrix between landraces using marker information defining covariance based on observed similarity at the genomic level as described by VanRaden (2007). This model captures a large additive genetic variance by accounting for genomic information and increases the heritability and prediction accuracy. Genomic heritability (h^2) was computed as the ratio between the genetic variance due to markers over the sum of the genetic variance plus the error variance. We have used for all models, as a more appropriate way, the complete dataset to estimate variance components (additive and residuals) and hence the genomic heritability.

Population structure might affect the estimation of heritability and the prediction accuracy in a genome wide prediction framework (Gou et al., 2014). To evaluate the impact of population structure in the performance of genomic prediction, we evaluated the following models:

(1) null model where no population structure was accounted for;
(2) accounting for population structure using discrete population resulting from DAPC with K number of subpopulations equal to 2 which is the first level of genetic separation (grp2);
(3) accounting for population structure using discrete population resulting from DAPC with K optimal number of subpopulations;
(4) accounting for population structure using 5 eigen vectors PC1 to PC5 resulting from PCA analysis. We have removed the population structure effect due to stratified populations using the population proxies (two discrete groups resulting from DAPC and PC1–PC5) as fixed effects in our models (Daetwyler et al., 2015). In addition, and to reduce the effect of population structure on the genomic prediction accuracy, we have also run predictions for separate subpopulations using groups resulted from DAPC analysis for K number of populations equal to 2.

Genomic predictions only consider the additive effects using the observed relationship between individuals using markers. It has been suggested that the estimation of non-additive effect can improve prediction accuracy (Varona et al., 2018). In this study, resemblance between landraces using environmental data from the site of the landrace's origin was used as a non-additive term, alone or in combination with the additive matrix, in the G-BLUP mixed model to account for the non-genetic effect.

To evaluate prediction model performance, cross-validation (CV30) where 30% of landraces were included in the validation set while the remaining 70% of landraces formed the training set, was employed. The process was repeated randomly 50 times. The prediction accuracy of a model was assessed using the Pearson correlation between genomic predictions and BLUP from the model using the full dataset.

All the above analyses were performed using a single stage analysis, where raw data from a single seed was used directly in the prediction models. Outliers were identified as data points with studentized residuals superior to 3.5 and removed from the final analysis. Models were fitted in ASReml v3.0-1 (Butler et al., 2009) for R v3.3.1 (R Core Team, 2016).

RESULTS

In this study, seven (7) seed traits were captured to determine grain shape, size and color for wheat landraces. The genomic heritability (**Table 1**) of the traits under this study ranged from moderate 0.47 for grain area and perimeter to high 0.78 for one of the color channels (ColCha1). As expected, large variation was found within the landraces used in this study (**Table 1**); the grain width showed less variability (range of 1 mm) than grain length with a range of 2 mm; and, grain area and perimeter ranged from 12 to 19.2 mm^2 and 17 to 22.3 mm, respectively.

For diversity analysis, PCA showed that the first five (5) eigen values explained 80% of the genetic variance. The set of landraces used in this study exhibited population structure as shown by plotting first versus second principal component (**Figure 1**). This structure was mainly due to the country of origin as landraces from the same country were clustered together. DAPC proposed $K = 5$ as the optimal number of

TABLE 1 | Summary statistics and genomic heritability for grain characteristics for the entire set collection and per subpopulation when $K = 2$ (grp1 and grp2) of the bread wheat landraces.

		Area	Perimeter	Length	Width	ColCha1	ColCha2	ColCha3
	h (Heritability)	0.47	0.47	0.52	0.54	0.77	0.78	0.68
	Mean	15.0	19.7	6.5	2.9	158.6	130.6	102.6
	Maximum	19.2	22.3	7.5	3.5	180.4	151.4	118.5
	Minimum	12.0	16.9	5.5	2.5	133.5	110.7	87.6
	SD	1.4	0.9	0.3	0.2	9.2	9.1	6.1
grp1	Mean	15.3	19.7	6.5	3.0	158.0	130.0	102.4
	Maximum	19.2	22.3	7.5	3.5	178.2	150.1	118.5
	Minimum	12.0	16.9	5.5	2.6	133.5	110.7	89.5
	SD	1.5	1.1	0.4	0.2	7.3	7.3	5.4
grp2	Mean	14.8	19.7	6.6	2.9	159.0	130.9	102.6
	Maximum	18.3	21.9	7.4	3.3	180.4	151.4	118.0
	Minimum	12.0	17.3	5.7	2.5	138.0	111.8	87.6
	SD	1.2	0.8	0.3	0.2	10.0	9.9	6.5

subpopulations as it presented the lowest Bayesian Criterion index value. Increasing K to more than five did not identify any further clear genetic group. The first level of separation $K = 2$ has clearly distinguished between landraces from Pakistan and Turkey and landraces from the other countries. However, when K was set to 5, the landraces were correctly classified into their agro-ecologies (**Figure 2**). The first subpopulation (red) is composed of landraces collected from hot environments, mainly from Pakistan, Egypt, and Oman. The second subpopulation (green) comprised of landraces collected in winter areas from Turkey. The third subpopulation (light blue) mostly made of landraces from Mediterranean environments and the spring type, whereas the fourth subpopulation (dark blue) composed of landraces collected in favorable Mediterranean environments. Finally, the last and fifth subpopulation (black) is the smallest one with 37 landraces originated mainly from North Africa and most probably are genetically similar to the beard wheat landraces from Southern Europe. The assignment of the wheat landraces to subpopulations for $K = 2$, 3, 4, and 5 can be found in the **Supplementary Table**: list of accessions.csv.

The grain characteristics have shown the same range of variation between the two subpopulations resulting from DAPC for $K = 2$ (**Table 1** and **Supplementary Figure 1**). Nevertheless, using the optimal number of subpopulations $K = 5$ revealed that some subpopulations (1 and 4) do not present as much variability as the other subpopulations (**Supplementary Figure 2**).

The prediction models showed medium to high accuracies for all grain traits (**Figure 3** and **Table 2**). The Perimeter showed the lowest prediction accuracy with an average of 0.64, followed by Length and ColCha3, ColCha1, Area and ColCha2 with an average accuracy of 0.66, 0.69, 0.7, and 0.74 respectively. The prediction accuracy reach its greatest value for grain width with 0.74 in average. Overall, the variability in accuracy between the 50 random cross-validations had similar trend as average accuracies, where the highest variation was identified in perimeter, length and the ColCha1 and ColCha2, and the lowest variation shown by grain area and grain width (**Figure 3**).

Prediction models using the climatic similarity matrix showed low values for prediction accuracy compared to the prediction model based on markers. On average for all traits, a maximum of 0.1 prediction accuracy was reached for the grain width (**Table 2**). The maximum prediction accuracy for the 50 random replicates reached more than 0.2 for all grain traits. Adding the climatic similarity to the genetic similarity in the prediction model has shown a slight increase in the prediction accuracy for all traits with a maximum increase of 0.06 (8%) achieved for the grain perimeter and ColCha1 and ColCha3.

Figure 4 displays genomic prediction accuracies comparing the null model with the ones that incorporated different population structure covariates. When accounting for population structure in the genomic prediction models, the change in the prediction accuracies showed different patterns depending on the variables used to correct for stratified populations and/or the trait under evaluation. Generally, when accounting for population structure using grp2, the accuracies were similar to the null models without accounting for population structure for all grain traits. However, when accounting for population structure using grp = 5, compared to the null model, there was a significant reduction of prediction accuracy at $\sim$0.06 (8%, p-value $\sim$ 0) found for all the traits; the lowest decrease was observed for the grain width (0.04). The most significant reduction in prediction accuracy was found when accounting for population structure using PC1–PC5. This decrease ranged from 0.33 for ColCh2 to as low as 0.11 and 0.14 for grain perimeter and area, respectively. Making genomic predicitions for grain characteristics for each subpopulation when K number of populations was equal to 2 has given contrasting results (**Table 2**). For the first subpopulation, increased prediction accuracy was obtained for the grain area, perimeter, and length, whereas prediction accuracies were found lower for the second subpopulation. A completely opposite pattern was observed for the width and the three color channels where subpopulation two showed a decreasing prediction accuracy for the area, length, and

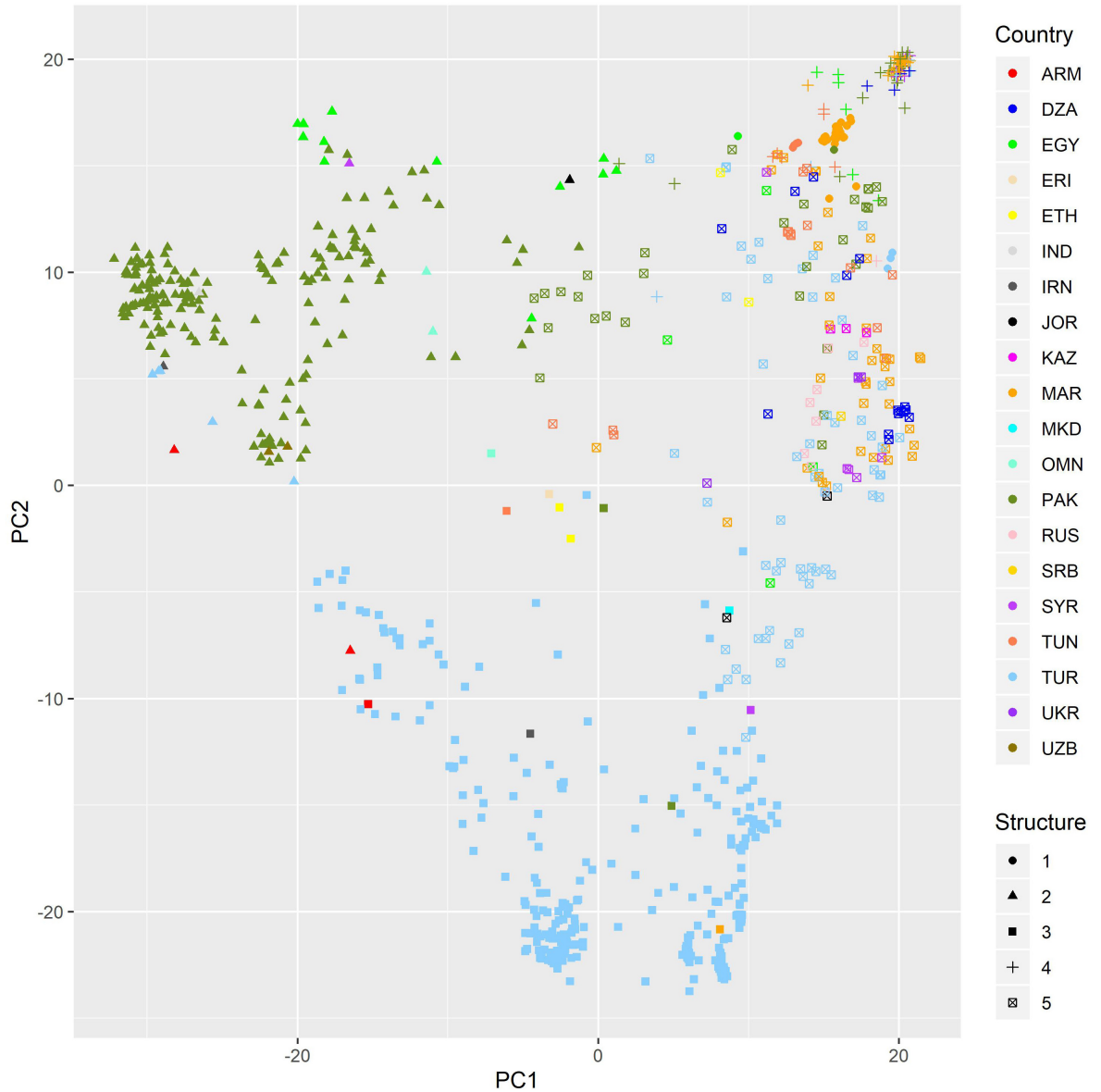


FIGURE 1 | PC plot of a set of bread wheat landraces from PCA using DarTseq markers. Colorsed by country of origin. Dot shape gives the assignment to subpopulations when $K = 5$.

perimeter and an increasing prediction accuracy for the three color channels.

DISCUSSION

Variation and Prediction Accuracies for Wheat Landrace's Grain Traits and Building on FIGS

Variability among bread wheat landraces was assessed for grain traits using image analysis. The seven grain traits exhibited medium to high heritability with considerable variation, at a similar order of heritability and scale of variabilities found in other studies (Gegas et al., 2010). Grain traits, especially grain shape and size, have a direct influence on yield and quality, and consequently, the market value of the wheat product. Ample evidence has also suggested that, compared to landraces and primitive wheat species, the significant reduction in grain shape and size of modern varieties is a result of domestication and breeding (Gegas et al., 2010). Landraces held in genebanks can have a crucial role in wheat breeding for grain traits because of their wide variability in terms of grain size and shape.

However, sending genebank's requesters the appropriate material to meet their demands is not a straightforward task. This means the genebank manager would have a complete description

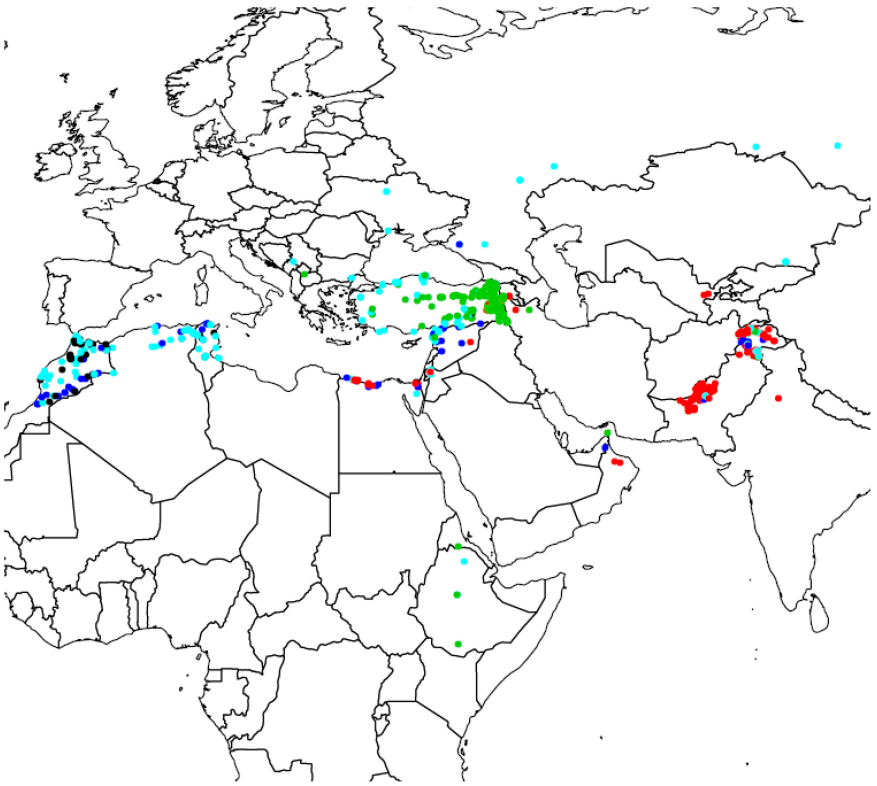


FIGURE 2 | Geographic distribution of the five subpopulations of bread wheat landraces found using Discriminant Analysis of Principal Component.

of all the genebank holdings. As a result, characterizing agronomically important grain traits has become an important activity within a genebank as it is essential to identify accessions with the desirable traits to be used as parental material in a breeding program. Characterization at the genebank level as well as at the breeding level is mainly based on 1000 kernel weight and hectoliter weight. However, characterization of large number of accessions held in the global collections is resources consuming. Several solutions have been developed and tested to address this issue, including core collections, FIGS, and GWAS (Anglin et al., 2018). The performance of core collection and GWAS to link a trait of interest to a genebank accession has been evaluated (Anglin et al., 2018); however, the application of FIGS has yet not been examined for quantitative morphological traits such as grain traits. In this research, we reported the efficiency of genome-wide prediction to predict the ICARDA genebank wheat landraces using high-density DartSeq^{TM} markers. This is done by characterizing a portion of the wheat collection for grain traits, including grain area, perimeter, length, and width and using DartSeq^{TM} and GS to predict the unevaluated genebank accessions. Our results suggest that genomic prediction is a useful tool for predictive characterization of genebank accessions, allowing phenotyping to be restricted to a portion of the collection in order to predict trait genomic estimated breeding value (GEBV) for the entire collection (Crossa et al., 2016; de Azevedo Peixoto et al., 2017; Thorwarth et al., 2017). We used GBLUP as a method of genomic predictions because of its performance stability and flexibility of applications regarding the genetic architecture (Meuwissen et al., 2001). Our study has shown that GS can be implemented within a genebank to predict important traits such as grain characteristics with accuracies of more than 0.7, more specifically for the trait with moderate to high heritability. Further work is needed to validate if those predictions are stable from year to year, knowing that our regeneration/characterization trials are done in the same experimental station and applying the same optimal field management practices.

We have shown that reasonable prediction accuracies for genomic predictions can be achieved using a randomly chosen subset from genebank wheat collection representing a wide genetic variability. These findings should encourage genebank managers to identify novel variation for potential use in breeding programs and facilitate broad, detailed phenotypic characterization of the entire genebank collection. Further, genotyping the entire *ex situ* collection is then needed to take full advantage of such technology.

Genomic Predictions in Stratified Populations

Genebank collections generally exhibit a wide array of genetic diversity, as well as the population structure due to the domestication process, including natural and farmer selection,

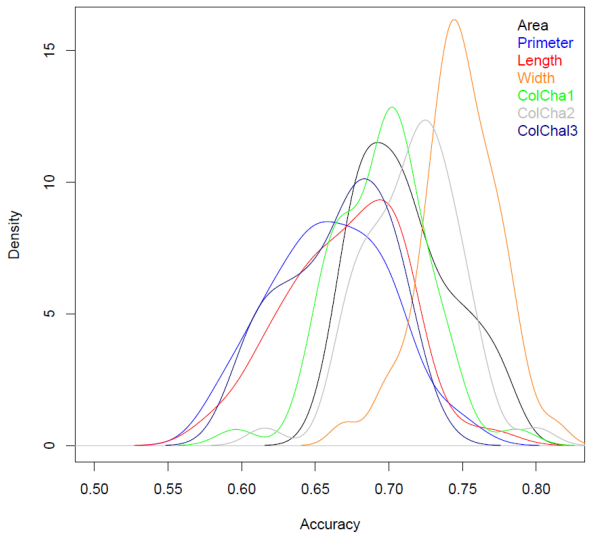


FIGURE 3 | Densities for prediction accuracy using validation set for grain traits of bread wheat genebank accessions.

genetic drift, and local adaptation. The knowledge of this diversity and structure is essential to genebanks when optimizing the collection's conservation policy to secure a continuum between the conservation and the use of the germplasm. As expected, a significant population structure in the collection of wheat landraces was identified in this study, with the first five principal components accounting for 80% of genetic variation. The strong population structure also showed a negative impact on performance in association studies and genomic prediction models, which was also found in other studies (Gou et al., 2014; Daetwyler et al., 2015). This degree of decrease was, however, dependent on way we accounted for the population structure in this study. For example, our study has noted that using a continuous axis from PCA analysis or discrete population assignment from structure or DAPC gave very contrasting results from almost no change in prediction accuracy to a significant reduction in prediction accuracy. Also, running genomic prediction for each of the subpopulation may or may not improve the accuracy depending on the subpopulation and the trait under study. Previous studies have shown that accounting for stratified populations is not an easy task in a genomic prediction models and this is generally done using the first five eigen values as covariates in a the GS model (Patterson et al., 2006; Daetwyler et al., 2014; Crossa et al., 2016; Norman et al., 2018). Further work and simulations should be undertaken to study the population structure effect carefully in the framework of genomic predictions.

Building on FIGS: FIGS +

FIGS has shown its relevance on delivering sources of resistance to diseases and sources of variation for important desirable trait to breeders worldwide for wheat and other ICARDA mandate crops (El Bouhssini et al., 2009, 2011; Khazaei et al., 2013). In recent years, FIGS is used by ICARDA to make predictive characterization for the genebank characterization traits for its collection (Azough et al., 2019). This is done by quantifying a relationship between collection site agro-climatic conditions and the presence of specific traits using machine learning algorithms. Moreover, the application FIGS has been successful for categorical traits such as growth stages, class of maturity, and tillering capacity. Other unpublished results have shown that the performance of these machine learning algorithms for quantitative traits was limited.

With rapid advances in genomics techniques, genetic resources users should be able to mine quickly genetic diversity as part of pre-breeding programs to achieve better and faster breeding outcomes and gains. More specifically, GS, which uses

TABLE 2 | Prediction accuracies for grain traits using only markers (All), passport information (Env) and combining both (All + Env), and for separate subpopulations grp1 and grp2 of bread wheat landraces.

		Area	Perimeter	Length	Width	ColCha1	ColCha2	ColCha3
All	Mean	0.70	0.64	0.66	0.74	0.69	0.71	0.66
	Maximum	0.76	0.73	0.74	0.80	0.78	0.80	0.73
	Minimum	0.64	0.55	0.57	0.67	0.60	0.62	0.59
All-Env	Mean	0.75	0.70	0.70	0.78	0.75	0.76	0.72
	Maximum	0.78	0.75	0.77	0.81	0.79	0.80	0.75
	Minimum	0.66	0.59	0.59	0.69	0.61	0.63	0.60
Env	Mean	0.06	0.04	0.03	0.10	0.03	0.05	0.04
	Maximum	0.20	0.19	0.22	0.29	0.20	0.25	0.24
	Minimum	−0.12	−0.19	−0.16	−0.07	−0.16	−0.13	−0.13
grp1	Mean	0.73	0.76	0.79	0.65	0.58	0.63	0.58
	Maximum	0.82	0.86	0.88	0.75	0.78	0.80	0.71
	Minimum	0.64	0.67	0.69	0.49	0.42	0.47	0.40
grp2	Mean	0.67	0.54	0.51	0.73	0.73	0.74	0.69
	Maximum	0.74	0.67	0.65	0.80	0.79	0.79	0.75
	Minimum	0.58	0.46	0.40	0.65	0.63	0.64	0.61

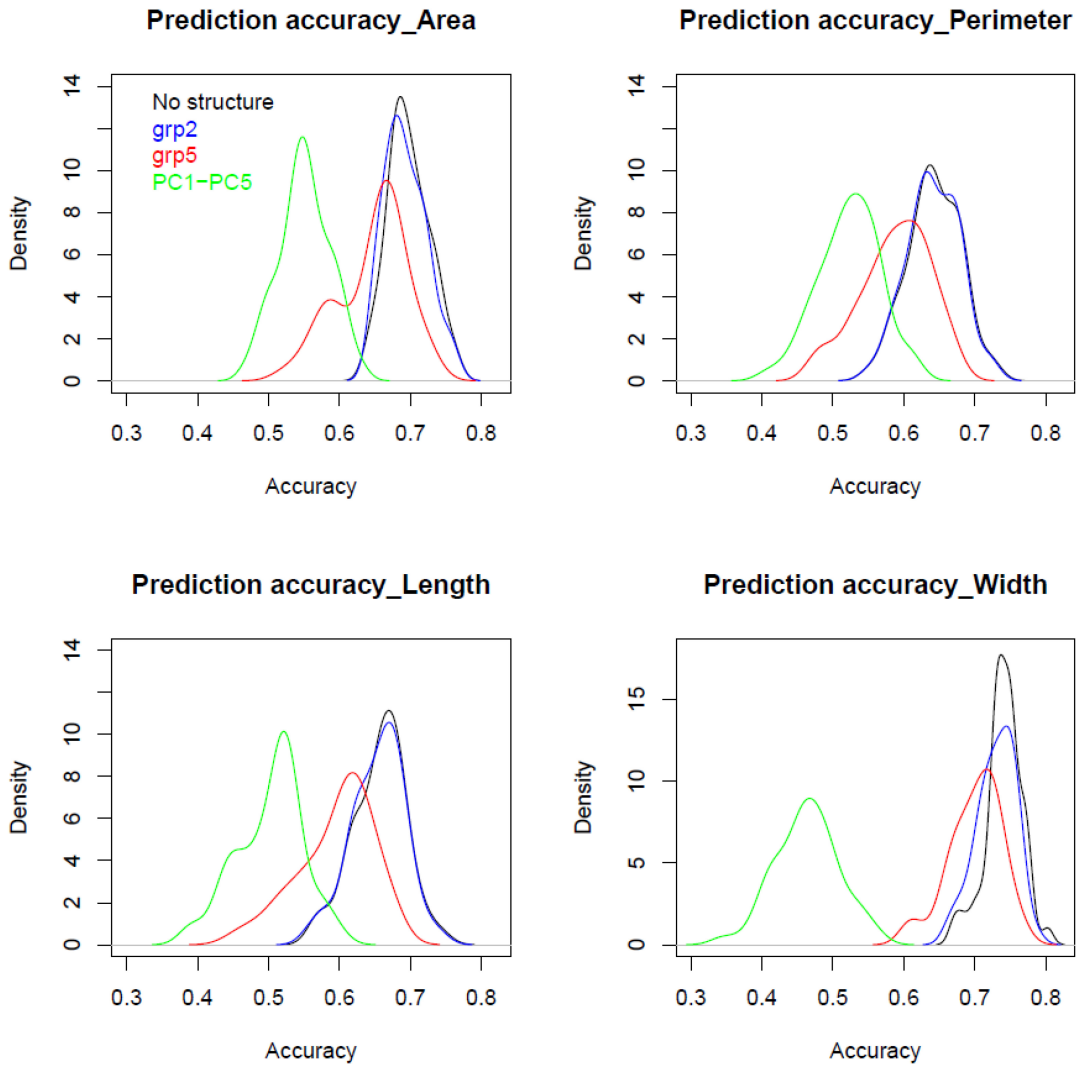


FIGURE 4 | Comparison between densities for prediction accuracy using validation set using different population structure covariates for grain area, perimeter, length, and width of bread wheat landraces.

a genomic relationship matrix to predict the performance of germplasm based on GEBVs, could be more reliable and useful to harness genetic gain from genetic resources (Bernardo, 2016). Since it has been shown that the environment of landrace's origin strongly influences gene flow and natural selection (Lin et al., 1975; Epperson, 1990), we have incorporated in this study an environmental similarity matrix based on landrace's passport data in addition to the genomic relationship matrix in the framework of GS. The increase in the prediction accuracy was noticeable but not significant. End-use and quality traits are the important factors that influence the market values, as well as the maintenance of landraces and then genetic diversity by the farmers (Negri, 2003; Seboka and van Hintum, 2006; Shewayrga and Sopade, 2011). Thus, we suspect that the grain traits used in this study were not only resulting from a natural selection but also affected by farmer selection and preferences. To summarize, genomic predictions for genebank accessions could benefit from using other characterization data such as phenology, morphology, and yield components.

Genebank Conservation and Use in the Era of Genomic Predictions

To safeguard future food, fiber and fuel resource, global germplasm conservation will increasingly rely on genomic technologies. Beyond the conservation aspects where identifying duplicates and redundancies between collections can be assessed by using genomics (Singh et al., 2019), there is an opportunity of using high-density markers to mine more efficiently genetic resources for better use of genebank accessions in pre-breeding programs (Rasheed et al., 2018). GS, for example, was identified as an optimal mining tool to identify genetic resources for quantitative traits, as also shown in the current study. Moving

forward, several challenges might limit the broad and routine use of GS, which include (1) the cost-effectiveness of genotyping, as the entire collection should be genotyped to take full advantage of GS; (2) aligning the genotyped and field-evaluated grains from the genebank; and finally, (3) dealing with population structure and forming the optimal training subset. The results in this study have shown that the use of passport information can be of a good start, but extra attention might be required for several collections that contain limited information on coordinates, especially for the old collections.

CONCLUSION

Evaluating the entire collection held by a genebank for all traits needed by breeding programs is resources consuming. Genebanks should stay innovative in the way where technologies could aid the identification of accessions that possess traits for new desirable variation. Our study demonstrated that genomic prediction has the potential of matching these outputs alone or augmented by passport information. This result will help breeders make better use of untapped genetic diversity.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to z.kehel@cgiar.org.

AUTHOR CONTRIBUTIONS

HA, AT, and AA delivered the seed and run the field experiment. AE and MS-G produced the grain scan data. ZK, MS-G, and CC analyzed the data. ZK wrote this article. AA, AT, MS-G, and CC provided insightful revisions and discussions. All authors reviewed the final version of the manuscript.

FUNDING

This study was funded by the CGIAR Genebank Platform and CRP Wheat.

ACKNOWLEDGMENTS

The genotyping characterization work was implemented by CIMMYT as part of the MasAgro Biodiversidad project in collaboration with ICARDA, made possible by the generous support of the Mexican Secretariat of Agriculture, Livestock, Rural Development, Fisheries and Food (SAGARPA) and CRP Wheat. Research work for CC is supported by the NSF-MRI 1626157.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00032/full#supplementary-material

REFERENCES

Anglin, N., Amri, A., Kehel, Z., and Ellis, D. (2018). A case of need: linking traits to genebank accessions. *Biopreserv. Biobank.* 16, 337–349. doi: 10.1089/bio.2018.0033

Aparicio, N., Villegas, D., Araus, J. L., Blanco, R., and Royo, C. (2002). Seedling development and biomass as affected by seed size and morphology in durum wheat. *J. Agric. Sci.* 139, 143–150. doi: 10.1017/s002185960020023 41411

Austin, R. B., Ford, M. A., and Morgan, C. L. (1989). Genetic improvement in the yield of winter wheat: a further evaluation. *J. Agric. Sci.* 112, 295–301. doi: 10.1007/s11032-017-0715-8

Azough, Z., Kehel, Z., Benomar, A., Bellafkih, M., and Amri, A. (2019). "Characterization of ICARDA genebank barley accessions," in *Intelligent Environments*, Vol. 26, Ambient Intelligence, 121–129.

Bernardo, R. (2016). Bandwagons I, too, have known. *Theor. Appl. Genet.* 129, 2323–2332. doi: 10.1007/s00122-016-2772-5

Brancourt-Hulmel, M., Doussinault, G., Lecomte, C., Berard, P., Le Buanec, B., and Trottet, M. (2003). Genetic improvement of agronomic traits of winter wheat cultivars released in France from 1946 to 1992. *Crop Sci.* 43, 37–45.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., and Bollmann, K. (2013). Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography* 36, 971–983. doi: 10.1111/j.1600-0587.2013.00138.x

Butler, D., Cullis, B., Gilmour, A., and Gogel, B. (2009). *ASReml R-Reference Manual*. Hemel Hempstead: VSN International Ltd.

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Jarquín, D., Franco, J., Juan, B., and Vikram, P. (2016). Genomic prediction of gene bank wheat landraces. *G3* 6, 1819–1834. doi: 10.1534/g3.116.029637

Curtis, B. C., Rajaram, S., and Gómez Macpherson, H. (2002). *Bread Wheat Improvement and Production*. Rome: FAO.

Daetwyler, H. D., Bansal, U. K., Bariana, H. S., Hayden, M. J., and Hayes, B. J. (2014). Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* 127, 1795–1803. doi: 10.1007/s00122-014-2341-8

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., and Hayes, B. J. (2015). Components of the accuracy of genomic prediction in a multi-breed sheep population. *J. Anim. Sci.* 90, 3375–3384. doi: 10.2527/jas2011-4557

de Azevedo Peixoto, L., Moellers, T. C., Zhang, J., Lorenz, A. J., Bhering, L. L., Beavis, W. D., et al. (2017). Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLoS One* 12:e0179191. doi: 10.1371/journal.pone.0179191

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

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi: 10.1534/genetics.109.101501

El Bouhssini, M., Street, K., Amri, A., Mackay, M., Ogbonnaya, F. C., Omran, A., et al. (2011). Sources of resistance in bread wheat to Russian wheat

aphid (*Diuraphis noxia*) in Syria identified using the focused identification of germplasm strategy (FIGS). *Plant Breed.* 130, 96–97. doi: 10.1111/j.1439-0523.2010.01814.x

El Bouhssini, M., Street, K., Joubi, A., Ibrahim, Z., and Rihawi, F. (2009). Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. *Genet. Resour. Crop Evol.* 56, 1065–1069. doi: 10.1007/s10722-009-9427-9421

Epperson, B. K. (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* 124, 757–771.

Evers, A. (2000). "Grain size and morphology: implications for quality," in *Wheat Structure, Biochemistry and Functionality*, ed. J. Schofield, (Cambridge, MA: The Royal Society of Chemistry), 19–24. doi: 10.1533/9781845698478.1.19

Evers, A. D., Cox, R. I., Shaheedullah, M. Z., and Withey, R. P. (1990). Predicting milling extraction rate by image analysis of wheat grains. *Asp. Appl. Biol.* 25, 417–426.

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

Gegas, V. C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., et al. (2010). A genetic framework for grain size and shape variation in wheat. *Plant Cell* 22, 1046–1056. doi: 10.1105/tpc.110.074153

González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J., Mahuku, G., et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125, 759–771. doi: 10.1007/s00122-012-1868-9

Gou, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., et al. (2014). The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet.* 127, 749–762. doi: 10.1007/s00122-013-2255-x

Grieve, C. M., and Francois, L. E. (1992). The importance of initial seed size in wheat plant response to salinity. *Plant Soil* 147, 197–205. doi: 10.1007/bf00029071

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Hu, X., Carver, B. F., Powers, C., Yan, L., and Chen, C. (2019). Genomic selectin and response to selection vy designed training population for grain yield and end-use quality traits in winter wheat variety development programs. *Plant Genome* 12. doi: 10.3835/plantgenomc2018.11.0090

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Khazaei, H., Street, K., Bari, A., Mackay, M., and Stoddard, F. (2013). The FIGS Focused Identification of Germplasm Strategy (FIGS) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources. *PLoS One* 8:e63107. doi: 10.1371/journal.pone.0063107

Levandi, T., Püssa, T., Vaher, M., Ingver, A., Koppel, R., and Kaljurand, M. (2014). Principal component analysis of HPLC–MS/MS patterns of wheat (*Triticum aestivum*) varieties. *Proc. Estonian Acad. Sci.* 63, 86–92.

Lin, W., Bradshaw, A. D., and Thurman, D. A. (1975). The potential for evolution of heavy metal tolerance in plants. III. The rapid evolution of copper tolerance in *Agrostis stolonifera*. *Heredity* 34, 165–187. doi: 10.1038/hdy.1975.21

Lopes, M. S., Reynolds, M. P., Jalal-Kamali, M. R., Moussa, M., Feltaous, Y., Tahir, I. S. A., et al. (2012). The yield correlations of selectable physiological traits in a population of advanced spring wheat lines grown in warm and drought environments. *Field Crops Res.* 128, 129–136. doi: 10.1016/j.fcr.2011.12.017

López-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *G3* 5, 569–582. doi: 10.1534/g3.114.016097

Mackay, M. C., and Street, K. (2004). "Focused identification of germplasm strategy—FIGS," in *Cereals 2004, Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly Cereal Chemistry Division*, eds C. K. Black, J. F. Panozzo, and G. J. Rebetzke, (Melbourne: Royal Australian Chemical Institute), 138–141.

Manickavelu, A., Hattori, T., Yamaoka, S., Yoshimura, K., Kondou, Y., Onogi, A., et al. (2017). Genetic nature of elemental contents in wheat grains and its genomic prediction: toward the effective use of wheat landraces from Afghanistan. *PLoS One* 12:e0169416. doi: 10.1371/journal.pone.0169416

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

Negri, V. (2003). Landraces in central Italy: where and why they are conserved and perspectives for their on-farm conservation. *Genet. Resour. Crop Evol.* 50, 871–885.

Nesbitt, M. (2002). "When and where did domesticated cereals first occur in southwest Asia?," in *The Dawn of Farming in the Near East. Studies in Early Near Eastern Production, Subsistence, and Environment*, eds R. T. J. Cappers, and S. Bottema, (Berlin: Ex Oriente), 113–132.

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3* 8, 2889–2899. doi: 10.1534/g3.118.200311

Pasam, R. K., Bansal, U., Daetwyler, H. D., Forrest, K. L., Wong, D., Petkowski, J., et al. (2017). Detection and validation of genomic regions associated with resistance to rust diseases in a worldwide hexaploid wheat landrace collection using BayesR and mixed linear model approaches. *Theor. Appl. Genet.* 130, 777–793. doi: 10.1007/s00122-016-2851-7

Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manes, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric models for genome-enabled prediction in wheat. *G3* 2, 1595–1605. doi: 10.1534/g3.112.003665

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F., He, Z., and Rajaram, S. (2018). Wheat genetic resources in the post-genomics era: promise and challenges. *Ann. Bot.* 121, 603–616. doi: 10.1093/aob/mcx148

Royo, C., Álvaro, F., Martos, V., Ramdani, A., Isidro, J., Villegas, D., et al. (2007). Genetic changes in durum wheat yield components and associated traits in Italian and Spanish varieties during the 20th century. *Euphytica* 155, 259–270. doi: 10.1007/s10681-006-9327-9

Sanchez-Garcia, M., Royo, C., Aparicio, N., Martín-Sánchez, J. A., and Álvaro, F. (2013). Genetic improvement of bread wheat yield and associated traits in Spain during the 20th century. *J. Agric. Sci.* 151, 105–118. doi: 10.1017/S0021859612000330

Seboka, B., and van Hintum, T. (2006). The dynamics of on-farm management of sorghum in Ethiopia: implication for the conservation and improvement of plant genetic resources. *Genet. Resour. Crop Evol.* 53, 1385–1403. doi: 10.1007/s10722-005-5676-9

Shewayrga, H., and Sopade, P. (2011). Ethnobotany, diverse food uses, claimed health benefits and implications on conservation of barley landraces in North Eastern Ethiopia highlands. *J. Ethnobiol. Ethnomed.* 7:19. doi: 10.1186/1746-4269-7-19

Siddique, K. H. M., Belford, R. K., Perry, M. W., and Tennant, D. (1989). Growth, development and light interception of old and modern wheat cultivars in a Mediterranean-type environment. *Austral. J. Agric. Res.* 40, 473–487.

Singh, N., Wu, S., Raupp, W. J., Sehgal, S., Arora, S., Tiwari, V., et al. (2019). Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep.* 9:410779. doi: 10.1038/s41598-018-37269-0

Thorwarth, P., Yousef, E. A. A., and Schmid, K. J. (2017). Genomic prediction and association mapping of curd-related traits in genebank accessions of cauliflower. *G3* 8, 707–718. doi: 10.1534/g3.117.300199

Title, P. O., and Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* 41, 291–307. doi: 10.1111/ecog.02880

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull. Bull.* 37, 33–36.

Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Front. Genet.* 9:78. doi: 10.3389/fgene.2018.00078

Whan, A. P., Smith, A. B., Cavanagh, C. R., Ral, J. P. F., Shaw, L. M., Howitt, C. A., et al. (2014). GrainScan: a low cost, fast method for grain size and colour measurements. *Plant Methods* 10, 1–10. doi: 10.1186/1746-4811-10-23

Zhang, X., Sallam, A., Gao, L., Kantarski, T., Poland, J., DeHaan, L. R., et al. (2016). Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. *Plant Genome* 9, 1–18. doi: 10.3835/plantgenome2015.07.0059

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*, 4th Edn. Oxford: Oxford University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Kehel, Sanchez-Garcia, El Baouchi, Aberkane, Tsivelikas, Charles and Amri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fgene.2020.00316

Combining QTL Analysis and Genomic Predictions for Four Durum Wheat Populations Under Drought Conditions

Meryem Zaïm[1,2†], Hafssa Kabbaj[1,2†], Zakaria Kehel[2], Gregor Gorjanc[3], Abdelkarim Filali-Maltouf[1], Bouchra Belkadi[1], Miloudi M. Nachit[2] and Filippo M. Bassi[2]*

[1] Laboratory of Microbiology and Molecular Biology, Faculty of Sciences, Mohammed V University, Rabat, Morocco, [2] ICARDA, Biodiversity and Integrated Gene Management, Rabat, Morocco, [3] The Roslin Institute, The University of Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

Edited by:
Alison Bentley,
National Institute of Agricultural Botany (NIAB), United Kingdom

Reviewed by:
Guanglin He,
Sichuan University, China
Changwei Shao,
Yellow Sea Fisheries Research Institute (CAFS), China

***Correspondence:**
Filippo M. Bassi
f.bassi@cgiar.org

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 21 September 2019
Accepted: 16 March 2020
Published: 06 May 2020

Citation:
Zaïm M, Kabbaj H, Kehel Z, Gorjanc G, Filali-Maltouf A, Belkadi B, Nachit MM and Bassi FM (2020) Combining QTL Analysis and Genomic Predictions for Four Durum Wheat Populations Under Drought Conditions. Front. Genet. 11:316. doi: 10.3389/fgene.2020.00316

Durum wheat is an important crop for the human diet and its consumption is gaining popularity. In order to ensure that durum wheat production maintains the pace with the increase in demand, it is necessary to raise productivity by approximately 1.5% per year. To deliver this level of annual genetic gain the incorporation of molecular strategies has been proposed as a key solution. Here, four RILs populations were used to conduct QTL discovery for grain yield (GY) and 1,000 kernel weight (TKW). A total of 576 individuals were sown at three locations in Morocco and one in Lebanon. These individuals were genotyped by sequencing with 3,202 high-confidence polymorphic markers, to derive a consensus genetic map of 2,705.7 cM, which was used to impute any missing data. Six QTLs were found to be associated with GY and independent from flowering time on chromosomes 2B, 4A, 5B, 7A and 7B, explaining a phenotypic variation (PV) ranging from 4.3 to 13.4%. The same populations were used to train genomic prediction models incorporating the relationship matrix, the genotype by environment interaction, and marker by environment interaction, to reveal significant advantages for models incorporating the marker effect. Using training populations (TP) in full sibs relationships with the validation population (VP) was shown to be the only effective strategy, with accuracies reaching 0.35–0.47 for GY. Reducing the number of markers to 10% of the whole set, and the TP size to 20% resulted in non-significant changes in accuracies. The QTLs identified were also incorporated in the models as fixed effects, showing significant accuracy gain for all four populations. Our results confirm that the prediction accuracy depends considerably on the relatedness between TP and VP, but not on the number of markers and size of TP used. Furthermore, feeding the model with information on markers associated with QTLs increased the overall accuracy.

Keywords: genomic selection, consensus map, drought, imputation, QTL analysis, fixed effect, consensus map, genotyping by sequencing (GBS)

INTRODUCTION

Durum wheat (*Triticum durum* Desf., 2n = 4x = 28, AABB) is grown annually on over 17 million hectares worldwide, and it represents one of the bases of the Mediterranean diet. This region is the largest consumer of durum wheat products and the most significant durum import market (Soriano et al., 2017). The Mediterranean basin is subject to frequent droughts and their occurrence is expected to raise in the near future, with a significant negative effect on crop development and production (Xiao et al., 2018). Breeding for durum genotypes that have an improved yield and tolerance to drought remains one of the most strategic methods to protect the harvest of this crop (Habash et al., 2009; Tadesse et al., 2016; Kuzmanoviæ et al., 2018). The use of genomic models to analyze the main drought adaptation traits can be deployed to significantly accelerate the breeding effort. Genetic linkage map and QTL mapping are useful tools for discovering genomic regions associated with traits of interest (Zhang et al., 2018). However, the significance of the identified QTLs is often linked to the specific parents used and it rarely proved useful for deployment in large scale breeding. One method to control for this error is to perform QTL discovery in multiple populations at the same time. The first step to achieve this is the development of genetic consensus maps that allow to bridge the discovery across populations. In fact, the development of consensus maps has already been shown to not only bridge the information between populations, but also to increase marker density, improve genome coverage, provide a validation of the marker ordering, and reduce markers gaps due to the absence of polymorphism between two parents (Marone et al., 2012; Maccaferri et al., 2014). Multiple genetic linkage maps have already been developed for wheat, and consensus genetic maps have been constructed for hexaploid wheat (Somers et al., 2004; Wang et al., 2014) and durum wheat (Maccaferri et al., 2014, 2015). Furthermore, high-throughput DNA sequencing technologies have now enabled the deployment of reliable and affordable marker coverage via genotype-by-sequencing (GBS), a methodology that relies on restriction enzymes to reduce the amount of genome to be sequenced (Poland et al., 2012; Edae et al., 2017). Numerous recent studies have used this marker system to identify quantitative trait loci (QTL) associated with yield, agronomic traits, and physiologic traits in drought and heat-stressed environments (Acuña-Galindo et al., 2015; Sukumaran et al., 2016; Edae et al., 2017; Hussain et al., 2017; Mwadzingeni et al., 2017; Asif et al., 2018; Bhatta et al., 2018; Roselló et al., 2019), in order to pyramid these QTLs via marker-assisted breeding (Edae et al., 2014).

Genomic selection (GS) builds on the concept of QTL analysis, but it explores the whole genome seeking large and small allelic effects (Bassi et al., 2016). Because of its capacity to better handle complex traits with several small effect alleles such as grain yield (GY), GS is now becoming the methodology of choice for incorporation into breeding strategies (Dekkers and Hospital, 2002; Crosbie et al., 2003; Bassi et al., 2016). GS analyzes jointly all markers to explain the total phenotypic variance through the sum of the markers effects (Meuwissen et al., 2001). Once a model is trained, an effect is assigned to each marker-allele, and the 'genomic estimated breeding value' (GEBVs; Meuwissen et al., 2001) can then be calculated for each individual as the sum of its allelic marker effects. The set of individuals used to train the model has both phenotypic and genotypic available and it is defined as the 'training population' (TP). The set of individuals from which the selection is made is defined as the 'breeding population' (BP), and only genotypic data are collected for it. The 'accuracy' of the predicted GEBV is determined by the correlation between GEBV and the true breeding value (TBV) calculated phenotypically for a 'validation population' (VP), which is genotyped and phenotyped, but not used to train the model. The value for accuracy is used to determine the overall success of the GS approach. Therefore, it is important to maintain a high degree of accuracy, and hence to use a TP that best fits the BP. The degree of relatedness between the two populations is often a good predictor of the accuracy that will be achieved. Cross-validation is used to train and develop the prediction models using different sampling techniques in the TP data sets ahead of estimating the GEBVs in the VP. The idea behind this approach is that breeders can derive predictions of the breeding value of an experimental line even before the line has been tested in the field. In turn, this would allow to make decisions on the use of the lines for yield testing or crossing already during the earlier generations (Crossa et al., 2010; Heffner et al., 2011; Bassi et al., 2016).

However, the integration of QTL analysis and GS remains severely understudied. In the present study, four recombinant inbreed lines (RILs) of durum wheat with different level of relatedness were field tested across environments. QTL analysis was performed for GY and TKW and the same populations were then used to assess different GS models for the two traits. The two methods were then combined by fixing the effect of the marker underlying the QTLs into GS models, to reveal a steep increase in the overall accuracy.

MATERIALS AND METHODS

Mapping Populations

Four F_9-derived RILs mapping populations were obtained by random selection of 200 individual durum spikes from each population at the F_4 generation, followed by single seed descent to F_9. At this generation, the individual plants were sampled for DNA extraction, and the seeds of each individual plant bulked. A different number of individuals for each population was then multiplied and used for yield trial to resemble the typical unbalanced dataset used by breeders. The four durum wheat crosses combining ICARDA's elite lines were: Icamor/Gidara2 (IC; 115 RILs) developed by combining the *Hessian fly* resistance of Icamor (F413J.S/3/Arthur71/Lahn//Blk2/Lahn/4/Quarmal) with the high yield potential of Gidara2 (Stojocri/Omrabi3) (see Bassi et al., 2019 for more details); the second population was Jennah Khetifa/Cham1//T.dicoccoides600545/2*Omrabi5 (DRO; 197 RILs) designed for pyramiding the drought tolerance of the Tunisian landrace Jennah Khetifa, wild emmer, and the ICARDA most successful variety Omrabi; the third population was SW Algia//Gidara1/Cham1 (SW; 93 RILs)

aimed at incorporating the *Septoria tritici* resistance of the Tunisian landrace SW Algia with Gidara1; the fourth population was Omrabi3/Omsnima1//Gidara2 (YG; 145 RILs) aimed at combining drought tolerance and yield potential. As indicated, these populations all have sibling relationships with Omrabi, Cham 1, and Gidara used as parental lines. Additional details are reported in **Table 1**.

Field Trials

Field trials were conducted during the 2014–2015 growing season. The experimental design used at all stations was an augmented complete block design with four common repeated checks, and a block size of 24 entries. The trials were conducted at three drought prone stations in Morocco (**Supplementary Figure S1**): Jemaat Shaim (JSH; 32°21′0″ N and 8°51′0″ W), Marchouch (MCH; 33°34′3.1″ N and 6°38′0.1″ W) and Sidi el Aidi (SAD; 33°9′36″ N and 7°24′0″ W); and one irrigated station in Lebanon: Terbol (TER; 33°48′29″ N and 35°59′22″ W) (**Table 2** and **Supplementary Figure S1**). All RILs and their parents were planted in plots of 4.2 m^2 at a seeding rate of 280 plants per m^2. The YG population was planted in MCH, JSH, SAD and TER; the DRO population was also planted in all stations except TER; the IC population was sown in two stations MCH and TER; the SW population in just MCH. Agronomic practices were done following standard procedures, with 80 units of nitrogen provided in 2 equal splits, and 40 units of potassium and phosphorous before planting. Weeds were control by tank mixtures of Derby and Pallas. Days to heading (DTH), days to maturity (DTM), plant height (PLH), and spike density per m^2 (SPK) were recorded in MCH and TER. At maturity, 3 m^2 of the plot were combine harvested and the weight was converted to grain yield as Kg ha^{-1}. At all stations except SAD, 1,000 kernels were weighted on a precision balance to derive 1,000-kernels weight (TKW) and express it in grams (g).

DNA Extraction and Genotyping

Leaf samples obtained from F_9 plants were freeze-dried and used for C-TAB DNA extraction. DNA quality was assessed on agarose gel and it was then equilibrated to 100 ng. The DNA was shipped to the Poland lab at Kansas State University for genotyping by sequencing following the protocol of Poland et al. (2012). Briefly, two restriction enzymes (*Pst*I and *Msp*I) were used for genome complexity reduction, followed by 96-multiplex sequencing by bar coding. Low-quality data filtering was carried out according to the following rules: heterozygous calls not superior to 2%, maximum of 30% missing data, and a minor allele frequency superior to 10%.

Consensus Map Procedure

Individual linkage maps for each population were constructed using the statistical software Carthagene v. 1.2.3 (De Givry et al., 2005) and QTL IciMapping V4.1 (Meng et al., 2015). First, all marker sequences were aligned to the available bread wheat genome assembly (Winfield et al., 2016; The International Wheat Genome Sequencing Consortium [IWGSC], 2018) by BLAST with an identity cut-off of 98% (1 SNP variant) and *E*-value of $5e^{-25}$. The *squeeze* function of Carthagene was used to eliminate markers that were wrongly ordered at LOD of 5 based on the genome alignment, followed by *flip* with window size of seven, LOD of 3, and zero iterations to determine the most plausible order of markers within each window. This framework map contained correctly aligned markers along the map and several unassigned markers. In QTL IciMapping, the framework markers were *anchored* while the unassigned markers were not. The *by anchor order* algorithm was used to assign to the different linkage groups the unassigned markers at a set LOD of 5, and then order them based on the position of the framework markers. This operation was then repeated using the newly developed framework map and reducing the LOD to 3. This methodology defined four individual genetic maps for each population.

The construction of the consensus map was performed chromosome by chromosome using the *consensus map from multiple linkage maps sharing common markers* (CMP) function of QTL IciMapping. First, by re-grouping markers at a distance of less than 20 cM to obtain one group for each chromosome, followed by the *by anchor order* option to measure the genetic distances between markers along the consensus map based on their relative positions on each individual map. Markers were then ordered based on their consensus map position in an Excel file. In several cases, a marker polymorphic in one population might be monomorphic in another. To avoid linkage distortions, the monomorphic scores were set to missing. At this point, imputation was done using AlphaIMpute option HMM (Hickey et al., 2012; Antolin et al., 2017) and confirmed with the BIP function of QTL IciMapping (Zhang et al., 2010).

Data Analysis and QTL Mapping

Statistical analysis of the phenotypic data was performed using the R software version 3.4.3 and Genstat program version 18. Best linear unbiased estimates (BLUEs) were estimated across all environments, assuming fixed effects for the genotype from a linear mixed-effects model using R package *lme4* (Bates et al., 2015; R Core Team, 2017).

The discriminant analysis of principal components (DAPC), was performed using the 'adegenet' package 1.4-1 (Jombart et al.,

TABLE 1 | Cluster analysis of the genetic diversity among four mapping populations using discriminant analysis of principal components (DAPC) with $k = 4$, their pedigrees, and maps features.

Pedigree	Individuals	Markers	Total length (cM)	Marker density (cM/Marker)
IC: Icamor/Gidara2	115	646	1720.1	5.3
DRO: Jennah Khetifa/Cham1// T.dicoccoides600545/2*Omrabi5	197	2291	1922.5	1.2
SW: SW Algia//Gidara1/Cham1	93	1212	1795.3	1.8
YG: Omrabi3/Omsnima1//Gidara2	145	521	1683.8	6.1

TABLE 2 | Description of the field testing environments during the 2014–2015 season.

Code	Site	Country	Coordinates	Altitude (m)	Soil type	Climate	Moisture	Annual rainfall (mm)
MCH15	Marchouch	Morocco	33° 34′ 3.1″ N, 6° 38′ 0.1″ W	398	Clay vertisol	Mediterranean/warm temperate	Rainfed	449
SAD15	Sidi el Aydi	Morocco	33° 9′ 36″ N, 7° 24′ 0″ W	226	Vertisol	Mediterranean/hot and temperate	Rainfed	237
JSH15	Jemhâa Shaim	Morocco	32° 21′ 0″ N, 8° 51′ 0″ W	196	Calcic Cambisols	Hot steppe	Rainfed	270
TER15	Terbol	Lebanon	33° 48′ 29″ N, 35° 59′ 22″ W	897	Chromic Vertisols	Mediterranean/temperate	Sprinkle	559

2010) in R studio V 3.4.3 (R Core Team, 2017). With DAPC, the hierarchical clustering among populations was determined by applying the R based package "hclust."

QTLs were searched for each individual population in each individual environment via composite interval mapping (CIM) analysis using R/qtl (Broman et al., 2003). The *cim* function was set to five markers covariates and a window size of 10 cM. LOD thresholds were calculated from QTL IciMapping by BIP functionality using 1,000 permutations with a maximum type 1 error probability of 0.05. Only QTLs that appeared at least in two environments and two populations were considered as valid. The distribution of QTLs and the marker density of the consensus and individual population maps were graphically presented on the fourteen chromosomes of durum wheat by a "Circos plot" using R/shiny application (Yu et al., 2018).

Genomic Prediction Modeling

A total of four genomic models were tested as a first step in this study:

(i) a baseline additive model without interactions of genotypic effect (G), environmental (E) effect, and error (ε) (G+E + ε).

(ii) a baseline multi-environment model (G+E + GxE + ε), which assumed interactions between the G and the E.

In both these models, all the effects were assumed to be random with a normal distribution N(0, σ) where σ is the term variance

(iii) the third model was a marker (M) effect model (G+E + GxE + M + ε), where the genotype effect is substituted by an approximation of the genotype's genomic value expressed as a regression on marker covariates.

In this case the model assumes that the genotype's genomic value follows a normal distribution N(0, G σ_g) where σ_g is the genetic variance and G is genomic relationship matrix.

(iv) the last model is the marker × environment model (G+E + GxE + M + MxE + ε) where the marker effect is composed by an effect common to all environment (main effect) plus a random deviation specific to a particular environment (Lopez-Cruz et al., 2015).

Testing of the different models' accuracies was done using DRO, IG and YG populations independently, and setting as cross-validation 80% of the individuals as TP and 20% as VP. The accuracies within and across environments were then calculated as a measure of good fit. The BGLR package (Pérez and de Los Campos, 2014) was used to run all models above from (i) to (vii) by Bayesian ridge regression (BRR) using 10,000 iterations and 5,000 burn in, and 50 replications (de los Campos et al., 2009, 2013). This model induces homogeneous shrinkage of all marker effects toward zero and yields a Gaussian distribution of marker effects. The 50 replications were used to define statistical differences between model accuracies following a one factor ANOVA.

The GxE + MxE model (iv) was selected and used to test additional hypothesis:

(v) the effect of markers number was investigated by comparing predictions using 100, 80, 60, 40, 20, and 10% of the total marker set in combination with reducing the TP population size to 20, 50, and 75% for GY and TKW. The TP individuals were selected randomly in 50 replications, and one factor ANOVA was used to determine significant differences.

(vi) the prediction accuracy of using half sibs vs. full sibs as TP was compared. Each population was set as TP for all others and itself using the whole population as TP and the whole other population as VP.

(vii) to compare the value of MAS and GS, the prediction accuracy was calculated using 50% as TP and 50% as VP for all markers, only markers associated with major effect QTLs, with 44 and 27 markers for GY and TKW, respectively, and by removing these markers linked to QTLs from the set. The TP individuals were selected randomly in 50 replications, and one factor ANOVA was used to determine significant differences.

(viii) the rr-BLUP package v4.6 (Endelman, 2011) was used to run a mixed model estimating the accuracy gain when using markers underlying the QTLs as fixed effects, and the remaining markers as random effects. For this analysis ten random subsets of 50% TP and 50% VP were selected in each population separately (DRO, IG, SW, and YG). QTL analysis was conducted again for each TP subset following the method described above. Those markers that resulted as underlying QTLs in each TP subset were fixed in the model. One factor ANOVA was run for the ten replicates of each population to determine significant differences.

RESULTS

Phenotypic Evaluation

Analysis of variance (ANOVA) showed significant differences for genetic (G) effect ($p < 0.05$) for all the traits across environments, indicating good levels of phenotypic within each population (**Table 3** and **Figure 1**). The genotype by environment interaction (GxE) effect was also significant ($p < 0.05$). The combined BLUE of TKW and GY differed greatly between the two parental lines of the four populations, displaying a normal distribution within RILs populations (**Figure 1**). Gidara 2 and Jk/Ch1 parents in populations IC, DRO and YG had smaller values of TKW than the average, whereas the Icamor parent in population IC had the maximum value (44 g). Similarly, for GY, Gidara 2 had a smaller value than the average GY, same for the parents Icamor and Younes. Cham1 parent of population DRO and SW had the highest recorded GY of this experiment. The population YG had the highest average TKW and GY. Among the four RILs populations, 50.2 g was the highest value recorded for TKW found in IC, and 3,304 kg ha^{-1} the highest GY for YG.

Individual and Consensus Linkage Maps

The GBS process resulted in 22,117 marker calls. Among these, 4,909 matched the curation criteria and were tentatively ordered via genetic mapping. The individual genetic maps contained 646 polymorphic markers covering 1,720.1 cM for the IC population, 2,291 markers spanned 1,922.5 cM for DRO, 1,212 markers were mapped along 1,795.2 cM in SW, and 521 markers over 1,683.7 cM for YG (**Table 4** and **Supplementary Table S1**). The final consensus map incorporated 3,202 markers assigned to 14 linkage groups corresponding to 1,883 unique loci, and spanned a total genetic distance of 2,705.7 cM, with a density of one marker each 0.85 cM (**Table 4**). The A genome, harbored 1,104 markers, covering a linkage distance of 1,133.8 cM, and the B genome 2,098 markers spanning a linkage distance of 1,572 cM. The largest chromosome was 2B, consisting of 540 markers and covering a genetic length of 243.5 cM, while the smallest chromosome in the map was 4A, covering a genetic length of 101.7 cM and consisting of 209 markers. The average size of markers gaps in the consensus map was 22.1 cM. The consensus map across four populations includes 550 RILS lines. Genetic diversity analysis revealed close kinship between IC and DRO, a lower relatedness with SW, and limited kinship to YG (**Table 1**).

QTL Analysis

The identified genetic and phenotypic variations were combined via QTL analysis across the 550 RILs for all measured traits. Significant QTLs were detected for all traits as summarized in **Figure 2** (**Supplementary Tables S2**, **S3**). A total of 31 QTLs

TABLE 3 | Rate of genetic effect across environments of four populations (IC, DRO, SW, and YG) for DTH, DTM, PLH, SPK, TKW, and GY and genotype by environment interactions (GxE) effects.

Pop	GY across env.		DTH			DTM			PLH			SPK	TKW		
	GxE	G	MCH	SAD	TER	MCH	SAD	TER	MCH	SAD	TER	MCH	MCH	TER	JSH
IC	–	0.93*	0.44*	–	0.74*	0.94*	–	0.85	0.93*	–	0.89*	0.95	0.85*	0.99*	–
DRO	0.45*	0.53*	0.90*	1.00*	–	1.00*	0.86*	–	0.99*	0.98*	–	0.97*	0.95*	–	0.97*
SW	–	0.81*	0.94*	–	–	0.74	–	–	0.96*	–	–	0.89	0.97*	–	–
YG	0.63*	0.36*	0.90*	–	0.93*	0.92*	–	0.76*	1.00*	–	0.89*	0.99*	0.95*	0.93*	0.98*

**Significant at 0.05 probability level; –, not available data, GxE, genotype by environment interaction effect; G genetic effect; DTH, days to heading; DTM, days to maturity; PLH, plant height; SPK, spike density; TKW, 1000 kernel weight; MCH, Marchouch; SAD, Sidi el Aidi; TER, Terbol; JSH, JemaatShaim.*

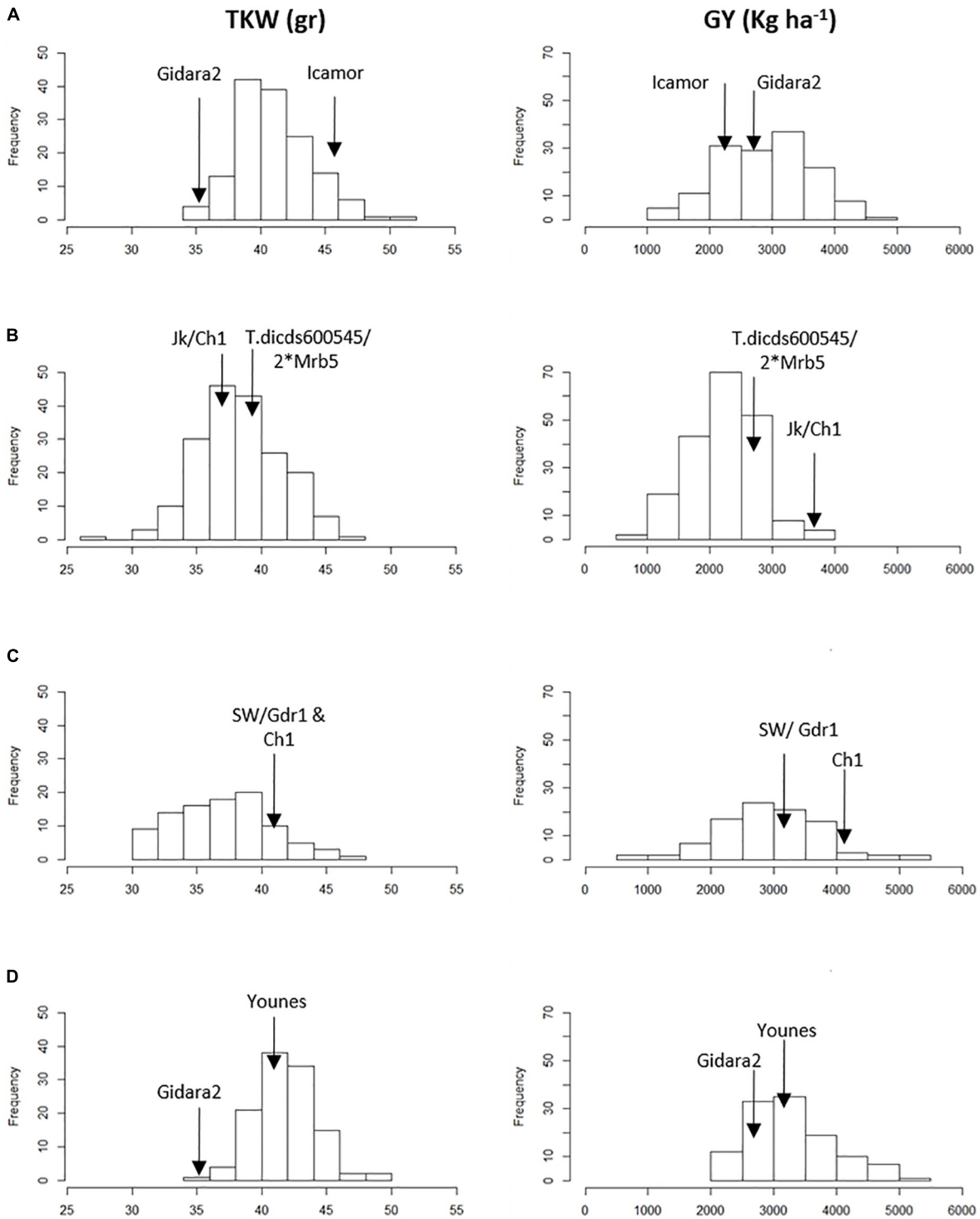


FIGURE 1 | Frequency distribution of 1,000 kernel weight (TKW) and grain yield (GY) in the parents and the four RIL populations. **(A)** IC, **(B)** DRO, **(C)** SW, and **(D)** YG.

were detected across the four populations, explaining from 3.9 to 81.3% of the PV and LOD diverging from 3.7 to 43.5. Six QTLs were found to be associated with GY and independent from the flowering time. In particular, on chromosomes 2B, 4A, and 5B the four independent populations identified consistently the same GY QTL. Six QTLs were detected for TKW on

TABLE 4 | Characteristics of the consensus map.

Chr.	Markers	Loci	Length (cM)	Marker density (cM/Marker)	Size of largest gap (cM)
1A	118	72	138.8	1.2	26.7
1B	257	106	228.5	0.9	24.9
2A	220	164	135.6	0.6	17.9
2B	540	361	243.5	0.5	16.6
3A	146	74	199.5	1.4	21.1
3B	302	189	238.1	0.8	32.6
4A	209	130	101.7	0.5	6.3
4B	197	125	208.3	1.1	29.9
5A	105	38	217.5	2.1	29.7
5B	302	162	245.0	0.8	16.6
6A	162	75	171.8	1.1	17.5
6B	246	155	181.9	0.7	16.6
7A	144	80	168.9	1.2	20.8
7B	254	152	226.6	0.9	31.7
A genome	1104	633	1133.8	1.0	29.7
B genome	2098	1250	1572.0	0.7	32.6

chromosomes 1B, 4B, 6A, 6B, and 7A, explaining 4.7–15.9 of PV and with maximum LOD of 6.1. Interestingly, loci controlling TKW were found to be also associated to GY on chromosome 2B, explaining 8.6 and 4.8% of PV, and LOD of 4.7 and 4.3 respectively.

Genomic Prediction: Identification of the Best Fitting Model (*i, ii, iii, iv*)

Four statistical models (*i, ii, iii, iv*) were tested to determine the best model to be used for each population (**Figure 3**). Non-significant differences could be identified for the IG population with average accuracies that ranged from 0.42 to 0.41. For DRO, the incorporation of the M effect resulted in a significant increase in accuracy from 0.47 to 0.49. The YG population was the most sensitive to the change of model ranging from 0.27 for models without M (*i* and *ii*), to 0.30 for model *iii*, to 0.33 for model *iv* incorporating GxE + MxE. Following these results, the model incorporating GxE + MxE was chosen to be the best suited for all three populations. For the SW population phenotypic data were available only for one environment, therefore a model using only markers effect (*iii*) was used to run genomic predictions for this population.

Genomic Prediction: Effect of Reducing TP and Marker Size (*v*)

The effect of marker number and TP size on prediction accuracies was tested for GY and TKW (v). **Figure 4** shows that when decreasing the number of markers from 3,202 to 320, a slight decrease in prediction accuracies was observed for the different set of TP. For GY, the reduction of markers caused a shift from 0.44 to 0.41 accuracy using 20% of TP, from 0.47 to 0.43 and from 0.49 to 0.44 for 50 and 75% of TP, respectively. For TKW, it dropped from 0.75 to 0.73 and from 0.76 to 0.74 for 20 and 50% of the TP, respectively, while no difference was observed for the 75% of TP between the total number of marker and 10% of it. Statistical analysis revealed no significant differences could be observed when reducing marker number and TP size for any of the two traits.

Genomic Prediction: Importance of Relatedness Between TP and VP (*vi*)

The four populations share common parents and have hence kinship relationships (**Table 1**). It was therefore evaluated if it would be possible to use one population as TP for the others (VP) which have half-sibs relationships. Using TP that were full sibs to the VP resulted in good accuracy values that ranged from 0.35 to 0.47, and from 0.92 to 0.30 for GY and TKW, respectively (**Table 5**). When the TP was not derived from the same cross of the BP (half sibs), the accuracies drop to values close to zero or even negative (**Table 5**). The only acceptable case for GY with an accuracy of 0.29 was obtained when SW was used as TP for IG, but this was not true when IG was used as TP for SW (accuracy of 0.08). The same was observed for TKW, with SW as TP ensuring an accuracy of 0.22, while YG as TP dropped to 0.09 accuracy. Interestingly, the two most genetically related populations, IG and DRO (**Table 1**) also resulted in very poor prediction accuracies when used as TP for each other.

Genomic Prediction: Effect of QTL Analysis on Model Accuracy (*vi, viii*)

Since QTL analysis and GS have been rarely combined, the last objective of this study was to determine if a step of QTL analysis could help improve the GS model's accuracy. A total of 44 and 27 markers were associated via QTL analysis to GY and TKW, respectively (**Figure 2**). To test their value alone, these were used as the only marker to perform genomic predictions and resulted in non-significant accuracies for GY for DRO (0.18), and IG (−0.02), while significant accuracies could be identified for YG (0.29), while an increased was observed for SW (0.54). Similarly, for TKW there was a loss significance for DRO (0.20), IG (0.11) and YG (0.09), while it again increased for SW (0.54) (**Figure 5**). The opposite attempt was also conducted by removing from the whole set all the markers associated with QTLs. In this case the GY and TKW accuracies became non-significant for all populations, except for SW for which it matched what was obtained when using the full marker dataset (**Figure 5**). With the exception of SW, for which the use of only markers associated to QTLs had a positive effect on the prediction accuracies, in all other populations the use of all markers combined was significantly superior.

As it can be expected, the sum of the accuracies of using markers associated to large and small effects does not equal to the accuracy of these combined. It then becomes interesting to assess a model that better incorporates these two by fixing the effect of markers associated to QTLs, while including the random effect of the small impact alleles (viii). To test the suitability to do so in a context that better resembles an actual breeding pipeline, QTL discovery was re-run for each random group of entries composing the TP, and only QTL that could be identified by the specific TP where fixed in the model. **Supplementary**

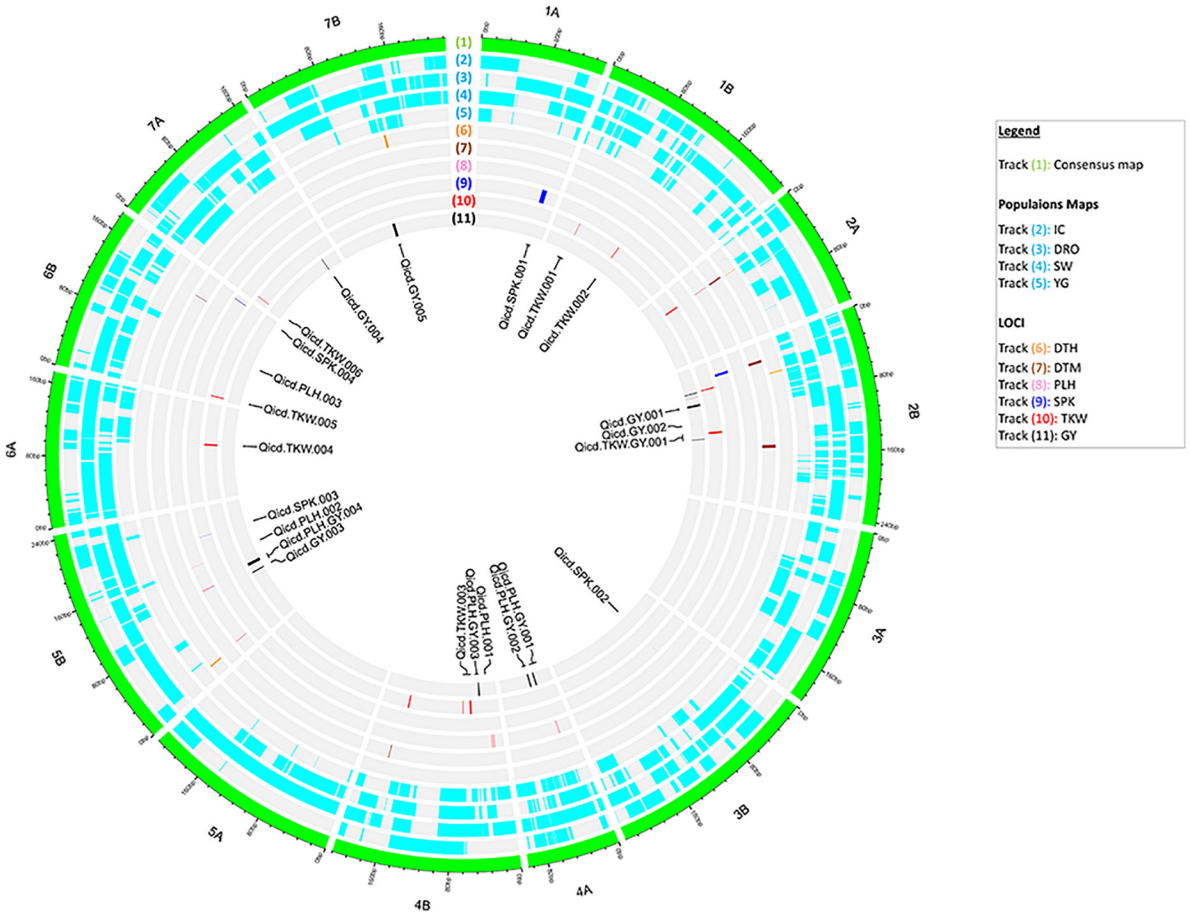


FIGURE 2 | Circos representation of the consensus map with aligned the marker density within each population, and the identified QTL for all traits. From outer to inner layer: **(1)** the consensus map with 14 chromosomes of durum wheat; **(2–5)** the genetic maps of the four populations: IC, DRO, SW and YG; **(6–11)** distribution of significant markers identified via QTL analysis for DTH, DTM, PLH, SPK, TKW, and GY. In the center, labels for QTLs of PLH, SPK, TKW, and GY independent from flowering time.

Table S4 reports how frequently the QTL associated with GY could be re-identified for each TP sub-set. The results of fixing the marker underlying the QTLs in the model is reported in **Figure 6**. For all four populations the accuracies increased significantly ($p < 0.05$) when the QTL-underlying markers were fixed in the model. The average accuracies shifted from 0.35 to 0.47, 0.38 to 0.44, 0.29 to 0.35, and 0.35 to 0.41, for the YG, DRO, IG, and SW populations, respectively. This represents a clear gain of 0.06–0.12 points of accuracy, superior than the 0.01–0.03 obtained by testing different statistical models (*i, ii, iii, iv*).

DISCUSSION

Rapid genetic gain for complex traits via traditional breeding selection is hampered by the difficulty of effectively controlling GxE in the field. Diverting the selection to the use of molecular markers promises to overcome this issue, if adequate models can be defined. Therefore, in our study we deployed four RILs populations that represented well a typical durum wheat breeding program to test the feasibility of replacing phenotypic selection with molecular selection. The four populations showed transgressive segregation when phenotyped for GY and TKW, indicating additive effect loci are present from both parents as it would be expected from a well-designed breeding cross.

A Reliable Consensus Map

To construct a high-density consensus genetic map, a combination of four genetic backgrounds was used by anchoring common markers, followed by imputation of the missing haplotypes. The consensus map of IC, DRO, SW, YG included 14 linkage groups and spanned 2,705 cM, similar to what defined in the four way cross NCCR population map (2,664 cM) of Milner et al. (2016), and the six elite × elite populations durum wheat consensus map (2,631 cM) presented by Maccaferri et al. (2015) and in agreement with other reports ranging from 1,352 cM to 3,598 cM (Blanco et al., 1998; Nachit et al., 2001; Elouafi and Nachit, 2004; Mantovani et al., 2008; Peleg et al., 2008; Patil et al., 2013). The consensus map length was higher by 34% of the average length of the four individual maps. In agreement with

previous studies (Nachit et al., 2001; Elouafi and Nachit, 2004; Peleg et al., 2008; Patil et al., 2013) and contrary to Maccaferri et al. (2015), the A and B genomes had different map lengths, with the B genome (1,572 cM) being longer than A genome (1,133.8 cM). However, similarly to Maccaferri et al. (2015), a smaller number of markers was mapped to the A genome (1,104) compared to the B genome (2,098). The marker density in the consensus map differed along the chromosomes. According to previous studies (Erayman et al., 2004; Saintenac et al., 2011; Maccaferri et al., 2015), this is probably due to the variation of recombination frequency and the potential to accumulate genetic diversity. Markers gaps of 10–33 cM were identified in all chromosomes, except chromosome 4A. Chromosome regions with reduced marker density in 1A, 2A, 3A, and 7A have also been reported in the consensus map of Maccaferri et al. (2014). Overall, the consensus map developed was well in line with previous reported examples and it was hence deemed adequate to perform the targeted study.

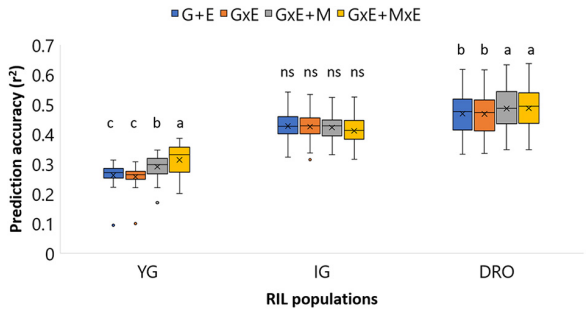


FIGURE 3 | Prediction accuracy for grain yield (GY) in YG, IG, and DRO populations using four different statistical models. G+E, genotype + environment effect; GxE, genotype by environment interaction; GxE + M, genotype by environment interaction + markers effect; GxE + MxE, genotype by environment interaction + markers by environment interaction. The horizontal line represents the average, the square indicates the 2nd and 3rd quartiles, the whiskers represent the 1st and 4th quartiles, the cross the median, and the dots are outliers. The letters indicated classes determined via LSD.

Identification of Major Effect Alleles by QTL Analysis

A total of 31 QTLs were identified for DTH, DTM, PLH, TKW, SPK, and GY, with most of them showing co-localization or pleiotropic effect. Consistent QTLs for GY were detected on chromosomes 2B (Qicd.TKW.DTH.GY.001, Qicd.GY.001, Qicd.GY.002, and Qicd.TKW.GY.001), 4A (Qicd.PLH.GY.001 and Qicd.PLH.GY.002), 4B (Qicd.PLH.GY.003), 5B (Qicd.GY.003 and Qicd.PLH.GY.004), 7A (Qicd.GY.004) and 7B (Qicd.GY.005). Chromosome 2B carries 10 individual QTLs, eight of which were found associated with GY, TKW, and SPK, explaining up to 33.4% of the phenotypic variance. This is in agreement with previous reports on QTLs identified on chromosome 2B associated with GY and its components (Huang et al., 2003; McCartney et al., 2005; Quarrie et al., 2005; Suenaga et al., 2005; Huang et al., 2006; Marza et al., 2006; Maccaferri et al., 2008; Golabadi et al., 2011). Six individual QTLs for TKW were found on chromosomes 1B, 4B, 6A, 6B, and 7A. Except for Qicd.TKW.006 on 7A, which we deem to have been reported here for the first time, the five remaining QTLs have been reported in previous studies by Blanco et al. (2011) and Patil et al. (2013). As indicated by Soriano et al. (2017), QTL influencing SPK were located on chromosomes 2B, 3B, and 5B. Assanga et al. (2017) have also found in winter wheat regions in 1A and 6B that are associated with the same trait.

Major genes associated with phenology were found to have a pleiotropic influence on trait measurement and QTL detection (Acuña-Galindo et al., 2015). Flowering time is a major trait in plant breeding and it provides the basis for plant adaptation. Chromosomes 2A, 2B, 4B, 5B, 6B, and 7B harbored QTLs linked to phenology traits. On 2A and 2B, two clusters of QTLs (Qicd.DTM.PL H.TKW.DTH.001 and Qicd.TKW.DTH.GY.001) were found in approximately the same position corresponding with Ppd-A1 and Ppd-B1 genes defined by several authors (Laurie, 1997; Maccaferri et al., 2008; Wilhelm et al., 2009; Maccaferri et al., 2011; Arjona et al., 2018). In our study, GY was associated to PLH in four QTLs located on chromosomes 4A, 4B, and 5B. Previous studies have also found that PLH genes are strongly associated with QTL for GY and its components (Quarrie et al., 2005; Crossa et al., 2007; Rebetzke et al., 2008; Acuña-Galindo et al., 2015). Borner et al. (2002), Huang et al. (2003, 2006), Blanco et al. (2012), and Patil et al. (2013) found that the short arm of chromosome 2A and its homologous harbor QTL influencing TKW, that was the case for clusters Qicd.DTM.PLH.TKW.DTH.001 and Qicd.TKW.DTH.GY.001. The cluster Qicd.DTM.PLH.TKW.DTH.001 for DTM, DTH, PLH (Soriano et al., 2017) and TKW on chromosome 2A confirms its agronomically important traits contribution as reported in Maccaferri et al. (2011) and Patil et al. (2013). On the homologous region on 2B, the cluster Qicd.TKW.DTH.GY.001 influences DTH, TKW and GY. On chromosome 5B cluster Qicd.DTH.PLH.001 could be related to Vrn-B1 as reported by Hanocq et al. (2004). On the long arm of chromosomes 2B, 4B, 6B, and 7B, the identified QTLs suggest important new regions controlling earliness. Soriano et al. (2017) have also identified a novel QTL on chr. 4B and 7B. In summary, the QTL analysis of these four populations has identified and validated several previously known loci and supports their use for molecular selection.

Selection of the Best Fitting Statistical Models for Genomic Predictions (*i, ii, iii, iv*)

The prediction analysis was conducted on the RILs population using models that account for the relationship matrix (G), environment effect (E), genotype by environment interaction (GxE), markers (M), and marker by environment interaction (MxE). The accuracy of breeding selection using only phenotypic data was computed (**Figure 3**) as G+E and GxE models (*i* and *ii*), to confirm that accuracies of 0.47-0.28 could be obtained via traditional breeding selection for GY. These results confirm what was reported by Crossa et al. (2014): that pedigree (population

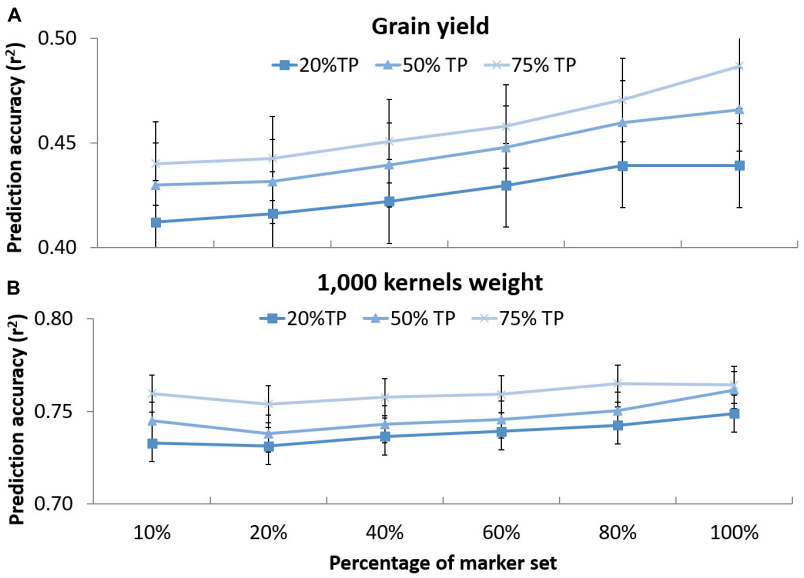


FIGURE 4 | Prediction accuracy for grain yield **(A)** and 1,000 kernel weight **(B)** using different randomly selected sub-sets of markers in decreasing order: 320 (10%), 640 (20%), 1,281 (40%), 921 (60%), 2,562 (80%), and 3,202 (100%) tested on DRO population using 20, 50, and 75% of the whole population as training set (TP) to predict the rest of the population (VP). The whiskers represent the standard errors.

TABLE 5 | Comparison of the prediction accuracies using full sibs and half sibs as training populations for grain yield and 1,000 kernel weight.

	DRO	IG	YG	SW	DRO	IG	YG	SW
	Grain yield				1,000-kernels weight			
DRO	0.47	−0.08	−0.11	0.07	0.76	−0.1	0.03	−0.26
IG	−0.09	0.41	0	0.08	−0.08	0.92	−0.02	0.09
YG	−0.07	−0.02	0.35	−0.08	0.12	0	0.83	0.14
SW	0.06	0.29	−0.13	0.37	−0.26	0.22	0.11	0.3

The columns represent the TP and the rows are the BP, the diagonal represents the full sibs relationships.

structure) accounts for a sizeable proportion of the prediction accuracy. These values were set as competitors to determine the success of replacing phenotypic selection with molecular selection. Interestingly, the GS models that incorporated marker effect (*iii, iv*) generated non-significantly different or superior accuracies than traditional breeding selection, indicating a strong role for GS in future breeding (**Figure 3**).

Size and Relatedness of the Training Population (*v, vi*)

Beside academical studies, breeders often have limited resources and tend to reduce costs whenever possible. A decrease in the size of the TP that needs to be both genotyped and phenotyped, and in the number of markers to be used for genotyping can represent important savings (Heffner et al., 2011; Crossa et al., 2014; Bassi et al., 2016). This possibility was tested by varying the proportion of individuals included in TP and VP from 75% TP and 25% VP, which is a very conservative and costly approach, to 50% TP and 50% VP, and even 25% TP and 75% VP. Interestingly, non-significant differences in accuracies could be observed for any of the reductions, for both high and low heritability traits (GY and TKW).

The relatedness between the TP and VP has been identified as a key consideration for predicting complex trait with low heritability. In an ideal scenario, breeders would like to accumulate information for a TP over time, using their normal yield trials as the source for this activity. By logic, the relatedness between such a TP and a BP under selection should be that of half-sibs. To test the feasibility of this approach, the four RIL populations that share half sib relationships were used to predict each other (**Table 5**). This resulted in severe losses of accuracy, reaching values close to zero for both high and low heritable traits (GY and TKW). This is in agreement with Windhausen et al. (2012), who also encountered accuracies close to zero when predicting far-related populations. The relatedness of a TP to the population to be predicted is hence one of the most critical aspect of GS in durum wheat. Therefore, small TP can be effectively deployed to accurately select BP only if these have full sibs relationships with the population to be selected. This is in good agreement with Bassi et al. (2016), who described several breeding schemes to deploy GS in a manner that would allow the TP to

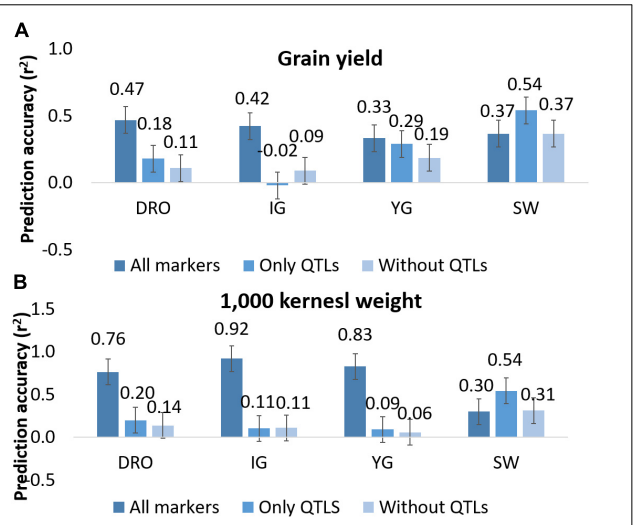


FIGURE 5 | Prediction accuracy for grain yield **(A)** and 1,000 kernel weight **(B)** using all markers, only markers linked to QTLs, and all markers except those identified as linked to QTLs. Whiskers represents the experiment wise LSD.

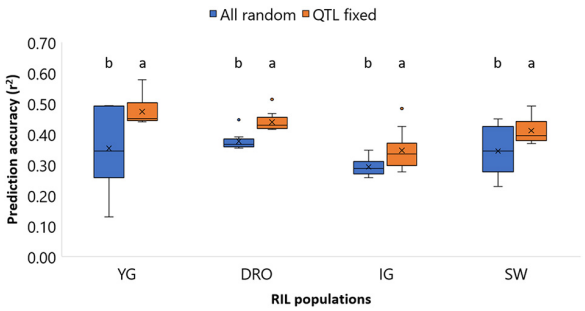


FIGURE 6 | Comparison of the prediction accuracies of grain yield (GY) for the four population YG, DRO, IG, and SW, using a model with all markers considered as random effect against a models that fixed markers underlying QTLs. The horizontal line represents the average, the square indicates the 2nd and 3rd quartiles, the whiskers represent the 1st and 4th quartiles, the cross the median, and the dots are outliers. The letters indicated classes determined via LSD.

be full-sib of the BP under selection, without excessive loss of genetic gain.

Does Markers Number Affect the Predictions? (*v*)

The possibility of deploying GS in breeding is still heavily hindered by the cost associated with genotyping huge populations. A way to reduce the cost of genotyping would be to reduce the number of markers used for the analysis. Here we tested the effect of the markers number to reveal that there was no significant difference in the prediction accuracies between using 3,202 or 320 SNPs as far as the TP and VP are full sibs (**Figure 4**). Hickey et al. (2014) also reported that when using information from related maize bi-parental populations high accuracies can be achieved using a small number of markers. Similarly, Haile et al. (2018) indicated that among advanced durum wheat breeding lines, the reduction from 9,000 to 500 markers did not cause a significant reduction in accuracies. However, it has to be noted that combining a decrease of TP size to 20% of the BP, and 10% of markers number caused the accuracy for GY to drop from 0.48 to 0.41 and for TKW from 0.77 to 0.74. This is a significant reduction of 0.07 and 0.03 points. Still, in the optic of practical application, the values of accuracies remain very close to what achieved using only phenotypic models (G+E and GxE) and hence it could be advisable to deploy small TP and small markers set in breeding if this makes GS a more affordable approach.

Is There an Advantage to Conduct QTL Analysis Before Genomic Predictions? (*vii, viii*)

QTL analysis and GS models rely on the same type of dataset. Therefore, it is of interest to define if there is additive contribution in combining both type of studies. Initially it was tested the effect of using only markers underlying QTLs to make prediction, as a way to simulate a MAS approach (**Figure 5**). The obtained accuracies reached between −0.02 and 0.54, depending on traits and populations. This would suggest that running prediction models using only few markers linked to known genes (44 and 27 for GY and TKW, respectively) could provide some degree of success. For confirmation, the opposite situation was also tested by removing any markers associated to QTL from the whole dataset. Once again, the accuracies dropped significantly for all traits and populations, except for SW. This result suggests that the marker number is not the only factor to ensure high accuracies, but that the ability to define the haplotype of major effect loci is also of critical importance.

The final test was designed to combine the extra information obtained via the definition of major allele effects by QTL analysis with the minor allele effects assessed via GS. Since the initial QTL discovery was conducted using the whole population, while GS models would instead use only sub-set of each population as TP and VP, QTL discovery was re-conducted for each TP subset. All initially identified QTLs were re-identified in 10–50% of the TP subsets (**Supplementary Table S4**) depending on the levels of allelic and phenotypic variation of each random subset. The marker underlying the re-identified QTLs were fixed for each TP subset and used to improve the prediction model. The results are extremely promising, since for all populations the combination of minor allele effects as GS random factor and major allele effects as QTL fixed factor resulted in a significant increase in prediction accuracies. Furthermore, the accuracies value were increased by 0.06–0.12 points, a major increase compared to the 0.02 points of reducing the TP size or changing statistical models. Our results are in partial agreement with Sarinelli et al. (2019) who demonstrated that major genes added as fixed effects always improved model predictive ability, with the greatest gains coming from combinations of multiple genes for days to heading and plant height in a winter wheat panel. Bian and Holland (2017)

also concluded that adding SNPs associated with a given trait as fixed effects resulted in higher predictive abilities when compared to models that only treated SNPs as random effects. Bernardo (2014) pointed out that the prediction accuracy of GS models can be increased by adding major genes as fixed effects when they represent a large proportion of the total variance associated with the trait under consideration (≥10%). Considering that GY remains often the main targeted trait, and also one of the most complex to predict, overall our results support the principle of incorporating fixed effect alleles into a prediction model, especially for markers accounting for a large part of the phenotypic variation. The idea of combining MAS using marker associated to known loci as fixed effects, and all other loci as random effect, becomes interesting for practical breeding applications. Furthermore, there appears to be an additive value in conducting a discovery step via QTL analysis before running genomic predictions, since the additional information can be strategically exploited to increase accuracies.

CONCLUSION

The results of this study provide a framework for better understanding and deploying molecular selection in durum wheat. The use of four populations to define a consensus linkage map allowed the precise identification of significant QTL for agronomic traits. Furthermore, these were incorporated into prediction models to reveal significant gains of accuracy for GY when integrated as fixed effects. Several critical considerations were also tested for their deployment in durum wheat breeding. The results presented here are in good agreement with previous literature and what suggested previously by us for breeding application of GS in wheat (Bassi et al., 2016). In practice, the use of half sibs or distantly related TP does not appear to be an exploitable methodology for GS in durum wheat. Instead, small size full sibs TP needs to be deployed and genotyping costs can be reduced by using just 200–300 SNPs. In addition, known loci linked to traits of interest should be also included in the marker set and used as fixed effects to increase prediction. Most importantly, all genomic prediction models were compared to the accuracy attainable by classical phenotypic selection to confirm that the same results could be achieved via molecular approaches. Altogether, our result provides strong support for the deployment of genomic prediction in durum wheat breeding.

DATA AVAILABILITY STATEMENT

The germplasm described here is available through ICARDA's genebank and can be requested here: https://www.genesys-pgr.org/wiews/SYR002. The genotypic and phenotypic data have been provided as **Supplementary Data Sheet 1**.

AUTHOR CONTRIBUTIONS

MZ, HK, FB, ZK, and GG analyzed the data. AF-M, BB, and MN provided insightful revision and discussions. MZ, HK, FB, and MN produced the data. MZ, HK, and FB wrote the manuscript. All authors reviewed the manuscript.

FUNDING

This study was funded by the Australian Grains Research and Development Corporation (GRDC) project ICA00012: Focused improvement of ICARDA/Australia durum germplasm for abiotic tolerance, while the salary of HK and FB was funded by the Swedish Research Council (Vetenskapsradet) U-Forsk2018 project 2017-05522, "Genomic prediction to deliver heat tolerant wheat to the Senegal River basin: phase II." The genotyping work was covered by the International Treaty on Plant Genetic Resources for Food and Agriculture 2014-2015-2B-PR-02-Jordan: "An Integrated Approach to Identify and Characterize Climate Resilient Wheat for the West Asia and North Africa." Funds for imputation by GG were obtained from BBSRC to The Roslin Institute (BBS/E/D/30002275).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00316/full#supplementary-material

FIGURE S1 | Cartographic location of research stations used for this study. Source: modified from Google Map.

TABLE S1 | Description of the four individual maps issued from IC, DRO, SW and YG populations.

TABLE S2 | Significant QTLs with LOD and phenotypic variance (PV) for the studied traits across environments of the four populations.

TABLE S3 | Significant QTLs with LOD and phenotypic variance (PV) for GY, TKW, SPK and PLH of the four populations.

TABLE S4 | Frequency of re-identifying QTL associated to grain yield in the ten subset of training population.

DATA SHEET S1 | Complete dataset, including phenotyping information for each mapping population presented as BLUEs per location, and full genotyping file for the consensus map.

REFERENCES

Acuña-Galindo, M. A., Mason, R. E., Subramanian, N. K., and Hays, D. B. (2015). Meta-analysis of wheat QTL regions associated with adaptation to drought and heat stress. *Crop Sci.* 55, 477–492. doi: 10.2135/cropsci2013.11.0793

Antolin, R., Nettelblad, C., Gorjanc, G., Money, D., and Hickey, J. M. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genet. Sel. Evol.* 49:30. doi: 10.1186/s12711-017-0300-y

Arjona, J. M., Royo, C., Dreisigacker, S., Ammar, K., and Villegas, D. (2018). Effect of Ppd-A1 and Ppd-B1 allelic variants on grain number and thousand kernel

weight of durum wheat and their impact on final grain yield. *Front. Plant Sci.* 9:888. doi: 10.3389/fpls.2018.00888

Asif, M. A., Schilling, R. K., and Tilbrook, J., Brien, C., Dowling, K., Rabie, H., et al., (2018). Mapping of novel salt tolerance QTL in an Excalibur × Kukri doubled haploid wheat population. *Theor. Appl. Genet.* 131:2179. doi: 10.1007/s00122-018-3146-y

Assanga, S. O., Fuentealba, M., Zhang, G., Tan, C., Dhakal, S., Rudd, J. C., et al. (2017). Mapping of quantitative trait loci for grain yield and its components in a US popular winter wheat TAM 111 using 90K SNPs. *PLoS One* 12:e0189669. doi: 10.1371/journal.pone.0189669

Bassi, F. M., Bentley, A. R., Charmet, G., and Ortiz, R. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* 242, 23–36. doi: 10.1016/j.plantsci.2015.08.021

Bassi, F. M., Brahmi, H., Sabraoui, A., Amri, A., Nsarellah, N., Nachit, M. M., et al. (2019). Genetic identification of loci for Hessian fly resistance in durum wheat. *Mol. Breed.* 39:24. doi: 10.1007/s11032-019-0927-1

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bernardo, R. (2014). Genome wide selection when major genes are known. *Crop Sci.* 54, 68–75. doi: 10.2135/cropsci2013.05.0315

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. S. (2018). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic Hexaploid wheat. *Int. J. Mol Sci.* 19:3011. doi: 10.3390/ijms19103011

Bian, Y., and Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity* 118, 585–593. doi: 10.1038/hdy.2017.4

Blanco, A., Bellomo, M. P., Cenci, A., De Giovanni, G., D'Ovidio, R., Lacono, E., et al. (1998). A genetic linkage map of durum wheat. *Theor. Appl. Genet.* 97, 721–728.

Blanco, A., Colasuonno, P., Gadaleta, A., Mangini, G., Schiavulli, A., Simeone, R., et al. (2011). Quantitative trait loci for yellow pigment concentration and individual carotenoid compounds in durum wheat. *Cereal Sci.* 54, 255–264. doi: 10.1016/j.jcs.2011.07.002

Blanco, A., Mangini, G., Giancaspro, A., Giove, S., Colasuonno, P., Simeone, R., et al. (2012). Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars. *Mol. Breeding* 30, 79–92. doi: 10.1007/s11032-011-9600-z

Borner, A., Schumann, E., Furste, A., Coster, H., Leithold, B., Roder, M. S., et al. (2002). Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (*Triticumaestivum* L). *Theor. Appl. Genet.* 105, 921–936. doi: 10.1007/s00122-002-0994-1

Broman, K. W., Wu, H., Sen, S., and Churchill, A. G. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Crosbie, T. M., Eathington, S. R., Johnson, G. R., Edwards, M., Reiter, R., Stark, S., et al. (2003). "Plant breeding: past, present, and future," in *Plant Breeding: The Arnel R. Hallauer International Symposium*, eds K. R. Lamkey and M. Lee (Oxford: Blackwell), 1–50.

Crossa, J., Burgueno, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S. A., Lillemo, M., et al. (2007). Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177, 1889–1913. doi: 10.1534/genetics.107.078659

Crossa, J., de los Campos, G., Perez, P., Gianola, D., Burgueno, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Peirez, P., Hickey, J., and BurguenTo, J. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

De Givry, S., Bouchez, M., Chabrier, P., Milan, D., and Schiex, T. (2005). CARHTAGENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* 21, 1703–1704. doi: 10.1093/bioinformatics/bti222

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi: 10.1534/genetics.109.101501

Dekkers, J. C. M., and Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* 3, 22–32. doi: 10.1038/nrg701

Edae, E. A., Byrne, P. F., Haley, S. D., Lopes, M. S., and Reynolds, M. P. (2014). Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. *Theor. Appl. Genet.* 127, 791–807. doi: 10.1007/s00122-013-2257-8

Edae, E. A., Olivera, P. D., Jin, Y., and Rouse, M. N. (2017). Genotyping-by-sequencing facilitates a high-density consensus linkage map for *Aegilopsumbellulata*, a wild relative of cultivated wheat. *G3* 7, 1551–1561. doi: 10.1534/g3.117.039966

Elouafi, I., and Nachit, M. M. (2004). A genetic linkage map of the Durum Triticum dicoccoides backcross population based on SSRs and AFLP markers, and QTL analysis for milling traits. *Theor. Appl. Genet.* 108, 401–413. doi: 10.1007/s00122-003-1440-8

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Erayman, M., Sandhu, D., Sidhu, D., Dilbirligi, M., and Gill, K. S. (2004). Demarcating the gene-rich regions of the wheat genome. *Nucleic Acids Res.* 32, 3546–3565. doi: 10.1093/nar/gkh639

Golabadi, M., Arzani, A., Maibody, S. M., Tabatabaei, B. S., and Mohammadi, S. (2011). Identification of microsatellite markers linked with yield components under drought stress at terminal growth stages in durum wheat. *Euphytica* 177, 207–221. doi: 10.1007/s10681-010-0242-8

Habash, D. Z., Kehel, Z., and Nachit, M. (2009). Genomic approaches for designing durum wheat ready for climate change with a focus on drought. *Exp. Bot.* 60, 2805–2815. doi: 10.1093/jxb/erp211

Haile, J. K., N'Diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., et al. (2018). Genomic selection for grain yield and quality trait in durum wheat. *Mol. Breed.* 38:75. doi: 10.1007/s11032-018-0818-x

Hanocq, E., Niarquin, M., Heumez, E., Rousset, M., and Le Gouis, J. (2004). Detection and mapping of QTL for earliness components in a bread wheat recombinant inbred lines population. *Theor. Appl. Genet.* 110, 106–115. doi: 10.1007/s00122-004-1799-1

Heffner, E., Jannink, J., Iwata, H., Souza, E., and Sorrells, M. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253

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

Hickey, J. M., Kinghorn, B. P., Tier, B., van der Werf, J. H., and Cleveland, M. A. (2012). A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet. Sel. Evol.* 44:9. doi: 10.1186/1297-9686-44-9

Huang, X. Q., Cloutier, S., Lycar, L., Radovanovic, N., Humphreys, D. G., Noll, J. S., et al. (2006). Molecular detection of QTL for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (*Triticum aestivum* L.). *Theor. Appl. Genet.* 113, 753–766. doi: 10.1007/s00122-006-0346-7

Huang, X. Q., Coster, H., Ganal, M. W., and Roder, M. S. (2003). Advanced backcross QTL analysis for the identification of quantitative trait loci alleles from wild relatives of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 106, 1379–1389. doi: 10.1007/s00122-002-1179-7

Hussain, W., Baenziger, P. S., Belamkar, V., Guttieri, M. J., Venegas, J. P., Easterly, A., et al. (2017). Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. *Sci. Rep.* 7:16394. doi: 10.1038/s41598-017-16006-z

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Kuzmanoviæ, L., Ruggeri, R., Able, J. A., Bassi, M. F., Maccaferri, M., Tuberosa, R., et al. (2018). Yield performance of chromosomally engineered durum wheat-Thinopyrum ponticum recombinant lines in a range of contrasting rain-fed environments across three countries. *bioRvix [Preprint]* doi: 10.1101/313825

Laurie, D. A. (1997). Comparative genetics of flowering time. *Plant Mol. Biol.* 35, 167–177. doi: 10.1023/A:1005726329248

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., et al. (2015). Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *G3 (Bethesda)* 5, 569–582. doi: 10.1534/g3.114.016097

Maccaferri, M., Cane, M. A., Colalongo, C., Massi, A., Clarke, F., Pozniak, C., et al. (2014). A consensus framework map of durum wheat (*Triticum durum* Desf.) suitable for linkage disequilibrium and genome-wide association mapping. *BMC Genome* 15:873. doi: 10.1186/1471-2164-15-873

Maccaferri, M., Mantovani, P., Tuberosa, R., Deambrogio, E., Giuliani, S., Demontis, A., et al. (2008). A major QTL for durable leaf rust resistance widely exploited in durum wheat breeding programs maps on the distal region of chromosome arm 7BL. *Theor. Appl. Genet.* 117, 1225–1240. doi: 10.1007/s00122-008-0857-5

Maccaferri, M., Ricci, A., Salvi, S., Milner, S. G., Noli, E., Martelli, P. L., et al. (2015). A high-density, SNP-based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding. *Plant Biotechnol.* 13, 648–663. doi: 10.1111/pbi.12288

Maccaferri, M., Sanguineti, M. C., Demontis, A., El-Ahmed, A., del Moral, G. L., Maalouf, F., et al. (2011). Association mapping in durum wheat grown across a broad range of water regimes. *Exp. Bot.* 62, 409–438. doi: 10.1093/jxb/erq287

Mantovani, P., Maccaferri, M., Sanguineti, M. C., Tuberosa, R., Catizone, I., Wenzl, P., et al. (2008). An integrated DArT-SSR linkage map of durum wheat. *Mol Breed.* 22, 629–648. doi: 10.1007/s11032-008-9205-3

Marone, D., Laidò, G., Gadaleta, A., Colasuonno, P., Ficco, D. B., Giancaspro, A., et al. (2012). A high-density consensus map of A and B wheat genomes. *Theor. Appl. Genet.* 125, 1619–1638. doi: 10.1007/s00122-012-1939-y

Marza, F., Bai, G. H., Carver, B. F., and Zhou, W. C. (2006). Quantitative trait loci for yield and related traits in the wheat population Ning7840 9 Clark. *Theor. Appl. Genet.* 112, 688–698. doi: 10.1007/s00122-005-0172-3

McCartney, C. A., Somers, D. J., Humphreys, D. G., Lukow, O., Ames, N., Noll, J., et al. (2005). Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross RL4452 × 'AC Domain'. *Genome* 48, 870–883. doi: 10.1139/g05-055

Meng, L., Li, H. H., Zhang, L. Y., and Wang, J. K. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

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

Milner, S. G., Maccaferri, M., Huang, B. E., Mantovani, P., Massi, A., Frascaroli, E., et al. (2016). A multiparental cross population for mapping QTL for agronomic traits in durum wheat (*Triticum turgidum* ssp. *durum*). *Plant Biotechnol.* 14, 735–748. doi: 10.1111/pbi.12424

Mwadzingeni, L., Shimelis, H., Jasper, D., Rees, G., and Tsilo, T. J. (2017). Genome-wide association analysis of agronomic traits in wheat under drought-stressed and non-stressed conditions. *PLoS One* 12:e0171692. doi: 10.1371/journal.pone.0171692

Nachit, M. M., Elouafi, I., Pagnotta, M. A., El Saleh, A., Lacono, E., Labhilili, M., et al. (2001). Molecular linkage map for an intraspecific recombinant inbred population of durum wheat (*Triticum turgidum* L. var. durum). *Theor. Appl. Genet.* 102, 177–186. doi: 10.1007/s001220051633

Patil, R. M., Tamhankar, S. A., Oak, M. D., Raut, A. L., Honrao, B. K., Rao, V. S., et al. (2013). Mapping of QTL for agronomic traits and kernel characters in durum wheat (*Triticum durum* Desf.). *Euphytica* 190, 117–129. doi: 10.1007/s10681-012-0785-y

Peleg, Z., Saranga, Y., Suprunova, T., Ronin, Y., Roder, M. S., Kilian, A., et al. (2008). High-density genetic map of durum wheat 9 wild emmer wheat based on SSR and DArT markers. *Theor. Appl. Genet.* 117, 103–115. doi: 10.1007/s00122-008-0756-9

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

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7:e32253. doi: 10.1371/journal.pone.0032253

Quarrie, S. A., Steed, A., Calestani, C., Semikhodskii, A., Lebreton, C., Chinoy, C., et al. (2005). A high-density genetic map of hexaploid wheat (*Triticum aestivum* L.) from the cross Chinese Spring × SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theor. Appl. Genet.* 110, 865–880. doi: 10.1007/s00122-004-1902-7

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rebetzke, G., Condon, A., Farquhar, G., Appels, R., and Richards, R. (2008). Quantitative trait loci for carbon isotope discrimination are repeatable across environments and wheat mapping populations. *Theor. Appl. Genet.* 118, 123–137. doi: 10.1007/s00122-008-0882-4

Roselló, M., Royo, C., Sanchez-Garcia, M., and Soriano, J. M. (2019). Genetic dissection of the seminal root system architecture in mediterranean durum wheat landraces by genome-wide association study. *Agronomy* 9:364. doi: 10.3390/agronomy9070364

Saintenac, C., Jiang, D., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12:R88. doi: 10.1186/gb-2011-12-9-r88

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., et al. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. *Theor. Appl. Genet.* 132, 1247–1261. doi: 10.1007/s00122-019-03276-6

Somers, J. D., Isaac, P., and Edwards, K. (2004). A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 109, 1105–1114. doi: 10.1007/s00122-004-1740-7

Soriano, J. M., Malosetti, M., Rosello, A. M., Sorrells, M. E., and Royo, C. (2017). Dissecting the old *Mediterranean durum* wheat genetic architecture for phenology, biomass and yield formation by association mapping and QTL meta-analysis. *PLoS One* 12:e0178290. doi: 10.1371/journal.pone.0178290

Suenaga, K., Khairallah, M., William, H. M., and Hoisington, D. A. (2005). A new intervarietal linkage map and its application for quantitative trait locus analysis of "gigas" features in bread wheat. *Genome* 48, 65–75. doi: 10.1139/g04-092

Sukumaran, S., Li, X., Zhu, C., Bai, G., Perumal, R., Tuinstra, M. R., et al. (2016). QTL mapping for grain yield, flowering time, and stay-green traits in sorghum with genotyping-by-sequencing markers. *Crop Sci.* 56, 1429–1442. doi: 10.2135/cropsci2015.02.0097

Tadesse, W., Nachit, M., Abdalla, O., and Rajaram, S. (2016). "Wheat breeding at ICARDA: achievements and prospects in the CWANA region," in *The World Wheat Book Volume 3. A History of Wheat Breeding*, eds A. Bonjean, B. Angus, and M. van Ginkel (Paris: Lavoiseier).

The International Wheat Genome Sequencing Consortium [IWGSC] (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Wang, S. C., Wong, D. B., Forrest, K., Allen, A., Chao, S. M., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol.* 12, 787–796. doi: 10.1111/pbi.12183

Wilhelm, E. P., Turner, A. S., and Laurie, D. A. (2009). Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (*Triticum durum* Desf.). *Theor Appl Genet.* 118, 285–294. doi: 10.1007/s00122-008-0898-9

Windhausen, V. S., Atlin, G. N., Crossa, J., Hickey, J. M., Grudloyma, P., Terekegne, A., et al. (2012). Effectiveness of genomic prediction of maize

hybrid performance in different breeding populations and environments. *Genes Genomes Genet.* 2, 1427–1436. doi: 10.1534/g3.112.003699

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L., Benbow, H. R., and Wilkinson, P. A. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol.* 14, 1195–1206. doi: 10.1111/pbi.12485

Xiao, D., Bai, H., and Liu, D. L. (2018). Impact of future climate change on wheat production: a simulated case for China's wheat system. *Sustainability* 10:1277. doi: 10.3390/su10041277

Yu, Y., Ouyang, Y., and Yao, W. (2018). ShinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* 34, 1229–1231. doi: 10.1093/bioinformatics/btx763

Zhang, J., Long, Y., Wang, L., Dang, Z., Zhang, T., Song, X., et al. (2018). Consensus genetic linkage map construction and QTL mapping for plant height-related traits in linseed flax (*Linumusitatissimum* L.). *BMC Plant Biol.* 18:160. doi: 10.1186/s12870-018-1366-6

Zhang, L., Liu, D., Guo, X., Yang, W., Sun, J., Wang, D., et al. (2010). Genomic distribution of quantitative trait loci for yield and yield-related traits in common wheat. *Integr. Plant Biol.* 52, 996–1007. doi: 10.1111/j.1744-7909.2010.00967.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zaïm, Kabbaj, Kehel, Gorjanc, Filali-Maltouf, Belkadi, Nachit and Bassi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 05 June 2020
doi: 10.3389/fgene.2020.00447

Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies

Stefano Pavan[1,2], Chiara Delvento[1], Luigi Ricciardi[1], Concetta Lotti[3], Elena Ciani[4] and Nunzio D'Agostino[5]**

[1] Department of Soil, Plant and Food Science, Section of Genetics and Plant Breeding, University of Bari Aldo Moro, Bari, Italy, [2] Institute of Biomedical Technologies, National Research Council (CNR), Bari, Italy, [3] Department of Agricultural, Food and Environmental Sciences, University of Foggia, Foggia, Italy, [4] Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari Aldo Moro, Bari, Italy, [5] Department of Agricultural Sciences, University of Naples Federico II, Naples, Italy

OPEN ACCESS

Edited by:
Hans D. Daetwyler, La Trobe University, Australia

Reviewed by:
Christian Werner, The University of Edinburgh, United Kingdom
Kai P. Voss-Fels, The University of Queensland, Australia

***Correspondence:**
Stefano Pavan
stefano.pavan@uniba.it
Nunzio D'Agostino
nunzio.dagostino@unina.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *19 October 2019*
Accepted: *14 April 2020*
Published: *05 June 2020*

Citation:
Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E and D'Agostino N (2020) Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. Front. Genet. 11:447. doi: 10.3389/fgene.2020.00447

High-throughput genotyping boosts genome-wide association studies (GWAS) in crop species, leading to the identification of single-nucleotide polymorphisms (SNPs) associated with economically important traits. Choosing a cost-effective genotyping method for crop GWAS requires careful examination of several aspects, namely, the purpose and the scale of the study, crop-specific genomic features, and technical and economic matters associated with each genotyping option. Once genotypic data have been obtained, quality control (QC) procedures must be applied to avoid bias and false signals in genotype–phenotype association tests. QC for human GWAS has been extensively reviewed; however, QC for crop GWAS may require different actions, depending on the GWAS population type. Here, we review most popular genotyping methods based on next-generation sequencing (NGS) and array hybridization, and report observations that should guide the investigator in the choice of the genotyping method for crop GWAS. We provide recommendations to perform QC in crop species, and deliver an overview of bioinformatics tools that can be used to accomplish all needed tasks. Overall, this work aims to provide guidelines to harmonize those procedures leading to SNP datasets ready for crop GWAS.

Keywords: crops, GWAS, genotyping, quality control, bioinformatics tools

INTRODUCTION

High-throughput genotyping, which leads to the identification of a large number of single-nucleotide polymorphisms (SNPs) is boosting the implementation of genome-wide association studies (GWAS), linking DNA variants to phenotypes of interest (Taranto et al., 2018). In crop species, GWAS enabled the mapping of genomic loci associated with economically important traits, including yield, resistance to biotic and abiotic stresses, and quality (Boyles et al., 2016; Pavan et al., 2017; Hou et al., 2018; Liu et al., 2018; He et al., 2019). This information has been further used to perform marker-assisted selection (MAS) in breeding programs and discover genes underlying phenotypic variation (Liu and Yan, 2019).

Several genotyping methods are available (reviewed by Scheben et al., 2017), which are usually performed by commercial parties upon the receipt of DNA samples. For application in GWAS, widely adopted genotyping options fall into three categories: whole genome resequencing (WGR), reduced representation sequencing (RRS), and SNP arrays. WGR and RRS are based on next-generation sequencing (NGS) technologies and bioinformatics pipelines that align reads to a reference genome and call both SNPs and genotypes (Nielsen et al., 2011). SNP arrays rely on allele-specific oligonucleotide (ASO) probes (including target SNP loci plus their flanking regions) fixed on a solid support, which are used to interrogate complementary fragments from DNA samples and infer genotypes based on the interpretation of the hybridization signal. Choosing the most appropriate (cost-effective) genotyping method for crop GWAS requires careful examination of several aspects, namely, the purpose and the scale of the study, crop-specific genomic features, and technical and economic matters associated with each genotyping method.

Raw SNP datasets resulting from genotyping experiments are typically inaccurate and incomplete. In addition, genes associated with phenotypes can have a small effect on genetic variance. In this scenario, quality control (QC) procedures are of pivotal importance to minimize false-positive or false-negative associations, referred to as type I and type II errors, respectively. QC includes filtering out poor-quality or suspected artifactual SNP loci, filtering out individuals in relation to missing data, anomalous genotype call and genetic synonymies, and the characterization of ancestral relationships among individuals of the GWAS population. Excellent reviews focused on QC of human SNP data (Turner et al., 2011; Marees et al., 2018); however, the QC procedure may be quite different for crop species. In this case, variables that need to be considered include the crop prevailing mating system (self- or open-pollinating) and the breeding history of the specific GWAS population.

This review aims to provide recommendations on how to plan genotyping experiments and best practices on how to perform QC in crop species.

CHOOSING THE CORRECT GENOTYPING METHOD

Genotyping methods differ with respect to the number of identifiable SNPs and the cost of the analysis *per* sample, and these two parameters are directly proportional. Given this premise, choosing the correct option for GWAS requires to have a clear idea on two key aspects, i.e., the number of SNPs that is sufficient/desirable to fulfill the GWAS goals and the cost associated with each genotyping alternative. In addition, genotyping methods come with different technical specifications that should be evaluated in relation to the particular GWAS experiment.

Whole Genome Resequencing

WGR allows the highest number of SNP calls, up to several millions as reported in peach (Cao et al., 2016) and cotton (Du et al., 2018). This is a clear advantage when, rather than MAS, gene isolation is the main aim of the GWAS project (Wang et al., 2016; Happ et al., 2019). Indeed, in high-resolution GWAS, SNP loci showing the highest evidence of association are usually in tight linkage, or may even coincide, with loci underlying phenotypic variation (e.g., Shang et al., 2014; Yano et al., 2016). However, it should be pointed out that, even with a very high marker density, the identification of causal polymorphisms can be difficult in the case of GWAS populations displaying slow decay of linkage disequilibrium (LD) (i.e., populations in which the allelic state at two loci on the same chromosome tends to be correlated even at high physical distance) (Korte and Farlow, 2013). As shown in **Table 1**, in populations of self-pollinating crops, such as wheat or soybean, the average square correlation coefficient (r^2) between pairs of loci may take several Mb to decay to values indicating substantial linkage equilibrium (0.2 or 0.1) (Vos et al., 2017).

WGR is especially desirable for GWAS populations displaying rapid LD decay. Indeed, in this case, low marker density may result in missing genomic regions associated with phenotypic traits. Extremely rapid LD decay (in the range of a few base pairs) has been reported for highly heterozygous populations of open-pollinating species (e.g., maize, carrot, olive), in which recombination is effective in breaking up haplotypes (**Table 1**). In this situation, even in the ideal case of equally spaced SNPs, millions of markers would be required to have a SNP distance lower than the LD decay distance. This is exactly the condition that enables one to detect associations for most genomic regions (**Table 1**).

WGR genotyping has been so far generally performed using paired-end Illumina technology (e.g., Zhou et al., 2015; Cao et al., 2016; Liang et al., 2019), which, according to our survey, roughly costs $400 per sample for a genome of 1 Gb and 10× average sequencing depth (this term indicating the number of times a base is sequenced on average). This implies that WGR-based GWAS, typically involving a few hundred individuals, may cost several hundred thousand dollars for crops with large genomes, as shown in **Table 1**. Decreasing the average sequencing depth can lower the cost of WGR; however, this may result in an unacceptable number of genotyping errors. This is especially the case of heterozygous loci, which are associated with a larger number of genotypic combinations (Kishikawa et al., 2019). In practice, WGR in crops has been usually performed with average sequencing depth ranging from ∼5×, as for cotton (Du et al., 2018), tomato (Lin et al., 2014), and peach (Cao et al., 2019), to ∼15×, as for watermelon (Guo et al., 2019) and grapevine (Liang et al., 2019). A notable exception is represented by strict self-pollinating species, such as rice and soybean, for which very low average sequencing depth (1× or lower) has been successfully applied (Wang et al., 2016; Happ et al., 2019). Indeed, homozygous populations of pure lines are effectively haploid, thus allowing easy reconstruction of haplotypes and, consequently, accurate imputation of missing data (Wang et al., 2016).

Reduced Representation Sequencing

RRS consists in sequencing only a small fraction of the genome, thus reducing the cost of the analysis with respect to WGR (Hirsch et al., 2014). Genotyping by sequencing

TABLE 1 | List of some genomic and economic aspects that should be taken into consideration when planning GWAS in crops.

Species	Genome size (Gb)	References	LD decay	References	Minimum number of SNPs for a distance < LD decay *	Estimated WGR cost on 100 individuals ($) **	SNP array			
							Name	Technology	Size	References
Brassicaceae										
Brassica napus	0.49	Chalhoub et al., 2014	800 Kb (r^2 = 0.2, A subgenome); 4.8 Mb (r^2 = 0.2, B subgenome)	Zhao et al., 2016	980 (subgenome A) 143 (subgenome B)	19.4 K	International Brassica SNP Consortium	Illumina Infinium BeadChip	52K	Clarke et al., 2016
Solanaceae										
Solanum lycopersicum	0.90	Sato et al., 2012	665 Kb (r^2 = 0.2)	Ruggieri et al., 2014	1353	36K	SolCAP Tomato 2013	Illumina Infinium BeadChip	9K	Sim et al., 2012
							Axiom Tomato Genotyping Array	Affymetrix Axiom	52K	Unpublished
Solanum tuberosum	0.84	Xu et al., 2011	1.5–0.6 Mb (r^2 = 0.1)	Vos et al., 2017	560–14,000	33.6K	SOLCAP Potato 2013	Illumina Infinium BeadChip	10K	Hamilton et al., 2011
							SolSTW array	Affymetrix Axiom	20K	Vos et al., 2015
Capsicum annuum	3.30	Kim et al., 2014; Qin et al., 2014	100 Kb (r^2 = 0.2)	Taranto et al., 2016	33,000	132K	UCD TraitGenetics Pepper (Capsicum) Consortium	Illumina Infinium BeadChip	19K	Ashrafi et al., 2012
							Pepper (Capsicum) SNP Genotyping Array	Affymetrix Axiom	640K	Unpublished
Cucurbitaceae										
Cucumis sativus	0.35	Huang et al., 2009	24 Kb (r^2 = 0.09)	Wang et al., 2018	14,583	14K	–	Fluidigm	35K	Rubinstein et al., 2015
			55–140.5 Kb (r^2 = 0.2)	Qi et al., 2013	6364–2491					
Cucumis melo	0.45	Garcia-Mas et al., 2012	100 Kb (r^2 = 0.2)	Gur et al., 2017	4500	18K				
			72–774 Kb (r^2 = 0.2)	Pavan et al., 2017	6250–581					
Fabaceae										
Phaseolus vulgaris	0.59	Schmutz et al., 2014	1 Mb (r^2 = 0.1)	Diniz et al., 2019	587	23.48K	BARCBean6K_1	Illumina Infinium BeadChip	5K	Song et al., 2015
Glycine max	1.12	Schmutz et al., 2010	8.5–15.5 Mb (r^2 = 0.1)	Liu Z. et al., 2017	131–72	44.6K	SoySNP50K	Illumina Infinium BeadChip	6K	Song et al., 2013
			5.9–7 Mb (r^2 = 0.1)	Mamidi et al., 2011	189–159		SoyaSNP180K Axiom	Affymetrix Axiom	180K	Lee et al., 2015
Apiaceae										
Daucus carota	0.47	Iorizzo et al., 2016	100–400 bp (r^2 = 0.2)	Ellison et al., 2018	4,730,000–1,182,500	18.92K				
Poaceae										
Oryza sativa	0.39	Sasaki, 2005	150 Kb (r^2 = 0.2)	Liu et al., 2020	2593	15.56K	RiceSNP50	Illumina Infinium BeadChip	50K	Chen et al., 2014
							RICE6K	Illumina Infinium BeadChip	6K	Yu et al., 2014

(Continued)

TABLE 1 | Continued

Species	Genome size (Gb)	References	LD decay	References	Minimum number of SNPs for a distance < LD decay *	Estimated WGR cost on 100 individuals ($) **	SNP array			
							Name	Technology	Size	References
							Axiom Rice Genotyping Array	Affymetrix Axiom	50K	Singh et al., 2015
Triticum aestivum	16.00	International Wheat Genome Sequencing and Consortium, 2014	8 Mb (r^2 = 0.08)	Liu J. et al., 2017	2000	640K	US/Australia 9K Wheat Consortium	Illumina Infinium BeadChip	9K	Cavanagh et al., 2013
							Wheat 90K iSelect	Illumina Infinium BeadChip	90K	Wang et al., 2014
							Axiom Wheat Breeders Genotyping Array	Affymetrix Axiom	35K	Allen et al., 2017
							Axiom Wheat HD Genotyping Arrays	Affymetrix Axiom	817K	Winfield et al., 2016
Zea mays	2.50	Schnable et al., 2009	6.34 Kb (r^2 = 0.2)	Dinesh et al., 2016	394,322	100K	MaizeSNP50 BeadChip	Illumina Infinium BeadChip	50K	Ganal et al., 2011
			500 bp (r^2 = 0.2)	Yan et al., 2009	5,000,000		Subset of MaizeSNP50	Illumina Infinium BeadChip	3K	Rousselle et al., 2015
			1.5 Kb (r^2 = 0.1)	Remington et al., 2001	1,666,667		Axiom Maize Genotyping Array	Affymetrix Axiom	600K	Unterseer et al., 2014
							Maize 55K Axiom	Affymetrix Axiom	55K	Xu et al., 2017
Rosaceae										
Malus domestica	0.74	Velasco et al., 2010	200 bp (r^2 = 0.2)	Larsen et al., 2019	7,420,000	29.68K	RosBREED Apple	Illumina Infinium BeadChip	8K	Chagné et al., 2012
							Fruitbreedomics Apple20k	Illumina Infinium BeadChip	20K	Bianco et al., 2014
							Axiom Apple Genotyping Array	Affymetrix Axiom	480K	Bianco et al., 2016
Prunus persica	0.27	Verde et al., 2013	1.2–3.2 Mb (r^2 = 0.1)	Li et al., 2013	221–83	10.6K	RosBREEDPeach	Illumina Infinium BeadChip	9K	Verde et al., 2012
Vitaceae										
Vitis vinifera	0.48	Jaillon et al., 2007	43 Kb (r^2 = 0.2)	Nicolas et al., 2016	11047	19K	GrapeReSeq Consortium	Illumina Infinium BeadChip	20K	Le Paslier et al., 2013
							GeneChip *Vitis vinifera* (Grape) Genome Array	Applied Biosystems	15K	Unpublished
Oleaceae										
Olea europaea	1.46	Unver et al., 2017	25 bp (r^2 = 0.05)	D'Agostino et al., 2018	58,400,000	58.4K				
Malvaceae										
Gossypium hirsutum	2.43	Li et al., 2015	3.2–3.3 Mb (r^2 = 0.1)	Yuan et al., 2018	759–736	97.2K	International Cotton SNP Consortium	Illumina Infinium BeadChip	70K	Hulse-Kemp et al., 2015
			900 Kb (r^2 = 0.1)	Wen et al., 2019	2700		Axiom Cotton Genotyping Array	Affymetrix Axiom	35K	Unpublished

For several main crop species belonging to different botanical families, the following information is reported: estimated haploid genome size; linkage disequilibrium (LD) decay; the minimum number of equally distributed SNPs providing a distance lower than the LD decay; estimated WGR cost on a panel of 100 individuals; the list of available SNP array(s).

(GBS) (Elshire et al., 2011), restriction site-associated DNA sequencing (RADseq) (Davey and Blaxter, 2011), and double digest RAD sequencing (ddRAD-seq) (Truong et al., 2012), which use restriction enzymes (REs) for the reduction of genome complexity, are currently the most popular RRS methods used to perform GWAS in crops, mainly due to their moderate cost. At a minimum, this is approximately $35 per sample independently from the genome size and including the application of bioinformatics pipelines for SNP and genotype calling (You et al., 2018). Another advantage of these RRS methods is their scalability, meaning that different combinations of restriction enzymes may be used to customize the percentage of the genome captured.

The number of SNPs identified by RRS genotyping typically varies from a few to several thousands (Pavan et al., 2018, 2019; Colonna et al., 2019), depending on the amount of genome sequenced and population diversity. As discussed above, this output can be largely sufficient in GWAS experiments whose main aim is to implement marker-assisted selection, and for crops displaying slow LD decay (**Table 1**).

A major technical limitation of RSS is that the genomic distribution of SNPs depends on the specific combination of REs used (D'Agostino and Tripodi, 2017). In addition, sequencing depth at individual SNP loci identified by RRS is typically uneven, leading to under-calling of heterozygous loci and many missing data. The latter issue can be mitigated by genotype imputation strategies; however, we highlight that the success of genotype imputation depends on the genetic makeup of the GWAS population, which influences, among other things, the occurrence of long homozygous segments useful to reconstruct haplotypes (Glaubitz et al., 2014).

SNP Arrays

SNP arrays for agrigenomics have been developed for over a decade to meet the needs for single research groups or consortia and are still widely used for GWAS in crops despite the decreasing cost of NGS-based technologies (LaFramboise, 2009; Rasheed et al., 2017; **Table 1**). In 2017, the two leader manufacturers Affymetrix and Illumina had developed 46 SNP array platforms for 25 crop species, associated with a number of markers ranging from 3K to 820K (Rasheed et al., 2017). Pricing of array genotyping is widely considered to exceed that of RRS; however, this is subject to fluctuations over time and is volume-dependent, as it varies with the number of samples and the array SNP density. Indeed, Darrier et al. (2019), considering a set of 1000 barley accessions, found that genotyping with the Illumina 50K iSelect SNP array was cheaper than GBS, with respect to both the cost per sample (£40 vs. £60.50) and the cost per marker (£0.001 vs. £0.003).

From a technical standpoint, SNP array genotyping has a series of advantages. First, genotype calls are generally accurate, even for highly heterozygous species (Bourke et al., 2018). In addition, polyploid crops represent an advantageous field of application of SNP genotyping arrays, as: for allopolyploids, NGS genotyping is complicated by sequence similarity among subgenomes, which hinders the alignment of reads to the reference genome; for autopolyploids, NGS genotyping requires very high sequencing depth and specific polyploid haplotyping algorithms, which make use of the sequence reads to determine the sequence of alleles along the same chromosomes (Motazedi et al., 2018). To date, array providers developed platforms for nine polyploid species ranging from the tetraploid potato to the dodecaploid sugarcane (reviewed by You et al., 2018), together with software solutions suitable to genotype polyploid datasets [i.e., Affymetrix's Power Tools (APT) and the Polyploid Genotyping Module within Illumina's GenomeStudio]. We highlight that while GWAS are commonly performed in allopolyploids, GWAS in autopolyploids are complicated by difficulties in the assessment of population structure and allele dosage (Rosyara et al., 2016).

A main disadvantage of SNP arrays is that they suffer from ascertainment bias (Lachance and Tishkoff, 2013), i.e., they cannot identify marker-trait associations in the case of SNPs that were not present in the population used for the development of the array. In addition, a typical drawback in the use of SNP arrays is the possibility that information (e.g., SNP chromosomal location) used for the design of the array is outdated and that there is no consistency in the use of SNPs among different genotyping array formats.

RECOMMENDATIONS FOR QUALITY CONTROL

Genotyping companies apply QC procedures depending on the method used. For NGS genotyping, these consist in removing loci with low sequencing depth (i.e., loci only supported by a few reads) and loci with low PHRED-like quality score (Q) (where Q indicates the probability that the base call is incorrect). As for array genotyping, these mainly consist in applying a clustering algorithm on fluorescence measurement data of ASO probes to distinguish samples into genotype clusters (allelic discrimination plot), and in assessing a set of QC scores on the goodness of cluster separation and signal-to-background ratio.

It should be clear that, in order to avoid bias and false signals in genotype-trait association tests, the QC procedures above mentioned are not enough and need to be complemented by others performed by the investigator, which are the focus of this paragraph. These include filtering procedures that are either common to any GWAS experiment or depend on the specific GWAS population type, as well as the characterization of the GWAS population for duplicated samples and ancestral relationships (**Figure 1**).

Application of Common Filters

A high rate of missing data at a SNP locus is considered an indication of inaccurate genotype calls (Turner et al., 2011). Therefore, filtering SNPs for call rate is typically the first step in QC. A standard rule is filtering for SNPs with call rates $\geq$95 or 99% (Anderson et al., 2010); however, a lower threshold might be chosen, especially in the case of NGS genotyping with low sequencing depth. For example, GBS-derived SNP data in crops have been filtered using call rate thresholds of 90% or lower

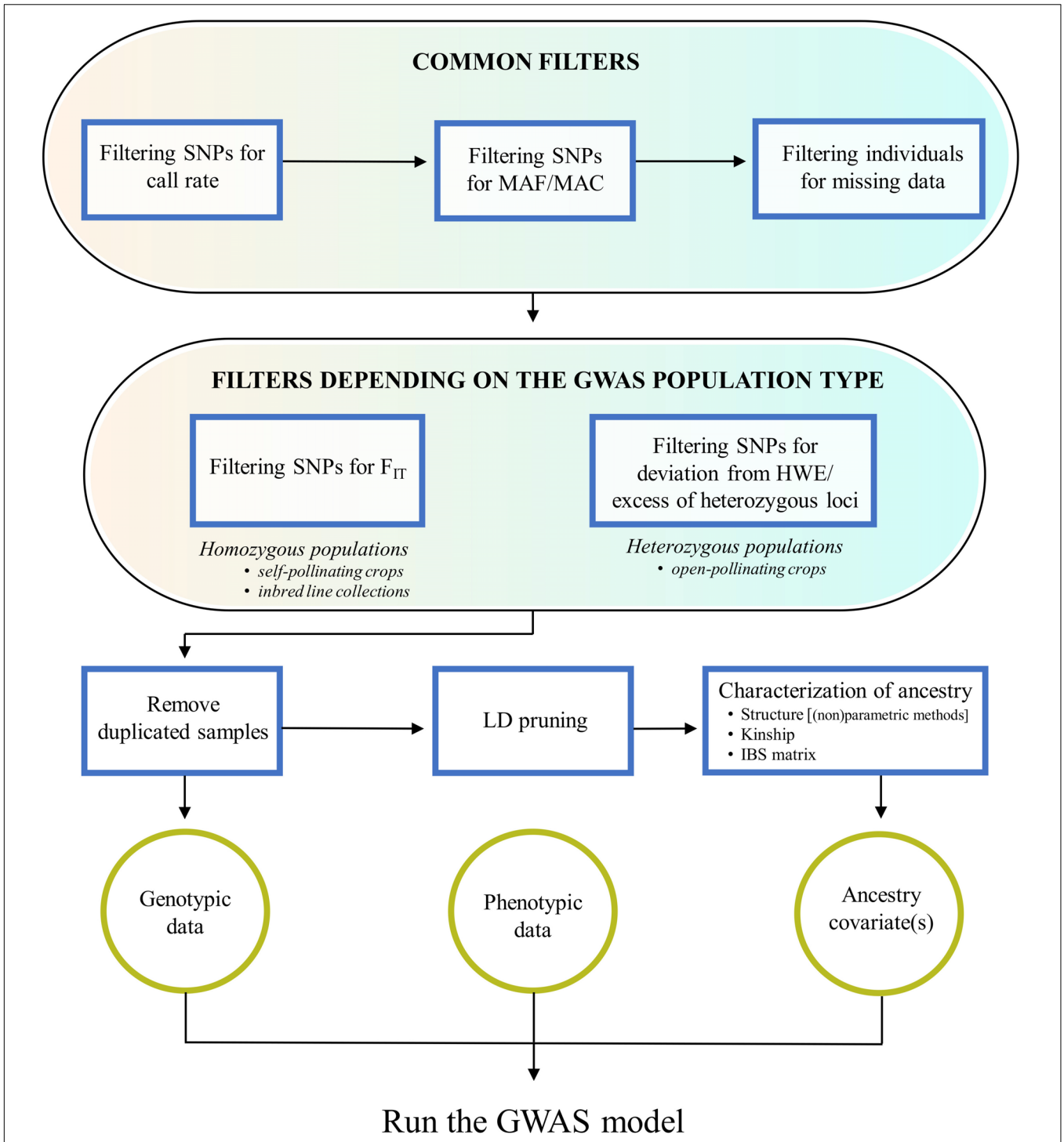


FIGURE 1 | Overview of quality control procedures for crop GWAS. These include: filtering steps that are common to any GWAS experiment; filtering steps depending on the GWAS population structure (homozygous or heterozygous); the removal of duplicated samples; the characterization of ancestral relationships, starting from a SNP dataset pruned for markers in linkage disequilibrium.

(e.g., Nimmakayala et al., 2014; Pavan et al., 2016, 2017). The overall distribution of call rates may be examined in order to set up a threshold value that eliminates classes occurring at suspiciously low frequency (**Figure 2A**).

SNP loci displaying rare variants may arise from genotyping errors and, in addition, have low statistical power to reveal association with phenotypic traits, thus they are commonly excluded by QC procedures. In this sense, a widely adopted

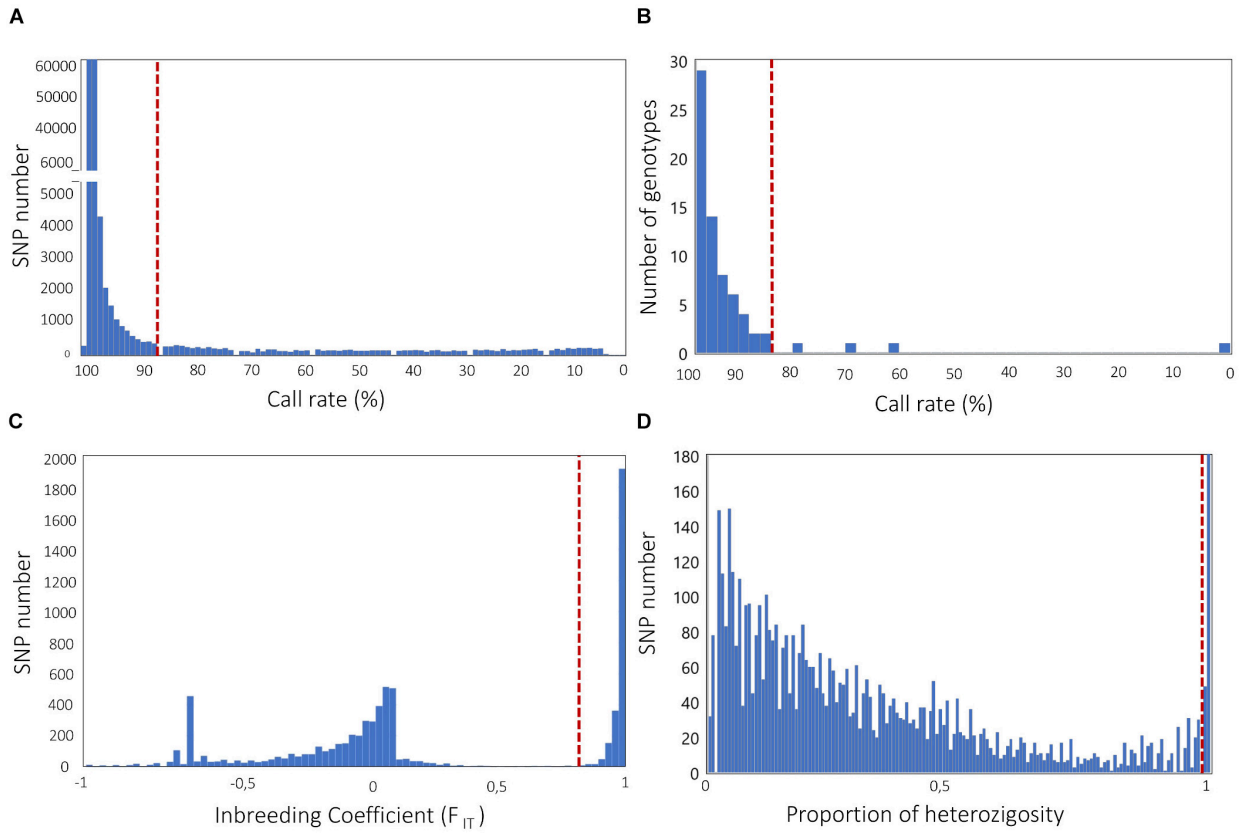


FIGURE 2 | Frequency distribution analysis to define filtering solutions for **(A)** SNP call rate; **(B)** genotype call rate; **(C)** SNP inbreeding coefficient (F_{IT}); **(D)** SNP proportion of heterozygosity. Dashed lines indicate possible filtering thresholds, based on classes occurring at suspiciously low **(A,B)** or high **(D)** frequency, and distribution gaps **(C)**. Genotypic data used to build histograms are all relative to published genotyping-by-sequencing experiments, carried out in the self-pollinated crops *Cicer arietinum* L. (Pavan et al., 2017, **A,C**) and *Lens culinaris* Medik (Pavan et al., 2019, **B**), and the open-pollinated crop *Cynara cardunculus* L. (Pavan et al., 2018, **D**).

solution is filtering for minor allele frequency (MAF). Filtering for MAF ≥1–5% has been commonly applied for crop GWAS involving populations of a few hundred individuals (Pavan et al., 2017; Yu et al., 2018), however the same thresholds might be too stringent for larger GWAS populations. Filtering for minor allele count (MAC) allows to set-up thresholds independent from the GWAS population size, commonly ranging from 5 to 10 (e.g., Taranto et al., 2016; Thomson et al., 2017).

As for loci, the presence of individuals with high rates of missing data is also suggestive of technical issues, often related with poor quality and/or quantity of DNA samples. This can generate inaccuracies and bias in downstream analyses. We emphasize that filtering for SNP missingness should normally precede filtering for individual missingness, as the opposite procedure may result in unnecessary removal of individuals. In literature, very different cutoff thresholds for individual missingness have been reported (Begum et al., 2015; Pavan et al., 2018). Our suggestion is to inspect the distribution of missing data across individuals and select a threshold that allows the elimination of classes occurring at suspiciously low frequency (**Figure 2B**). In addition, for binary traits (e.g., the response to a pathogen, for which individuals of the GWAS population can be classified in either resistant or susceptible), it is of main importance that there are no systematic differences of call rate between the two groups, in order to avoid bias in association tests.

Application of Filters Depending on the GWAS Population Type

SNP loci characterized by excessive heterozygosity should be filtered out, as they are indicative of technical artifacts or paralogous/repetitive regions that could not be distinguished through the genotyping procedure (Glaubitz et al., 2014). Therefore, specific SNP filters are applied based on the extent of heterozygosity expected in the GWAS population. For crops, this depends on the natural mating system, which may favor self-pollination or open-pollination, and anthropic interventions, such as artificial inbreeding.

Natural populations of self-pollinating crops, as well as populations of inbred lines, are highly homozygous. Therefore, in these cases, even loci with modest heterozygosity rates are suspicious. Glaubitz et al. (2014) suggested the use of the F_{IT} inbreeding coefficient (given by $1\text{-}H_o/H_E$, with H_o and H_E being the observed heterozygosity and the expected heterozygosity

at Hardy–Weinberg equilibrium, respectively) to filter SNPs in homozygous populations, and applied a minimal F_{IT} threshold of 0.8 in case of a large population of maize inbred lines. The identification of gaps in the distribution of F_{IT} across all loci may help to set up a threshold that allows the elimination of most of the genotyping errors while retaining the highest possible number of loci (**Figure 2C**).

For natural populations of open-pollinating crops, filtering SNPs that significantly deviate from the Hardy–Weinberg equilibrium (HWE) (e.g., through chi-square or exact tests) can be performed to remove excessively heterozygous loci. In accordance with GWAS on human genotypic data, the HWE filter in open-pollinating crops has been generally applied using a threshold p-value of 10^{-4}, e.g., in, cassava, olive and watermelon (Anderson et al., 2010; Nimmakayala et al., 2014; D'Agostino et al., 2018; Zhang et al., 2018). We stress here that, in crops, the HWE filter should be used with care, as there is the risk of a significant and unnecessary loss of the GWAS resolution power. Indeed, it should be firstly noticed that the HWE assumption of random mating is not respected when the population has strong genetic structure (see next paragraph) and contains some inbred individuals. Secondly, loci under selection violate by definition the HWE, thus the HWE filter might exclude loci associated with important traits under investigation. All of this considered, solutions might be to (i) adopt a relaxed threshold to eliminate markers, e.g., $p < 10^{-6}$, as previously performed on apple and globe artichoke (Bianco et al., 2016; Pavan et al., 2018); (ii) apply the HWE filter separately to each sub-population identified by the analysis of genetic structure; (iii) apply the HWE filter only to individuals not showing the phenotype supposedly under selection, in case of GWAS on binary traits. In other circumstances, including that of partially outbreeding crops, it might be advisable to avoid the HWE filter and, as a possible alternative, to eliminate SNPs with unexpected high levels of heterozygosity (**Figure 2D**).

Checking for Sample Duplication and Ancestral Relationships

In the case of crops, GWAS populations might contain several genetically identical samples. This is often caused by the occurrence, in germplasm collections, of unintended duplication of anonymous accessions and/or the occurrence of synonymous accessions. For example, genotyping with the 9K SNP array of the USDA grapevine collection revealed that 568 out of 950 accessions (58%) were genetically identical to at least another accession (Myles et al., 2011).

The identification and removal of duplicated samples is usually performed on the basis of pairwise identity-by-state (IBS) or identity-by-descent (IBD). Pairwise IBS refers to the proportion of alleles shared by two individuals, whereas pairwise IBD refers to the proportion of two individuals' genome tracing back to the same recent common ancestor (Purcell et al., 2007; Manichaikul et al., 2010). The latter is commonly estimated from pairwise IBS and allele frequency using a method-of-moment algorithm (Purcell et al., 2007). Many studies have used IBS/IBD thresholds of 95 or 99% to declare samples as identical (Myles et al., 2011). The examination of the IBS/IBD distribution associated with a few known identical samples, included on purpose in the GWAS population, might also be used to set up a threshold to estimate identity (Pavan et al., 2019).

Ancestral relationships generate LD between unlinked loci, so they are considered in the GWAS model to limit spurious associations (Astle and Balding, 2009). Therefore, a crucial step in the QC procedure is the characterization of ancestry within the GWAS population. Genetic structure (i.e., the occurrence of sub-populations with different allele frequencies) reflects remote differences in ancestry; in crops, it often originates from physical barriers to random mating and anthropic selection for specific traits, such as seed/fruit size and phenological features (Pavan et al., 2017, 2019; Siol et al., 2017). Instead, kinship reflects recent ancestry, often related to pedigree connections among modern cultivars (Taranto et al., 2020).

Starting from genotypic data, the analysis of population structure can be carried out through different approaches. Parametric methods, such as those implemented in the popular software STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009), typically estimate the allele frequency of each sub-population jointly with the membership of individuals to each sub-population, using maximum-likelihood or Bayesian statistics. The resulting matrix (known as Q-matrix), which indicates, for each individual, the proportion of the genome referable to various sub-populations, can be conveniently incorporated in GWAS models. However, it should be noticed that parametric methods are based on several genetic assumptions, including those of linkage equilibrium (LE) among markers and HWE within sub-populations. Approximate LE from the original SNP dataset can be obtained by removing markers through LD pruning algorithms (Joiret et al., 2019); on the other hand, HWE may not be met even in populations of open-pollinating crops, due to displacements, breeding activities, and clonal propagation (Campoy et al., 2016).

Non-parametric methods such as principal component analysis (PCA) and multidimensional scaling (MDS) can be used to account for population structure, using coordinates of each individual along the main PCA/MDS axes as covariates in association models (Wang et al., 2009). While non-parametric methods have the advantage of being independent on genetic assumptions, they also come with a number of issues that need to be considered. Importantly, the top PCA/MDS axes may not adequately capture variation due to population structure in the presence of other strong sources of variation, such as outlier sub-populations/individuals or family groups (Price et al., 2010; Liu et al., 2013). These latter may be frequent when the GWAS population contains many cultivars with similar pedigrees. Finally, as for parametric models, it is advisable to perform LD pruning prior to non-parametric analysis, in order to avoid noise from correlated marker data (Liu et al., 2013).

Kinship ultimately depends on the proportion of the genome that is identical-by-descent. Therefore, in order to account for kinship, the GWAS model can use IBD estimates from pedigree notes. However, it is clear that pedigrees of crop species might be in several cases unknown or inaccurate. As mentioned above in this paragraph, methods to estimate pairwise IBD from genotypic

data have been also developed. These yield a kinship matrix, also referred to as K-matrix, which has been widely used together with the Q-matrix or PCA/MDS covariates to implement the so-called Q + K and P + K GWAS models (Yu et al., 2006; Zhao et al., 2007).

We finally highlight that several works showed that a simple pairwise IBS matrix could efficiently capture both remote and recent ancestry (Zhao et al., 2007). Therefore, many GWAS models today accommodate the IBS matrix in the framework of linear mixed models, under the assumption that phenotypic variation is positively correlated with genetic distance (e.g., Kang et al., 2008, 2010).

BIOINFORMATICS TOOLS TO PERFORM QC

QC can be carried out using several bioinformatics tools, which may differ with respect to the specific action(s) performed and the file requested as input. Therefore, the investigator is often called to the conversion of genotypic data among different formats, the most common being variant call format (VCF), haplotype map (hapmap), pedigree/map (ped/map), binary (bed/bim/fam), Affymetrix chip (chp), Illumina sample map and final report, and structure. PGDSpider[1] (Lischer and Excoffier, 2012) is a dedicated tool for the conversion of genotypic data among a wide range of formats. Among other powerful conversion tools, we mention the one implemented in the software suite TASSEL (Bradbury et al., 2007), which deals with the most common formats associated with NGS genotyping, and the *gene_converter* function within the R package radiator (Gosselin, 2017), accepting and delivering 13 and 29 file formats, respectively.

Several open-source software suites are available for QC. Among the most widely used, PLINK (Purcell et al., 2007), starting from common genotypic data file formats (ped/map, bed/bim/fam and VCF), enables the application of all the SNP and individual filters presented in Sections "Application of Common Filters" and "Application of Filters Depending on the GWAS Population Type," with the exception of the F_{IT} filter. In relation to the study of genetic ancestry, it has options for LD pruning and MDS, and for the estimation of pairwise IBS and IBD.

Compared with PLINK, the abovementioned TASSEL (Bradbury et al., 2007) accepts a wider range of file formats (also including hapmap) and does not perform filtering for HWE departure. On the other hand, having been developed for GWAS on maize inbred lines, TASSEL provides the possibility to perform the F_{IT} filter. As for the genetic ancestry options, it can perform PCA/MDS and estimate pairwise IBS. While PLINK is based on command lines, thus requiring specific training by the user, TASSEL also implements a graphical user interface. Another important feature of TASSEL is the possibility to easily build histograms for SNP and individual missingness and SNP heterozygosity, which, as discussed above, are useful to set up cutoff thresholds specific for each GWAS experiment.

[1]http://www.cmpg.unibe.ch/software/PGDSpider/

Investigators with some bioinformatics skills may be interested in QC tools also enabling filtering procedures depending on the genotyping method, which, as stated above, are commonly performed through external services. For NGS genotyping, we cite VCFtools (Danecek et al., 2011), a command line software suite developed for the VCF format, which allows, among other options, filtering SNP sites and individuals based on sequencing depth and PHRED-quality score. For array genotyping, we cite the following: (i) the proprietary packages GenomeStudio and Axiom Analysis Suite, for data generated on Illumina or Affymetrix SNP array platforms, respectively; (ii) freeware tools that directly accept raw data in the original format generated by array genotyping platforms, including fluorescence intensity data necessary for QC of genotype calls. Among the many available options, we cite here the R packages argyle (Morgan, 2016) and SNPQC (Gondro et al., 2014), and the Python package ASSIsT (Di Guardo et al., 2015), for data generated on Illumina SNP array platforms, and AffyPipe (Nicolazzi et al., 2014), for data generated on Affymetrix SNP array platforms.

Finally, concerning the study of genetic structure, besides the above mentioned STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009), the EIGENSOFT utilities SMARTPCA and SMARTEIGENSTRAT are popular bioinformatics tools for, respectively, detecting and analyzing population structure via PCA, and correcting for population stratification in association studies (Price et al., 2006).

CONCLUSION

This work is thought to provide researchers, who mainly focus on the biology and breeding of crop species, with essential technical and economic aspects required to plan and carry out cost-effective and accurate GWAS. To the best of our knowledge, this is the first work specifically addressing the issue of QC in crop species, so we expect that it may contribute to the future harmonization of the procedures leading to the obtainment of high-quality SNP datasets ready for GWAS.

AUTHOR CONTRIBUTIONS

SP, ND'A, and EC conceived the review. SP, ND'A, EC, and CD wrote the manuscript. CL and LR critically revised the manuscript.

FUNDING

This research has been performed within the project "LEgume GEnetic REsources as a tool for the development of innovative and sustainable food TEchnological system" supported under the "Thought for Food" Initiative by Agropolis Fondation (through the "*Investissements d'avenir*" programme with reference number ANR-10-LABX-0001-01), Fondazione Cariplo, and Daniel & Nina Carasso Foundation.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L. A., et al. (2017). Characterization of a wheat breeders' array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (Triticum aestivum). *Plant Biotechnol. J.* 15, 390–401. doi: 10.1111/pbi.12635

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc.* 5, 1564–1573. doi: 10.1038/nprot.2010.116

Ashrafi, H., Hill, T., Stoffel, K., Kozik, A., Yao, J., Chin-Wo, S. R., et al. (2012). De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. *BMC Genomics* 13:571. doi: 10.1186/1471-2164-13-571

Astle, W., and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* 24, 451–471. doi: 10.1214/09-STS307

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., et al. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS One* 10:e0119873. doi: 10.1371/journal.pone.0119873

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., et al. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant J.* 86, 62–74. doi: 10.1111/tpj.13145

Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., et al. (2014). Development and validation of a 20K Single nucleotide polymorphism (SNP) whole genome genotyping array for apple (Malus × domestica Borkh). *PLoS One* 9:e0110377. doi: 10.1371/journal.pone.0110377

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., and Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* 9:513. doi: 10.3389/fpls.2018.00513

Boyles, R. E., Cooper, E. A., Myers, M. T., Brenton, Z., Rauh, B. L., Morris, G. P., et al. (2016). Genome-wide association studies of grain yield components in diverse sorghum germplasm. *Plant Genome* 9, 1–17. doi: 10.3835/plantgenome2015.09.0091

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Campoy, J. A., Lerigoleur-Balsemin, E., Christmann, H., Beauvieux, R., Girollet, N., Quero-García, J., et al. (2016). Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of *Prunus avium* L. landraces and bred cultivars. *BMC Plant Biol.* 16:49. doi: 10.1186/s12870-016-0712-9

Cao, K., Li, Y., Deng, C. H., Gardiner, S. E., Zhu, G., Fang, W., et al. (2019). Comparative population genomics identified genomic regions and candidate genes associated with fruit domestication traits in peach. *Plant Biotechnol. J.* 17, 1954–1970. doi: 10.1111/pbi.13112

Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, P., et al. (2016). Genome-wide association study of 12 agronomic traits in peach. *Nat. Commun.* 7:13246. doi: 10.1038/ncomms13246

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., et al. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One* 7:e31745. doi: 10.1371/journal.pone.0031745

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the Post-neolithic brassica napus oilseed. *Genome Sci.* 345, 950–953. doi: 10.1126/science.1253435

Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., et al. (2014). A high-density snp genotyping array for rice biology and molecular breeding. *Mol. Plant* 7, 541–553. doi: 10.1093/mp/sst135

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* 129, 1887–1899. doi: 10.1007/s00122-016-2746-7

Colonna, V., D'Agostino, N., Garrison, E., Albrechtsen, A., Meisner, J., Facchiano, A., et al. (2019). Genomic diversity and novel genome-wide association with fruit morphology in *Capsicum*, from 746k polymorphic sites. *Sci. Rep.* 9:10067. doi: 10.1038/s41598-019-46136-5

D'Agostino, N., Taranto, F., Camposeo, S., Mangini, G., Fanelli, V., Gadaleta, S., et al. (2018). GBS-derived SNP catalogue unveiled wide genetic variability and geographical relationships of Italian olive cultivars. *Sci. Rep.* 8:15877. doi: 10.1038/s41598-018-34207-y

D'Agostino, N., and Tripodi, P. (2017). NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. *Diversity* 9:38. doi: 10.3390/d9030038

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., et al. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front. Plant Sci.* 10:554. doi: 10.3389/fpls.2019.00544

Davey, J. W., and Blaxter, M. L. (2011). RADSeq: next-generation population genetics. *Brief. Funct. Genomics* 9, 416–423. doi: 10.1093/bfgp/elr007

Di Guardo, M., Micheletti, D., Bianco, L., Koehorst-Van Putten, H. J. J., Longhi, S., Costa, F., et al. (2015). ASSIsT: an automatic SNP scoring tool for in- and outbreeding species. *Bioinformatics* 31, 3873–3874. doi: 10.1093/bioinformatics/btv446

Dinesh, A., Patil, A., Zaidi, P. H., Kuchanur, P. H., Vinayan, M. T., and Seetharam, K. (2016). Genetic diversity, linkage disequilibrium and population structure among CIMMYT maize inbred lines, selected for heat tolerance study. *Maydica* 61, 1–7.

Diniz, A. L., Giordani, W., Costa, Z. P., Margarido, G. R. A., Perseguini, J. M. K. C., Benchimol-Reis, L. L., et al. (2019). Evidence for strong kinship influence on the extent of linkage disequilibrium in cultivated common beans. *Genes* 10:5. doi: 10.3390/genes10010005

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Xiongfeng, M., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi: 10.1038/s41588-018-0116-x

Ellison, S. L., Luby, C. H., Corak, K. E., Coe, K. M., Senalik, D., Iorizzo, M., et al. (2018). Carotenoid presence is associated with the or gene in domesticated carrot. *Genetics* 210, 1497–1508. doi: 10.1534/genetics.118.301299

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6:e28334. doi: 10.1371/journal.pone.0028334

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., Gonźalez, V. M., et al. (2012). The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci. U.S.A.* 109, 11872–11877. doi: 10.1073/pnas.1205415109

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346. doi: 10.1371/journal.pone.0090346

Gondro, C., Porto-Neto, L. R., and Lee, S. H. (2014). SNPQC-an R pipeline for quality control of Illumina SNP genotyping array data. *Anim. Genet.* 45, 758–761. doi: 10.1111/age.12198

Gosselin, T. (2017). *Radiator: RADseq Data Exploration, Manipulation and Visualization Using R. R Package Version 0.0.5.* Available at: https://github.com/thierrygosselin/radiator (accessed May 15, 2018).

Guo, S., Zhao, S., Sun, H., Wang, X., Wu, S., Lin, T., et al. (2019). Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.* 51, 1616–1623. doi: 10.1038/s41588-019-0518-4

Gur, A., Tzuri, G., Meir, A., Sa'Ar, U., Portnoy, V., Katzir, N., et al. (2017). Genome-wide linkage-disequilibrium mapping to the candidate gene level in melon (*Cucumis melo*). *Sci. Rep.* 7:9770. doi: 10.1038/s41598-017-09987-4

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., et al. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* 12:302. doi: 10.1186/1471-2164-12-302

Happ, M. M., Wang, H., Graef, G. L., and Hyten, D. L. (2019). Generating high density, low cost genotype data in soybean [*Glycine max* (L.) Merr.]. *G3 Genes Genom Genet.* 9, 2153–2160. doi: 10.1534/g3.119.400093

He, Y., Yan, L., Ge, C., Yao, X., Han, X., Wang, R., et al. (2019). Pinoid is required for formation of the stigma and style in rice. *Plant Physiol.* 180, 926–936. doi: 10.1104/pp.18.01389

Hirsch, C. D., Evans, J., Buell, C. R., and Hirsch, C. N. (2014). Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. *Brief. Funct. Genom* 13, 257–267. doi: 10.1093/bfgp/elt051

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* l.). *Front. Plant. Sci.* 9:1276. doi: 10.3389/fpls.2018.01276

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber. *Cucumis sativus* L. *Nat. Genet.* 41, 1275–1281. doi: 10.1038/ng.475

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3-Genes Genom. Genet.* 5, 1187–1209. doi: 10.1534/g3.115.018416

International Wheat Genome Sequencing and Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum Aestivum) Genome. *Science* 345:1251788. doi: 10.1126/science.1251788

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* 48, 657–666. doi: 10.1038/ng.3565

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Joiret, M., Mahachie John, J. M., Gusareva, E. S., and Van Steen, K. (2019). Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min.* 12:11. doi: 10.1186/s13040-019-0199-7

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J. M., Lee, H.-A., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* 46, 270–278. doi: 10.1038/ng.2877

Kishikawa, T., Momozawa, Y., Ozeki, T., Mushiroda, T., Inohara, H., Kamatani, Y., et al. (2019). Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.* 9:1784. doi: 10.1038/s41598-018-38346-0

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29.

Lachance, J., and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays* 35, 780–786. doi: 10.1002/bies.201300014

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic. Acids. Res.* 37, 4181–4193. doi: 10.1093/nar/gkp552

Larsen, B., Migicovsky, Z., Jeppesen, A. A., Gardner, K. M., Toldam-Andersen, T. B., Myles, S. D., et al. (2019). Genome-wide association studies in apple reveal loci for aroma volatiles, sugar composition, and harvest date. *Plant Genome* 12:180104. doi: 10.3835/plantgenome2018.12.0104

Le Paslier, M.-C., Choisne, N., Bacilieri, R., Bounon, R., Boursiquot, J.-M., Brunel, D., et al. (2013). "The GrapeReSeq 18k Vitis genotyping chip," in *Proceeding of the Ninth International Symposium on Grapevine Physiologyand Biotechnology*, La Serena.

Lee, Y.-G., Jeong, N., Kim, J. H., Lee, K., Kim, K. H., Pirani, A., et al. (2015). Development, validation and genetic analysis of a large soybean SNP genotyping array. *Plant J.* 81, 625–636. doi: 10.1111/tpj.12755

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, X.-W., Meng, X.-Q., Jia, H.-J., Yu, M.-L., Ma, R.-J., Wang, L.-R., et al. (2013). Peach genetic resources: diversity, population structure and linkage disequilibrium. *BMC Genet.* 14:84. doi: 10.1186/1471-2156-14-84

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., et al. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nat. Commun.* 10:1190. doi: 10.1038/s41467-019-09135-8

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* 46, 1220–1226. doi: 10.1038/ng.3117

Lischer, H. E. L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28, 298–299. doi: 10.1093/bioinformatics/btr642

Liu, H., and Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. *Plant J.* 97, 8–18. doi: 10.1111/tpj.14139

Liu, H., Zhan, J., Li, J., Lu, X., Liu, J., Wang, Y., et al. (2020). Genome-wide association study (GWAS) for mesocotyl elongation in rice (*Oryza sativa* L.) under multiple culture conditions. *Genes* 11:49. doi: 10.3390/genes11010049

Liu, J., He, Z., Rasheed, A., Wen, W., Yan, J., Zhang, P., et al. (2017). Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). *BMC Plant Biol.* 17:220. doi: 10.1186/s12870-017-1167-3

Liu, L., Zhang, D., Liu, H., and Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics* 14:132. doi: 10.1186/1471-2105-14-132

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, J., Liu, A., et al. (2018). Gwas analysis and qtl identification of fiber quality traits and yield components in upland cotton using enriched high-density snp markers. *Front. Plant. Sci.* 9:1067. doi: 10.3389/fpls.2018.01067

Liu, Z., Li, H., Wen, Z., Fan, X., Li, Y., Guan, R., et al. (2017). Comparison of genetic diversity between Chinese and American soybean (*Glycine max* (L.)) accessions revealed by high-density SNPs. *Front. Plant. Sci.* 8:2014. doi: 10.3389/fpls.2017.02014

Mamidi, S., Chikara, S., Goos, R. J., Hyten, D. L., Annam, D., Moghaddam, S. M., et al. (2011). Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. *Plant Genome* 4, 154–164. doi: 10.3835/plantgenome2011.04.0011

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi: 10.1093/bioinformatics/btq559

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27:e1608. doi: 10.1002/mpr.1608

Morgan, A. P. (2016). argyle: an R package for analysis of illumina genotyping arrays. *G3 Genes Genom.Genet.* 6, 281–286. doi: 10.1534/g3.115.023739

Motazedi, E., Finkers, R., Maliepaard, C., and de Ridder, D. (2018). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Brief. Bioinform* 19, 387–403. doi: 10.1093/bib/bbw126

Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., et al. (2011). Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3530–3535. doi: 10.1073/pnas.1009363108

Nicolas, S. D., Péros, J.-P., Lacombe, T., Launay, A., Le Paslier, M.-C., Bérard, A., et al. (2016). Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L.) diversity panel newly designed for association studies. *BMC Plant Biol.* 16:74. doi: 10.1186/s12870-016-0754-z

Nicolazzi, E. L., Iamartino, D., and Williams, J. L. (2014). AffyPipe: an open-source pipeline for Affymetrix Axiom genotyping workflow. *Bioinformatics* 30, 3118–3119. doi: 10.1093/bioinformatics/btu486

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Nimmakayala, P., Levi, A., Abburi, L., Abburi, V. L., Tomason, Y. R., Saminathan, T., et al. (2014). Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. *BMC Genomics* 15:767. doi: 10.1186/1471-2164-15-767

Pavan, S., Bardaro, N., Fanelli, V., Marcotrigiano, A. R., Mangini, G., Taranto, F., et al. (2019). Genotyping by sequencing of cultivated lentil (lens culinaris medik.) highlights population structure in the mediterranean gene pool associated with geographic patterns and phenotypic variables. *Front. Genet.* 10:872. doi: 10.3389/fgene.2019.00872

Pavan, S., Curci, P. L., Zuluaga, D. L., Blanco, E., and Sonnante, G. (2018). Genotyping-by-sequencing highlights patterns of genetic structure and domestication in artichoke and cardoon. *PLoS One* 13:e0205988. doi: 10.1371/journal.pone.0205988

Pavan, S., Lotti, C., Marcotrigiano, A. R., Mazzeo, R., Bardaro, N., Bracuto, V., et al. (2017). A distinct genetic cluster in cultivated chickpea as revealed by genome-wide marker discovery and genotyping. *Plant Genome* 10, 1–9. doi: 10.3835/plantgenome2016.11.0115

Pavan, S., Schiavulli, A., Marcotrigiano, A. R., Bardaro, N., Bracuto, V., Ricciardi, F., et al. (2016). Characterization of low-strigolactone germplasm in pea (*pisum sativum* L.) resistant to crenate broomrape (orobanche crenata forsk.). *Mol. Plant Microbe Interact.* 29, 743–749. doi: 10.1094/MPMI-07-16-0134-R

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11, 459–463. doi: 10.1038/nrg2813

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1111/j.1471-8286.2007.01758.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X. et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* 45, 1510–1515. doi: 10.1038/ng.2801

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5135–5140. doi: 10.1073/pnas.1400975111

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., et al. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant.* 10, 1047–1064. doi: 10.1016/j.molp.2017.06.008

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., et al. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11479–11484. doi: 10.1073/pnas.201394398

Rosyara, U. R., de Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome.* 9, 1–10. doi: 10.3835/plantgenome2015.08.0073

Rousselle, Y., Jones, E., Charcosset, A., Moreau, P., Robbins, K., Stich, B., et al. (2015). Study on essential derivation in maize: III. selection and evaluation of a panel of single nucleotide polymorphism loci for use in European and North American germplasm. *Crop Sci* 55, 1170–1180. doi: 10.2135/cropsci2014.09.0627

Rubinstein, M., Katzenellenbogen, M., Eshed, R., Rozen, A., Katzir, N., Colle, M., et al. (2015). Ultrahigh-density linkage map for cultivated cucumber (*Cucumis sativus* L.) using a single-nucleotide polymorphism genotyping array. *PLoS One.* 10:e0124101. doi: 10.1371/journal.pone.0124101

Ruggieri, V., Francese, G., Sacco, A., D'Alessandro, A., Rigano, M. M., Parisi, M., et al. (2014). An association mapping approach to identify favourable alleles for tomato fruit quality breeding. *BMC Plant Biol.* 14:337. doi: 10.1186/s12870-014-0337-9

Sasaki, T. (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Scheben, A., Batley, J., and Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotecnol. J.* 15, 149–161. doi: 10.1111/pbi.12645

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713. doi: 10.1038/ng.3008

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., et al. (2014). Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 346, 1084–1088. doi: 10.1126/science.1259215

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., van Deynze, A., et al. (2012). Development of a large snp genotyping array and generation of high-density genetic maps in tomato. *PLoS One* 7:e40563. doi: 10.1371/journal.pone.0040563

Singh, N., Jayaswal, P. K., Panda, K., Mandal, P., Kumar, V., Singh, B., et al. (2015). Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. *Sci. Rep.* 5:11600. doi: 10.1038/srep11600

Siol, M., Jacquin, F., Chabert-Martinello, M., Smıkal, P., Le Paslier, M., Aubert, G., et al. (2017). Patterns of genetic structure and linkage disequilibrium in a large collection of pea germplasm. *G3-Genes Genom Genet.* 7, 2461–2471. doi: 10.1534/g3.117.043471

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* 8:e54985. doi: 10.1371/journal.pone.0054985

Song, Q., Jia, G., Hyten, D. L., Jenkins, J., Hwang, E.-Y., Schroeder, S. G., et al. (2015). SNP assay development for linkage map construction, anchoring whole-genome sequence, and other genetic and genomic applications in common bean. *G3-Genes Genom Genet.* 5, 2285–2290. doi: 10.1534/g3.115.020594

Taranto, F., D'Agostino, N., Greco, B., Cardi, T., and Tripodi, P. (2016). Genome-wide SNP discovery and population structure analysis in pepper (*Capsicum annuum*) using genotyping by sequencing. *BMC Genomics* 17:943. doi: 10.1186/s12864-016-3297-7

Taranto, F., D'Agostino, N., Rodriguez, M., Pavan, S., Minervini, A. P., Pecchioni, N., et al. (2020). Whole genome scan reveals molecular signatures of divergence and selection related to important traits in durum wheat germplasm. *Front. Genet.* 11:217. doi: 10.3389/fgene.2020.00217

Taranto, F., Nicolia, A., Pavan, S., De Vita, P., and D'Agostino, N. (2018). Biotechnological and digital revolution for climate-smart plant breeding. *Agronomy* 8:277. doi: 10.3390/agronomy8120277

Thomson, M. J., Singh, N., Dwiyanti, M. S., Wang, D. R., Wright, M. H., Agosto Perez, F. et al. (2017). Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. *Rice* 10:40. doi: 10.1186/s12284-017-0181-2

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J. A., Huvenaars, K. H. J., et al. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One* 7:e37565. doi: 10.1371/journal.pone.0037565

Turner, S., Armstrong, L. L., Bradford, Y., Carlsony, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* 68, 1.19.1–1.19.18. doi: 10.1002/0471142905.hg0119s68

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15:823. doi: 10.1186/1471-2164-15-823

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.* 42, 833–839. doi: 10.1038/ng.654

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., et al. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494. doi: 10.1038/ng.2586

Verde, I., Bassil, N., Scalabrin, S., Gilmore, B., Lawley, C. T., Gasic, K., et al. (2012). Development and evaluation of a 9k snp array for peach by internationally coordinated snp detection and validation in breeding germplasm. *PLoS One* 7:e35668. doi: 10.1371/journal.pone.0035668

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130, 123–135. doi: 10.1007/s00122-016-2798-8

Vos, P. G., Uitdewilligen, J. G. A. M. L., Voorrips, R. E., Visser, R. G. F., and van Eck, H. J. (2015). Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* 128, 2387–2401. doi: 10.1007/s00122-015-2593-y

Wang, D., Sun, Y., Stang, P., Berlin, J. A., Wilcox, M. A., and Li, Q. (2009). Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proc.* 3(Suppl. 7):S109. doi: 10.1186/1753-6561-3-s7-s109

Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., et al. (2016). The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. *Mol. Plant.* 9, 975–985. doi: 10.1016/j.molp.2016.04.018

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, X., Bao, K., Reddy, U. K., Bai, Y., Hammar, S. A., Jiao, C., et al. (2018). The USDA cucumber (*Cucumis sativus* L.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development. *Hortic. Res.* 5:64. doi: 10.1038/s41438-018-0080-8

Wen, T., Dai, B., Wang, T., Liu, X., You, C., and Lin, Z. (2019). Genetic variations in plant architecture traits in cotton (*Gossypium hirsutum*) revealed by a genome-wide association study. *Crop J.* 7, 209–216. doi: 10.1016/j.cj.2018.12.004

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L. A., Benbow, H. R., Wilkinson, P. A., et al. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* 14, 1195–1206. doi: 10.1111/pbi.12485

Xu, C., Ren, Y., Jian, Y., Guo, Z., Zhang, Y., Xie, C., et al. (2017). Development of a maize 55 K SNP array with improved genome coverage for molecular breeding. *Mol. Breed.* 37:20. doi: 10.1007/s11032-017-0622-z

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., et al. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Yan, J., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., and Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 4:e8451. doi: 10.1371/journal.pone.0008451

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.-C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* 9:104. doi: 10.3389/fpls.2018.00104

Yu, H., Xie, W., Li, J., Zhou, F., and Zhang, Q. (2014). A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol. J.* 12, 28–37. doi: 10.1111/pbi.12113

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Yu, Y., Fu, J., Xu, Y., Zhang, J., Ren, F., Zhao, H., et al. (2018). Genome re-sequencing reveals the evolutionary history of peach fruit edibility. *Nat. Commun.* 9:5404. doi: 10.1038/s41467-018-07744-3

Yuan, Y., Wang, X., Wang, L., Xing, H., Wang, Q., Saeed, M., et al. (2018). Genome-wide association study identifies candidate genes related to seed oil composition and protein content in *Gossypium hirsutum* L. *Front. Plant Sci.* 9:1359. doi: 10.3389/fpls.2018.01359

Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., et al. (2018). Genome-wide association studies of 11 agronomic traits in cassava (*Manihot esculenta* crantz). *Front. Plant Sci.* 9, 503. doi: 10.3389/fpls.2018.00503

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An Arabidopsis example of association mapping in structured sample. *PLoS Genet.* 3:e4. doi: 10.1371/journal.pgen.0030004

Zhao, X., Li, B., Zhang, K., Hu, K., Yi, B., Wen, J., et al. (2016). Breeding signature of combining ability improvement revealed by a genomic variation map from recurrent selection population in *Brassica napus*. *Sci. Rep.* 6:29553. doi: 10.1038/srep29553

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Pavan, Delvento, Ricciardi, Lotti, Ciani and D'Agostino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 October 2020
doi: 10.3389/fgene.2020.499094

Marker Selection in Multivariate Genomic Prediction Improves Accuracy of Low Heritability Traits

Jaroslav Klápště[1]*, *Heidi S. Dungey*[1], *Emily J. Telfer*[1], *Mari Suontama*[1,2], *Natalie J. Graham*[1], *Yongjun Li*[1,3] *and Russell McKinley*[1]

[1] Scion (New Zealand Forest Research Institute Ltd.), Rotorua, New Zealand, [2] Skogforsk, Umeå, Sweden, [3] Agriculture Victoria, AgriBio Center, Bundoora, VIC, Australia

Multivariate analysis using mixed models allows for the exploration of genetic correlations between traits. Additionally, the transition to a genomic based approach is simplified by substituting classic pedigrees with a marker-based relationship matrix. It also enables the investigation of correlated responses to selection, trait integration and modularity in different kinds of populations. This study investigated a strategy for the construction of a marker-based relationship matrix that prioritized markers using Partial Least Squares. The efficiency of this strategy was found to depend on the correlation structure between investigated traits. In terms of accuracy, we found no benefit of this strategy compared with the all-marker-based multivariate model for the primary trait of diameter at breast height (DBH) in a radiata pine (*Pinus radiata*) population, possibly due to the presence of strong and well-estimated correlation with other highly heritable traits. Conversely, we did see benefit in a shining gum (*Eucalyptus nitens*) population, where the primary trait had low or only moderate genetic correlation with other low/moderately heritable traits. Marker selection in multivariate analysis can therefore be an efficient strategy to improve prediction accuracy for low heritability traits due to improved precision in poorly estimated low/moderate genetic correlations. Additionally, our study identified the genetic diversity as a factor contributing to the efficiency of marker selection in multivariate approaches due to higher precision of genetic correlation estimates.

Keywords: multivariate mixed model, genomic prediction, variable selection, PLS, *Pinus radiata*, *Eucalyptus nitens*

OPEN ACCESS

Edited by:
Charles Chen, Oklahoma State University, United States

Reviewed by:
Freddy Mora-Poblete, University of Talca, Chile
Fernando H. Toledo, International Maize and Wheat Improvement Center, Mexico

***Correspondence:**
Jaroslav Klápště
jaroslav.klapste@scionresearch.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *19 September 2019*
Accepted: *18 September 2020*
Published: *30 October 2020*

Citation:
Klápště J, Dungey HS, Telfer EJ, Suontama M, Graham NJ, Li Y and McKinley R (2020) Marker Selection in Multivariate Genomic Prediction Improves Accuracy of Low Heritability Traits. Front. Genet. 11:499094. doi: 10.3389/fgene.2020.499094

1. INTRODUCTION

Heritability is one of the most important genetic parameters to consider for breeding, defined as the proportion of phenotypic variance explained by underlying genetic factors (Falconer and Mackay, 1996). Trait heritability is affected by changes in allelic frequencies due to selection or inbreeding, introduction of new alleles through mutation or migration (Latta, 2010), or due to changes in genetic effect due to altered genetic backgrounds or environmental conditions (Chandler et al., 2017). Quantitative traits normally present low to moderate heritability, as a result of their genetic control and the high degree of environmental influence on the expression of these traits. In tree breeding, important quantitative traits, such as height, diameter at breast height and stem volume generally have relatively low to moderate heritability estimates, ranging from 0.09 to 0.3 (Ukrainetz et al., 2008; Chen et al., 2018; Hayatgheibi et al., 2019). Furthermore, the magnitude and precision of these heritability estimates vary with the testing effort (such as sample size, experimental, and mating design) and the ontogenetic stage of individuals in the population being tested (Bouvet et al., 2003; Mihai and Mirancea, 2016). Reports of low heritability for productivity traits is not

surprising, as they are assumed to be essential for individual tree survival and thus likely close to fixation (King, 1990; Merilä and Sheldon, 2000; Blows and Hoffmann, 2005). Unfortunately, both low heritability and less accurate estimates of breeding values makes selection decisions challenging for such traits and slows progress in genetic improvement.

The current rapid development of genomic resources in forest tree species (Neale and Kremer, 2011; Nystedt et al., 2013; Myburg et al., 2014; Neale et al., 2014) has improved forest tree breeding practices through the implementation of genomic prediction (Meuwissen et al., 2001; Grattapaglia and Resende, 2011; Isik, 2014; Grattapaglia et al., 2018). Genomic best linear unbiased prediction (GBLUP) is the most popular method for genomic prediction, due to the simple substitution of the average numerator relationship matrix (Wright, 1922) with a marker-based relationship matrix (Nejati-Javaremi et al., 1997; VanRaden, 2008; Hayes et al., 2009). Such a relationship matrix allows tracking of both recent and historical relatedness (Powell et al., 2010), as well as Mendelian segregation (Visscher et al., 2006) and linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs) (Habier et al., 2013). The ultimate goal of genomic prediction is the development of model using mainly LD between markers and QTLs which would support predictive ability stable across generations. Sun et al. (2016) found that the accuracy of such model across generations is high only when the historical LD between markers and QTLs is high. Alternatively, the capture of co-segregation improves accuracy of the prediction when effective population is relatively small. Additionally, the accuracy of genomic prediction critically depends on the level of relatedness between the training and validation populations (Scutari et al., 2016).

While genetic correlations often represent evolutionary constraints (Clark, 1987), they are also a means to improve the accuracy of genetic parameters (Calus and Veerkamp, 2011) and reduce bias of estimated breeding values caused by selection on correlated trait through use of a multivariate instead of univariate approach (Pollak et al., 1984). The use of multivariate linear mixed models in genetic evaluations provides a basis for inference about traits' integration (Armbruster et al., 2014) as well as evolutionary response to selection (Sedlacek et al., 2016). Additionally, these types of models could deliver improvements in the accuracy of genetic parameters, especially where traits with low heritability can be analyzed together with traits of high heritability, and genetic covariances can be taken into consideration (Jia and Jannink, 2012; Marchal et al., 2016). Guo et al. (2014) reported an advantage to using multi-trait genomic predictions over single-trait alternatives when traits had low heritability or if phenotypic records were lacking. The traits with low heritability (Stejskal et al., 2018) benefited the most from the implementation of genomic information in the genetic analysis (Meuwissen et al., 2001). Therefore, a combination of both approaches in a genomic-based multivariate mixed linear model might provide the best results. However, both approaches have their drawbacks. Multivariate analysis can provide benefits to low heritability traits only in cases where there are strong genetic correlations with other traits, while no benefit or even reductions in breeding values accuracy can result when genetic correlations are weak (Jia and Jannink, 2012). Furthermore, optimization of the population sample size, effective population size and the level of genetic diversity captured is required to reach statistically significant genetic correlations (Bijma and Bastiaansen, 2014).

The majority of complex quantitative traits follows Fisher's infinitesimal model (Fisher, 1918) where each QTL contributes by only small fraction of total genetic variance. Such traits require genomic prediction models using large amount of genetic marker densely populating whole genome (Meuwissen et al., 2001; Guo et al., 2010). However, some traits show a positive response in prediction accuracy as a result of marker selection (Resende et al., 2012), depending on the structure of the training population and the genetic complexity of the investigated trait (Berger et al., 2015). Bayesian models have proven an efficient way to consider different variances for the distribution of marker effects which might result in an improvement in genomic predictions over classical GBLUP, especially in cases where the underlying genetic architecture of a trait involves large-effect QTLs (Cole et al., 2009).

Alternatively, construction of a trait-specific relationship matrix, considering marker-specific weights, provides a viable alternative (Zhang et al., 2010; Su et al., 2014). Lippert et al. (2013) investigated the ratio of causal and non-causal variants present in genomic data, and found that the most precise genetic parameter estimates are obtained when only causal variants are included in the prediction model. de los Campos et al. (2015) argued that a large number of markers in imperfect LD with QTLs can produce false inferences about heritability due to instability in likelihood estimates, especially when LD decays rapidly. Additionally, using an exhaustive amount of genomic information in genetic analyses can potentially reduce the precision of genetic parameters and the accuracy of genomic estimated breeding values (Habier et al., 2007, 2013).

Similar to single-trait genomic prediction models, several marker selection strategies have been developed within multi-trait genomic prediction models. Classical multiple regression models assign effects to every marker, which is not necessarily biologically true. Cheng et al. (2018), therefore, developed a Bayesian multi-trait model which allows for the assumption that each marker affects only one or a few traits, and has no effect on other traits. Karaman et al. (2018) applied an alternative approach using posterior estimates of marker effect covariances to weight their contribution to the marker-based relationship matrix, implemented in the GBLUP model. They found a further advantage to this weighted marker-based relationship matrix when weights were assigned to blocks of 100 SNPs, rather than to each marker separately.

The aim of this study is the improvement of genomic prediction for traits with relatively low heritability and poor prediction accuracy, such as those related to forest tree productivity (Gamal El-Dien et al., 2015; Ratcliffe et al., 2015), through the implementation of multi-trait models using a relationship matrix based only on prioritized markers. Our primary trait under investigation was diameter at breast height (DBH) for radiata pine (*Pinus radiata* D.Don) and shining gum [*Eucalyptus nitens* (H. Deane & Maiden) Maiden], a proxy for productivity in forest trees and thus considered the most

economically important trait for those species. Non-target traits involved in the multivariate analysis represent operationally measured attributes related to stem form and wood quality.

2. MATERIALS AND METHODS

2.1. Plant Material

2.1.1. Radiata Pine (*Pinus radiata*)

The *P. radiata* population used in this study included 523 vegetatively propagated individuals (four ramets per individual genotype), structured into 42 full-sib families each represented by ∼10 individual genotypes, part of The New Zealand Radiata Pine Breeding Company's (RPBC) program, selected for growth and form attributes. The field experiment was established as an incomplete block design containing nine blocks, each comprising six families with five replicates per family. All individuals were evaluated for the following traits: branch cluster frequency (BR9), visually assessed using a 9-point scale from 1 (uninodal) to 9 (extremely multinodal) (Carson, 1991); stem straightness (ST9), visually assessed using a 9-point scale from 1 (crooked) to 9 (very straight) (Carson, 1986); diameter at breast height (DBH [cm]) measured with diameter tape; wood density (WD, [kg/m^3]), measured as basic wood density through the maximum moisture content method (Smith, 1954); and predicted modulus of elasticity (PME [GPa]), inferred from acoustic wave velocity using HITMAN (HM200) (Carter et al., 2007).

Genomic data were generated through an exome capture-based Genotype-By-Sequencing (GBS) platform (Neves et al., 2013), developed using in-house genomic resources (Telfer et al., 2018). The captured markers were filtered using a previously reported bioinformatics pipeline (Telfer et al., 2019). In brief, markers were removed if heterozygosity in haploid megagametophyte tissues was higher than 5%, average read depth was <10 (mean average read depth per marker was ∼60 in our data) and have more than 1 alternative allele. Individual datapoints were classified as missing if the ratio between the reference and alternative allele was lower than 0.1 and the number of read was <10 (Telfer et al., 2019). In total, 80,160 SNPs passed the criteria, and were further filtered to remove SNPs with minor allele frequencies (MAF) <0.05 and a SNP call rate <0.6. The average proportion of SNP missing data was 9.9%. The genotype mean was used to impute missing data and 58,636 SNPs were used in downstream analysis.

2.1.2. Shining Gum (*Eucalyptus nitens*)

The *E. nitens* population used in this study included 691 individuals, part of the third generation of open-pollinated progeny established within New Zealand's breeding program. The experimental design contained 30 replications of randomized complete blocks of these "sets" with each replication of the "set" comprising the same families but different individuals within these families (Klápště et al., 2019). Missing relatedness information in this population was recovered using sib-ship reconstruction as genomic information was not available for all possible parents (Klápště et al., 2017). This sib-ship reconstruction-based relationship matrix was used in both the genomic-based and pedigree-based scenarios in this study. The individuals within the open-pollinated progeny trial were phenotyped for diameter at breast height at age 6 (DBH [mm]) and for wood quality traits, such as wood density (WD [kg/m^3]), wood stiffness (ST [km/s]), growth strain (GS [mm]), and average tangential air-dry shrinkage (TS [%]) measured on two different logs: log 1 from 1.4 to 3 m (index 1) and log 2 from 3 to 6 m (index 2) at the age of 7 (Klápště et al., 2017). Diameter at breast height was measured with diameter tape, wood density was measured as basic wood density through the maximum moisture content method (Smith, 1954), wood stiffness was measured indirectly as acoustic wave velocity using HITMAN (HM200) (Carter et al., 2007), growth strain was assessed by ripping logs with a chainsaw and measuring the resulting openings at the end of the log and average tangential air-dry shrinkage was measured following standard wood quality assessment protocols (Treloar and Lausberg, 1997).

Genomic data were generated using the EUChip60K SNP chip (Silva-Junior et al., 2015). SNP genotypes were called using the *Maidenaria* section specific cluster files (Silva-Junior et al., 2015) and filtered using Illumina metrics genTrain score >0.5 and GenCall >0.15, in addition to MAF >0.01 and call rate >0.6. The average proportion of SNP missing data was 5.8%. The genotype mean was used to impute missing data, with 9,697 SNPs used in downstream analysis.

2.2. Statistical Analysis

A univariate model was used to estimate variance components and derive narrow-sense heritability for both species using the following mixed linear model implemented in statistical package ASReml-R (Butler et al., 2009):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{Zb} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of individual-tree trait measurements, $\boldsymbol{\beta}$ is the vector of fixed effects (intercept and replicate, as well as seed orchard in the case of *E.nitens*), $\boldsymbol{g}$ is the vector of random additive genetic values following var($\boldsymbol{g}$)∼N(0, $\boldsymbol{A}\sigma_g^2$), where σ_g^2 is the genotypic variance and $\boldsymbol{A}$ is the average numerator relationship matrix (Wright, 1922), $\boldsymbol{b}$ is the vector of random block effects nested within replication effects following var($\boldsymbol{b}$)∼N(0, $\boldsymbol{I}\sigma_b^2$), where σ_b^2 is block nested within replication variance, $\boldsymbol{e}$ is the vector of random residual effects following var($\boldsymbol{e}$)∼N(0, $\boldsymbol{I}\sigma_e^2$), and where σ_e^2 is the residual variance.

Additionally, a univariate model was used to estimate best linear unbiased estimates (BLUEs) for genotype in *P. radiata* as well as to correct phenotypes for design effects in the *E. nitens* population using the following mixed linear model implemented in statistical package ASReml-R (Butler et al., 2009):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of individual-tree trait measurements, $\boldsymbol{\beta}$ is the vector of fixed effects (intercept, replicates and block nested within replicates, and genotype in the case of *P. radiata*), $\boldsymbol{e}$ is the vector of random residual effects following var($\boldsymbol{e}$)∼N(0, $\boldsymbol{I}\sigma_e^2$), and where σ_e^2 is the residual variance.

The BLUE estimates for genotypes for *P. radiata* and corrected phenotypes for *E. nitens* were used along with the genomic data

to estimate marker weights prior to construction of the marker-based relationship matrix. Weights for marker selection were derived through two blocks of canonical partial least squares (PLS-CA) (Tenenhaus, 1998) implemented using the "plsca" function from the R package "plsdepot" (Sanchez and Sanchez, 2012). The algorithm computes sequences of pairs of vectors of latent scores which are orthogonal by maximization of Cov($\boldsymbol{Xu}$, $\boldsymbol{Yv}$), where $\boldsymbol{X}$ is the scaled matrix of marker genotypes and $\boldsymbol{Y}$ is the scaled matrix of clonal values for measured traits, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of coefficients maximizing the covariance. The coefficients in $\boldsymbol{u}$ measure the importance of variables in $\boldsymbol{X}$ (genetic markers) to latent variables, and were therefore used as criteria for selection of markers to calculate the marker-based relationship matrix. Since prior knowledge of genetic architecture in studied traits and complexity of pleiotropy and QTL collocation is usually lacking, exploration of the whole matrix of combinations of selection intensity for potentially informative genetic markers was required. First, marker coefficients in the vector $\boldsymbol{u}$ associated with each component were truncated by the 90th, 80th, 70th, 60th, and 50th percentiles, and loadings for selected markers were transformed to either 1 or 0. For each percentile level, different numbers of components were included into the marker selection process.

Univariate models using corrected phenotypes and pedigree (BLUP) or marker information (GBLUP) were used to estimate narrow-sense heritability (the proportion of additive to total genetic variance in the case of *P. radiata*) and prediction accuracy using the "BGLR" statistical R package (Pérez and de Los Campos, 2014), as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{e}$$

where y is the vector of corrected phenotypes/genotypic values, $\boldsymbol{\beta}$ is the vector of fixed effects (overall mean), $\boldsymbol{g}$ is the vector of additive genetic effects following var(g)∼N(0, $\boldsymbol{A}\sigma_g^2$), where $\boldsymbol{A}$ is the average numerator relationship matrix (Wright, 1922) in the BLUP analysis, and is substituted by marker-based relationship $\boldsymbol{G}$ (VanRaden, 2008) in the GBLUP analysis, σ_g^2 is additive genetic variance, $\boldsymbol{e}$ is the vector of residuals following var(e)∼N(0, $\boldsymbol{I}\sigma_e^2$), where $\boldsymbol{I}$ is the identity matrix and σ_e^2 is residual variance.

Since the aim of the algorithm is the maximization of covariance among genomic and phenotypic data, the first scenario selects only markers with the highest positive coefficients, which have an associated positive effect with the underlying covariance/correlation structure (positive pleiotropy) (scenario MVGBLUP1). However, the relationship between traits is not driven only by markers acting in the same direction; some markers act in the same direction only for certain sets of traits, and in opposite directions for other traits (negative pleiotropy). To investigate the impact of such markers, we tested a second scenario where markers involved in the construction of the relationship matrix were selected from both positive and negative tails of the loading distribution (scenario MVGBLUP2). For example, in the 90th percentile scenario, markers were selected from both above the 90th percentile and from below the 10th percentile. The other scenarios continued to select the markers having loadings closer to the middle of their distribution. Again, this marker selection strategy was applied across the variable number of components included in this study. The improved marker-based estimates of genetic correlation were performed using marker weights implemented in the construction of a trait-specific marker-based relationship matrix (Zhang et al., 2010) as follows:

$$\boldsymbol{G_w} = \frac{\boldsymbol{ZWZ'}}{\sum w_i}$$

where $\boldsymbol{G_w}$ is the marker-based relationship matrix, $\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P}$, where $\boldsymbol{M}$ is the matrix of genotypes coded as 0, 1, and 2 for reference allele homozygotes, heterozygotes and the alternative allele homozygotes, respectively, $\boldsymbol{P}$ is the vector of doubled allelic frequencies for the alternative allele, $\boldsymbol{W}$ is the diagonal matrix of weights and w_i is the weight for the ith marker. The effect of SNP selection on the precision of genetic parameters and prediction accuracy of genomic estimated breeding values was investigated through multivariate mixed linear modeling using Gibbs sampling, performed in the "MTM" package (de los Campos and Grüneberg, 2016) implementing algorithms from the "BGLR" statistical R package (Pérez and de Los Campos, 2014), as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is a matrix of phenotypes, $\boldsymbol{a}$ is the vector of random genomic breeding values following var($\boldsymbol{a}$)∼N(0,G1), where G1 is a variance-covariance structure for additive genetic effects following G1=$\begin{bmatrix} \sigma_{a_1}^2 & \cdots & \sigma_{a_1 a_n} \\ \vdots & \ddots & \vdots \\ \sigma_{a_n a_1} & \cdots & \sigma_{a_n}^2 \end{bmatrix} \otimes \boldsymbol{G}$, where $\sigma_{a_1}^2$ and $\sigma_{a_n}^2$ are additive genetic variances for the 1st and nth trait, respectively, $\sigma_{a_1 a_n}$ and $\sigma_{a_n a_1}$ are additive genetic covariances between the 1st and nth trait, $\otimes$ is the Kronecker product and $\boldsymbol{G}$ is the marker-based relationship matrix estimated either as follows:

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\sum p_i(1-p_i)}$$

where p_i is the frequency of the alternative allele at the ith loci, or estimated on the basis of weighted markers ($\boldsymbol{G_w}$) as defined above, $\boldsymbol{e}$ is the vector of random residual effects following var($\boldsymbol{e}$)∼N(0,R), where R is the residual variance-covariance structure following R = $\begin{bmatrix} \sigma_{e_1}^2 & \cdots & \sigma_{e_1 e_n} \\ \vdots & \ddots & \vdots \\ \sigma_{e_n e_1} & \cdots & \sigma_{e_n}^2 \end{bmatrix} \otimes \boldsymbol{I}$, where $\sigma_{e_1}^2$ and $\sigma_{e_n}^2$ are residual variances for the 1st and nth trait, and $\sigma_{e_1 e_n}$ and $\sigma_{e_n e_1}$ are residual covariances between the 1st and nth trait. The number of iterations in BGLR was set to 300,000, with a burn-in period of 50,000 iterations, thinning to 10. Given the different percentiles of marker loadings and numbers of latent variables used in marker selection, the best scenario was identified on the basis of the deviance information criterion (DIC). Additionally, single-trait model (scenarios BLUP and GBLUP) were implemented for each investigated trait to evaluate

the benefit of the multivariate model over univariate analysis. Trait heritability was estimated following:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

where σ_a^2 is additive genetic and σ_e^2 is residual variance. Genetic correlations were estimated through Pearson's product moment as follows:

$$r_G = \frac{\sigma_{a_x a_y}}{\sqrt{\sigma_{a_x}^2 \sigma_{a_y}^2}}$$

where $\sigma_{a_x a_y}$ is the additive genetic covariance between the *x*th and *y*th trait, and $\sigma_{a_x}^2$ and $\sigma_{a_y}^2$ are the additive genetic variances for the xth and yth trait, respectively. The multivariate scenarios using all available markers (MVGBLUP) or pedigree/sib-ship reconstruction (MVBLUP) were considered as benchmarks in this study.

Independent evaluation was performed using a 10-fold cross-validation. Nine-folds formed the training population, where PLS-CA was performed to obtain marker weights and construct the marker-based matrix from selected markers. The 10th-fold was used as the validation population to predict genomic breeding values (GEBV). The prediction accuracy was estimated as correlations between EBVs and GEBVs predicted through cross-validation. The statistical significance of difference in prediction accuracy between benchmark and the best scenario using selected markers, non-parametric Wilcoxon rank test was implemented (Wilcoxon, 1992).

3. RESULTS

3.1. Genetic Parameters

Discriminant analysis of principal components (DAPC) (Jombart et al., 2010) was performed to investigate population structure. We found almost no support for population stratification in *E. nitens* and scenario with two clusters showed the best fit of the data (**Supplementary Figure 1**). This scenario identified clusters associated to the each seed orchard progeny. The same approach applied in *P. radiata* selected seven clusters as the best scenario considering fit of the data (**Supplementary Figure 1**). The exploration of marker-based relationship matrices within each population through principal component analysis (PCA) found relatively weak stratification, mostly due to the separation of families accounting for 1.5–2.04% (*E. nitens*) and 3.44–3.79% (*P. radiata*) of the total variance attributed to the first two principal components (**Supplementary Figure 2**, upper plots). The distribution of relatedness showed that the majority of matrix elements had no or very weak relatedness. Additionally, there is a peak around 0.2, representing half-sibs in the *E. nitens* population, and two peaks around 0.2 and 0.4 in the *P. radiata* population, representing half-sibs and full-sibs (**Supplementary Figure 2**, bottom plots) corresponding to the mating strategy implemented at each population. The mean sample observed heterozygosity was ∼0.29 in *E. nitens* and ∼0.19 in *P. radiata*. The self-relatedness was distributed around 1 in *P. radiata*, but shifted to around 0.75 in *E. nitens* due to the higher level of inbreeding (**Supplementary Figure 3**).

Trait heritabilities were estimated using variance components inferred from a sib-ship reconstruction-based (BLUP) as well as marker-based (GBLUP) univariate model in *E. nitens*, and from a pedigree-based (BLUP) as well as marker-based (GBLUP) univariate model in *P. radiata*. Heritability estimates were moderate to high, ranging from 0.093 (ST2) to 0.282 (WD) using sib-ship (BLUP) and from 0.089 (DBH) to 0.559 (WD) using markers (GBLUP) in *E. nitens*, and from 0.046 (ST9) to 0.588 (WD) using pedigree (BLUP) and from 0.126 (ST9) to 0.529 (WD) using markers (GBLUP) in *P. radiata* (**Table 1**). In general, marker-based analysis (GBLUP) resulted in higher heritability estimates than pedigree/sib-ship based (BLUP) analysis.

In *E. nitens*, genetic correlations ranged from −0.459 (between WD and GS2) to 0.859 (between GS1 and GS2) using sib-ship (MVBLUP) (**Figure 1**—left plot below diagonals), and from −0.113 (between WD and GS2) to 0.929 (between GS1 and GS2) using markers (MVGBLUP) (**Figure 1**—left plot above diagonals). In *P. radiata*, genetic correlations ranged from −0.978 (between DBH and WD) to 0.548 (between WD and PME) using the pedigree (MVBLUP) (**Figure 1**—right plot below diagonals), and from −0.987 (between DBH and WD) to 0.602 (between WD and PME) using markers (MVGBLUP) (**Figure 1**—right plot above diagonals). Genetic correlations showed a more complex pattern in *E. nitens* compared with *P. radiata* (**Figure 2**).

3.2. Marker Selection

Using PLS-CA resulted in the construction of marker-based relationship matrices using different numbers of markers. When only markers with positive loadings (MVGBLUP1) were used, the number of selected markers ranged from 970 to 9,627 in *E. nitens* and from 5,864 to 56,809 in *P. radiata*. Scenarios which considered markers with both positive and negative loadings (MVGBLUP2) resulted in the number of selected markers ranging from 1,940 to 9,697 in *E. nitens* and from 9,838 to 58,636 in *P. radiata* (**Table 2**).

The most intensive marker selection in the *E. nitens* population resulted in the worst model fit in terms of deviance information criteria (DIC). The model fit continually improved with more relaxed parameters on marker loadings. This pattern was observed for both tested strategies (MVGBLUP1 and MVGBLUP2). The best scenario appeared close to the one using all markers (MVGBLUP) (using seven components and the 40th percentile) (**Supplementary Table 1**). There was no real pattern to the number of markers selected in the *P. radiata* population, with the best model fit found for the scenario that used four latent components and the 50th percentile (**Supplementary Table 1**).

Comparison of the marker-based relationship matrix using all markers with matrices using only selected subsets of markers showed correlations from 0.73 to 0.99 in *E. nitens*. Similarly, in *P. radiata*, correlations reached values from 0.57 to 0.99. In both populations, the genetic correlations increased as the number of components as well as the proportion of markers selected within components increased (**Figure 3**).

TABLE 1 | Heritability estimates and their 95% confidence limits using variance components inferred from the sib-ship reconstruction-based univariate model (BLUP) in *E. nitens* and from using the pedigree-based univariate model (BLUP) in *P. radiata* as well as marker-based univariate models (GBLUP).

	E. nitens		*P. radiata*	
Trait	**Pedigree**	**Markers**	**Pedigree**	**Markers**
TS	0.242 (0.147–0.338)	0.539 (0.389–0.689)	NA	NA
WD	0.282 (0.193–0.371)	0.559 (0.420–0.699)	0.588 (0.292–0.884)	0.529 (0.400–0.658)
DBH	0.138 (0.030–0.245)	0.089 (−0.049–0.228)	0.134 (0.024–0.244)	0.131 (0.052–0.210)
ST1	0.210 (0.107–0.313)	0.394 (0.229–0.559)	NA	NA
ST2	0.093 (−0.001–0.187)	0.199 (0.044–0.354)	NA	NA
GS1	0.248 (0.139–0.357)	0.309 (0.149–0.469)	NA	NA
GS2	0.211 (0.103–0.319)	0.318 (0.154–0.481)	NA	NA
ST9	NA	NA	0.046 (−0.010–0.102)	0.126 (0.034–0.218)
BR9	NA	NA	0.128 (0.019–0.237)	0.177 (0.073–0.282)
PME	NA	NA	0.224 (0.055–0.393)	0.397 (0.250–0.544)

NA represents the case where data were not available for a particular species and trait.

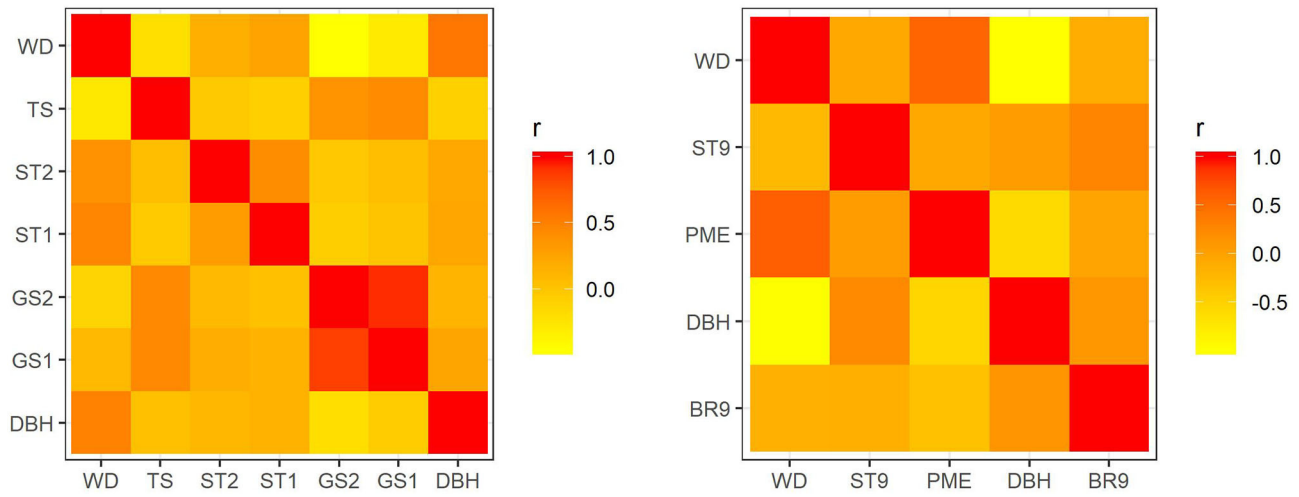


FIGURE 1 | Genetic correlations using variance components and covariances inferred from use of a sib-ship reconstruction-based multivariate model (MVBLUP) (below diagonals) and marker-based relationship matrix (MVGBLUP) (above diagonals) in the *E. nitens* population (left plot) and using variance components and covariances inferred from use of a pedigree-based multivariate model (MVBLUP) (below diagonals) and marker-based relationship matrix (MVGBLUP) (above diagonals) in the *P. radiata* population (right plot).

3.3. Prediction Accuracy

Prediction accuracy in the pedigree/sib-ship based model (BLUP) ranged from 0.246 (DBH) to 0.782 (WD) in *E. nitens*, and from 0.441 (DBH) to 0.653 (BR9) in *P. radiata*. In marker-based models (GBLUP), this ranged from 0.183 (DBH) to 0.764 (WD) in *E. nitens*, and from 0.388 (DBH) to 0.645 (WD) in *P. radiata*. In general, the implementation of single-trait models (BLUP and GBLUP) resulted in lower prediction accuracies when the marker-based model (GBLUP) was compared to the pedigree/sib-ship based model (BLUP) (**Tables 3**, **4**).

The prediction accuracies from the multi-trait model (MVBLUP and MVGLUP) were higher compared to the single-trait model (BLUP and GBLUP). Prediction accuracy in the pedigree/sib-ship based model (MVBLUP) ranged from 0.541 (DBH) to 0.754 (WD) in *E. nitens*, and from 0.553 (PME) to 0.679 (BR9) in *P. radiata*. In the marker-based model (MVGBLUP), this ranged from 0.529 (DBH) to 0.768 (WD) in *E. nitens*, and from 0.435 (ST9) to 0.618 (WD) in *P. radiata*. Generally, the implementation of multi-trait models (MVBLUP and MVGBLUP) followed a similar pattern as the single-trait model, in that the pedigree/sib-ship based model (MVBLUP) mostly outperformed the marker-based model (MVGBLUP), with a few exceptions, such as WD in *E. nitens* and DBH and WD in *P. radiata* (**Tables 3**, **4**).

Prediction accuracy of the models with markers selected using only positive loadings (MVGBLUP1) ranged from 0.434 (ST2) to 0.759 (WD) in *E. nitens* and from 0.446 (ST9) to 0.627 (WD) in *P. radiata*. For models with markers selected using both positive and negative loadings (MVGBLUP2), prediction accuracies ranged from 0.414 (ST2) to 0.766 (WD) in *E. nitens*, and from 0.436 (ST9) to 0.631 (WD) in *P. radiata*. The marker-based models using marker selection (MVGBUP1 and

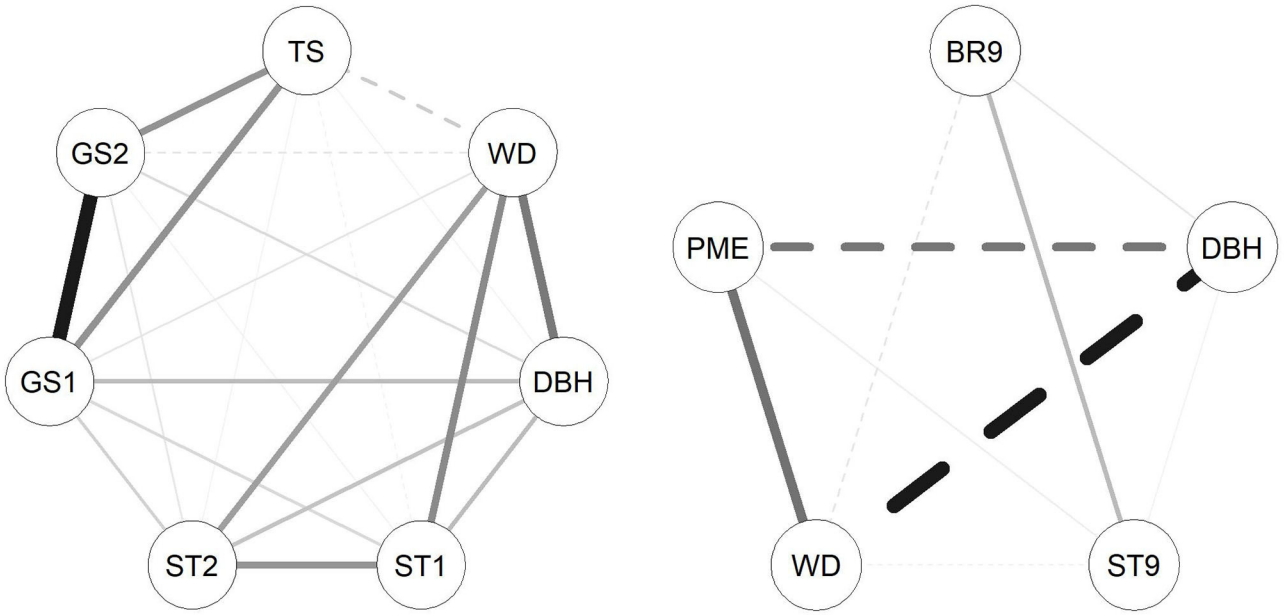


FIGURE 2 | Correlation networks between traits investigated in the *E. nitens* (left) and *P. radiata* (right) populations based on genetic correlations estimated in multivariate model using marker-based relationship matrix. Solid lines represent positive correlations and dashed lines represent negative genetic correlations; the thickness of the lines represents magnitude of correlations.

TABLE 2 | Number of markers selected in different scenarios using only positive (upper part) or both positive and negative (bottom part) marker loadings obtained from PLS-CA procedure.

	Species	*E. nitens*					*P. radiata*				
Scen	**Prop**	**P10**	**P20**	**P30**	**P40**	**P50**	**P10**	**P20**	**P30**	**P40**	**P50**
	C1	970	1,940	2,909	3,879	4,849	5,864	11,728	17,591	23,455	29,318
	C2	1,824	3,513	4,999	6,292	7,348	11,364	21,448	30,668	38,650	45,014
	C3	2,574	4,704	6,371	7,634	8,510	15,856	28,529	38,725	46,448	51,793
Pos	C4	3,318	5,776	7,456	8,555	9,180	20,773	35,762	45,963	52,179	55,740
	C5	3,898	6,492	8,049	8,992	9,419	24,128	39,697	49,288	54,523	57,198
	C6	4,515	7,188	8,631	9,312	9,567	NA	NA	NA	NA	NA
	C7	4,997	7,632	8,896	9,452	9,627	NA	NA	NA	NA	NA
	C1	1,904	3,840	5,825	7,792	9,659	10,574	22,848	35,438	47,511	58,636
	C2	3,377	6,131	8,103	9,282	9,697	19,871	37,712	49,848	56,706	58,636
Pos	C3	4,578	7,502	9,048	9,612	9,697	2,8337	46,493	55,238	58,271	58,636
+	C4	5,558	8,314	9,418	9,680	9,697	34,662	51,277	57,188	58,542	58,636
Neg	C5	6,303	8,801	9,562	9,694	9,697	40,064	54,329	58,047	58,618	58,636
	C6	6,946	9,108	9,639	9,696	9,697	NA	NA	NA	NA	NA
	C7	7,425	9,300	9,665	9,697	9,697	NA	NA	NA	NA	NA

NA represents the case not applicable for a particular species.

MVGBLUP2) resulted in increased prediction accuracy of the primary trait while maintaining similar accuracies for other traits in *E. nitens*. No impact of marker-selection on prediction accuracy of the primary trait was observed in *P. radiata* (**Tables 3**, **4**, **Figure 4**). The highest prediction accuracy for each trait was obtained using different marker selection scenarios, with no one scenario allowing for the highest prediction accuracy in all investigated traits simultaneously (**Supplementary Tables 2**–**5**). The significance of the improvement in prediction accuracy through marker selection was tested with the Wilcoxon non-parametric test, and a significant improvement was found only for DBH in *E. nitens* when the MVGBLUP2 model was implemented (**Table 3**).

The prediction accuracies estimated for each trait and marker selection scenario were correlated with DIC and number of selected markers. The correlations between prediction accuracy

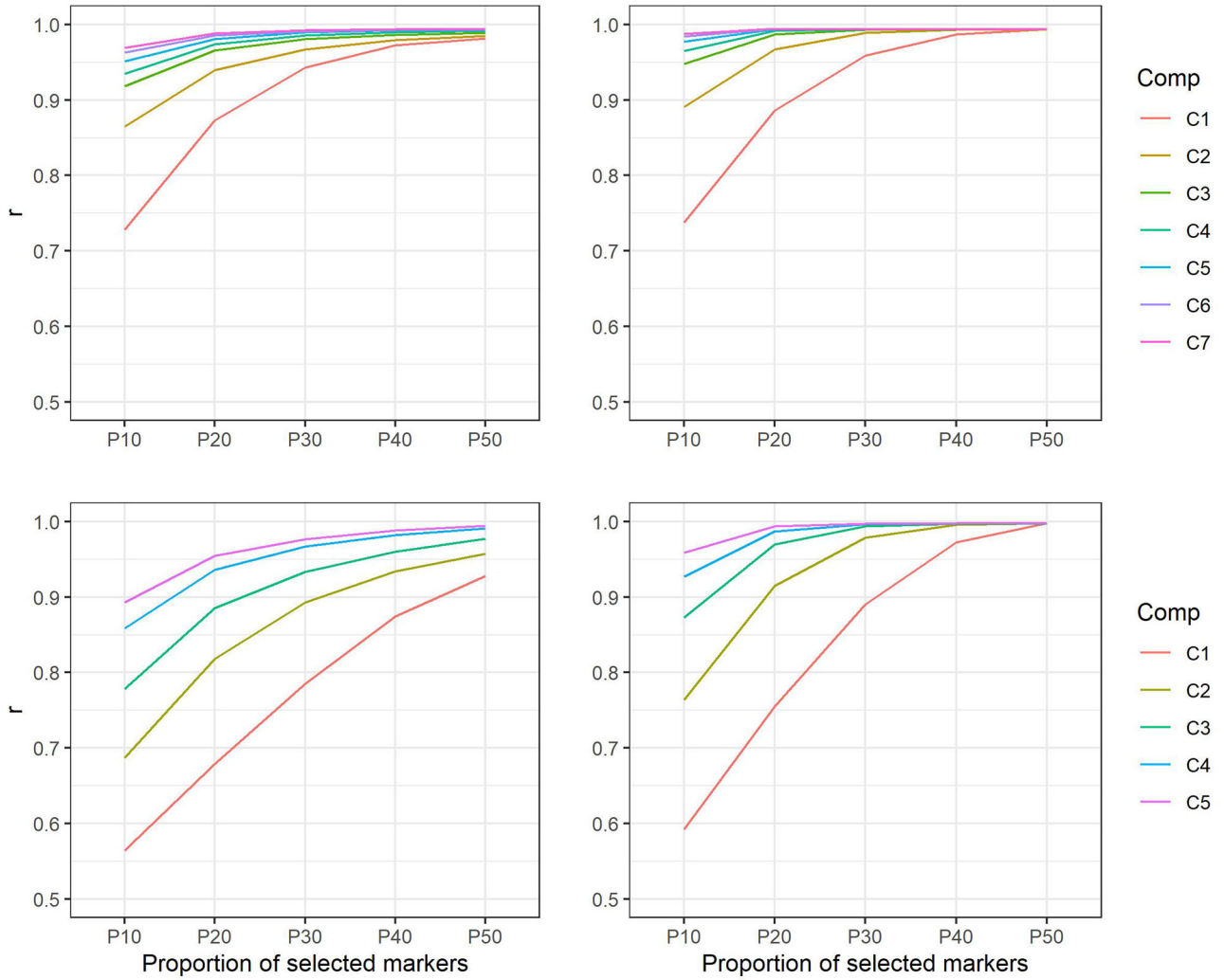


FIGURE 3 | Correlations between marker-based relationship matrices using markers selected on the basis of positive loadings only and marker-based relationship matrix using all markers in *E. nitens* (upper left) and in *P. radiata* (bottom left) populations and correlations between marker-based relationship matrices using markers selected on the basis of both positive and negative loadings and marker-based relationship matrix using all markers in *E. nitens* (upper right) and in *P. radiata* (bottom right) populations. Each line represent scenario for different number of latent variables considered in marker selection (e.g., C1—only the first latent variable is considered, C2—only the first two latent variables are considered, etc.).

and DIC were strong for *E. nitens*, reaching values from −0.952 (TS) to −0.559 (DBH) in scenarios where marker selection was based on positive marker loadings, and from −0.951 (TS) to −0.332 (WD) in scenarios where marker selection was based on both positive and negative marker loadings. The correlations between prediction accuracy and DIC were relatively weak in *P. radiata* reaching values from −0.721 (WD) to 0.115 (BR9) in scenarios where marker selection was based on positive marker loadings, and from −0.583 (DBH) to 0.623 (BR9) in scenarios where marker selection was based on both positive and negative marker loadings.

The correlations between prediction accuracy and number of selected markers were strong in *E. nitens*, reaching values from 0.467 (DBH) to 0.910 (ST1) in scenarios where marker selection was based on positive marker loadings and from 0.274 (WD) to 0.923 (TS) in scenarios where marker selection was based on both positive and negative marker loadings. Conversely, the correlations between prediction accuracy and number of selected markers were rather weak in *P. radiata* reaching values from −0.235 (BR9) to 0.841 (DBH) where marker selection was based on positive marker loadings and from −0.613 (BR9) to 0.439 (DBH) where marker selection was based on both positive and negative marker loadings. For our primary trait (DBH), in both species the opposite pattern was found between prediction accuracy and number of selected markers compared with prediction accuracy and DIC (**Table 5**).

TABLE 3 | Prediction accuracies and their standard deviations (in parenthesis) obtained from multivariate mixed models in the *E. nitens* population when using, a relationship matrix derived from sib-ship reconstruction (MVBLUP), a marker-based relationship matrix using all markers (MVGBLUP), a marker-based relationship matrix using selected SNPs having only positive loadings (MVGBLUP1), or a marker-based relationship matrix using selected SNPs having both positive and negative loadings (MVGBLUP2).

Trait	BLUP	GBLUP	MVBLUP	MVGBLUP	MVGBLUP1	MVGBLUP2
TS	0.737 (0.039)	0.656 (0.069)	0.754 (0.034)	0.665 (0.071)	0.650^{NS} (0.047)	0.642^{NS} (0.059)
WD	0.782 (0.060)	0.764 (0.054)	0.658 (0.068)	0.768 (0.049)	0.759^{NS} (0.053)	0.766^{NS} (0.035)
DBH	0.246 (0.132)	0.183 (0.117)	0.541 (0.251)	0.529 (0.336)	0.576^{NS} (0.241)	0.595^{**} (0.353)
ST1	0.613 (0.056)	0.523 (0.098)	0.621 (0.072)	0.545 (0.085)	0.525^{NS} (0.074)	0.523^{NS} (0.078)
ST2	0.571 (0.140)	0.448 (0.131)	0.582 (0.137)	0.442 (0.134)	0.434^{NS} (0.137)	0.414^{NS} (0.107)
GS1	0.683 (0.045)	0.558 (0.071)	0.720 (0.062)	0.609 (0.082)	0.604^{NS} (0.072)	0.604^{NS} (0.085)
GS2	0.603 (0.068)	0.547 (0.076)	0.737 (0.068)	0.651 (0.081)	0.650^{NS} (0.065)	0.660^{NS} (0.073)

Predicted EBV/GEBVs were correlated with EBVs estimated when using the multivariate mixed model using either documented pedigree or relationships inferred from sib-ship reconstruction.
***represents a statistically significant while NS represents a statistically non-significant test at α level 0.05.*

TABLE 4 | Prediction accuracies and their standard deviations (in parenthesis) obtained from multivariate mixed model in *P. radiata* population when using the documented pedigree (MVBLUP), a marker-based relationship matrix using all markers (MVGBLUP), a marker-based relationship matrix using selected SNPs having only positive loadings (MVGBLUP1), or a marker-based relationship matrix using selected SNPs having both positive and negative loadings (MVGBLUP2).

Trait	BLUP	GBLUP	MVBLUP	MVGBLUP	MVGBLUP1	MVGBLUP2
BR9	0.653 (0.088)	0.550 (0.121)	0.679 (0.095)	0.570 (0.136)	0.586^{NS} (0.134)	0.589^{NS} (0.123)
DBH	0.441 (0.103)	0.388 (0.133)	0.573 (0.069)	0.611 (0.062)	0.616^{NS} (0.058)	0.626^{NS} (0.048)
ST9	0.638 (0.147)	0.415 (0.148)	0.646 (0.126)	0.435 (0.135)	0.446^{NS} (0.149)	0.436^{NS} (0.119)
WD	0.642 (0.043)	0.645 (0.056)	0.610 (0.045)	0.618 (0.064)	0.627^{NS} (0.064)	0.631^{NS} (0.044)
PME	0.565 (0.118)	0.554 (0.119)	0.553 (0.109)	0.530 (0.116)	0.542^{NS} (0.113)	0.543^{NS} (0.108)

***represents a statistically significant while NS represents a statistically non-significant test at α level 0.05.*

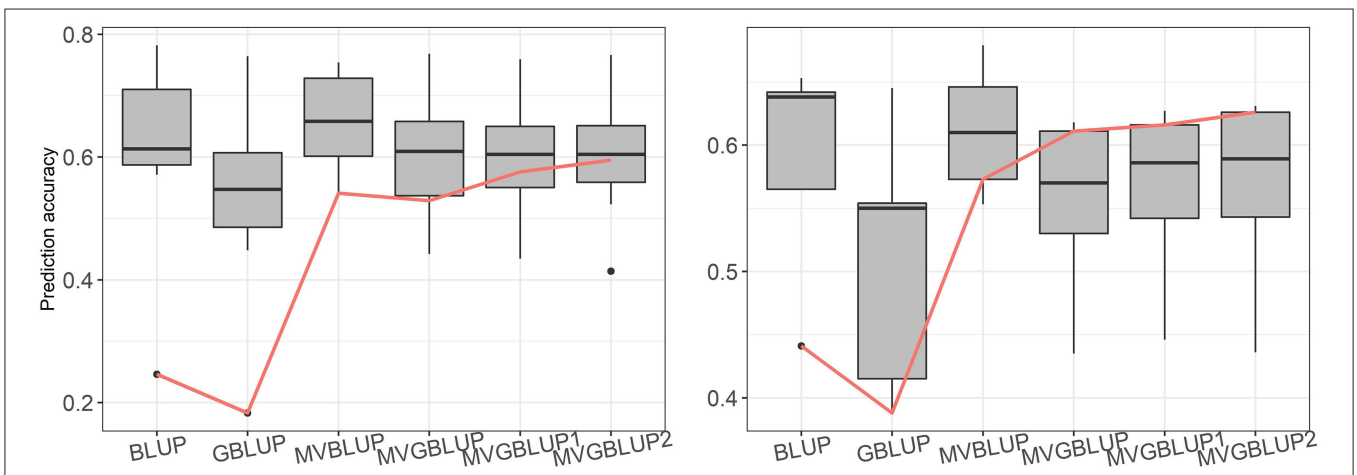


FIGURE 4 | Boxplot of prediction accuracies for each tested scenario in *E. nitens* (left plot) and in *P. radiata* (right plot) populations, red line represents prediction accuracy for primary trait (DBH).

4. DISCUSSION

4.1. Effect of Phenotypic Integration

Any complex trait is the end-product of many pathways, with many of the genes involved contributing to multiple pathways (i.e., pleiotropy). The efficient coordination of the pathways responsible for each particular attribute requires a certain level of organization in space and time, developed through modularity in the biological processes (Wagner et al., 2007). Therefore, pathways to achieving certain phenotypic characteristics can be structured into different modules comprising a number of different levels of shared pathways. The characteristics within each module show a high level of phenotypic integration while the characteristics from different modules show a low level of

TABLE 5 | Correlations between prediction accuracy and Deviance Information Criterion (DIC) and between prediction accuracy and number of selected markers.

Trait	*E. nitens*				*P. radiata*			
	Pos		Pos + Neg		Pos		Pos + Neg	
	DIC	$N_{Markers}$	DIC	$N_{Markers}$	DIC	$N_{Markers}$	DIC	$N_{Markers}$
TS	−0.952	0.849	−0.951	0.923	NA	NA	NA	NA
WD	−0.702	0.544	−0.332	0.274	−0.650	0.551	−0.557	0.364
DBH	−0.559	0.467	−0.409	0.358	−0.664	0.841	−0.583	0.439
ST1	−0.955	0.910	−0.902	0.855	NA	NA	NA	NA
ST2	−0.582	0.657	−0.455	0.504	NA	NA	NA	NA
GS1	−0.906	0.777	−0.756	0.701	NA	NA	NA	NA
GS2	−0.905	0.816	−0.635	0.600	NA	NA	NA	NA
ST9	NA	NA	NA	NA	−0.223	0.029	0.147	0.246
BR9	NA	NA	NA	NA	0.115	−0.235	0.623	−0.613
PME	NA	NA	NA	NA	−0.721	0.251	−0.194	0.082

NA represents the case where no data were available for particular species and trait.

integration (Wagner et al., 2007; Armbruster et al., 2014). Such stratification allows for effective independent evolution between modules, while the genetic correlations within the modules represent evolutionary constraints (Clark, 1987).

We proposed searching for markers that represent genomic regions involved in the shared pathways underlying the traits of interest. Our strategy for identifying such markers was through the alignment of the covariance structure within traits with the covariance structure within genetic markers, using a PLS-CA approach. This creates latent variables that collectively represent the studied attributes at each block (phenotypes on one side and genetic markers on the other side) through their shared variances (i.e., covariances). Since the method maximizes covariance between the latent variables from each block (phenotypes vs. genetic markers) through the coefficients in vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, it is possible to emphasize the shared variance caused by genetics (i.e., the part of the phenotypic covariance associated with genetic markers). Markers with strong associations to this alignment (large loadings) are likely positioned within the genomic regions showing pleiotropy or an accumulation of QTLs responsible for studied traits. Due to evolutionary trade-offs of gene functions on overall fitness, pleiotropy can act in opposite directions for affected traits (Guillaume and Otto, 2012). As a result, markers with negative association with the alignment (large negative loadings) are also likely to be involved in the underlying genetic architecture of covariances between traits. Watanabe et al. (2019) found that 90% of genes identified in human genome-wide association studies (GWAS) were associated with multiple traits, emphasizing how commonly pleiotropy plays a part in the genetic architecture of complex traits. However, where the complexity of genetic covariances between studied traits is unknown, a range of selection intensity in genetic markers is needed. We thus adopted a marker selection strategy based on quantiles derived from the distribution of their loadings.

Our analysis found there was a benefit to using marker selection (MVGBLUP1 and MVGBLUP2) in the multivariate analysis in the *E. nitens* population. Including more traits with no strong relationships (**Figure 1**—left plot) increased the prediction accuracy for DBH beyond that observed for the model using all available markers. On the other hand, using a multivariate model with marker selection (MVGBLUP1 and MVGBLUP2) in the *P. radiata* population did not improve prediction accuracy of low heritability DBH beyond that observed for the model using all available markers, possibly due to strong genetic correlation between DBH and WD (**Figure 1**—right plot). Since the precision of genetic correlations estimates depends on the strength and both size and structure of the sampled population (Bijma and Bastiaansen, 2014), the prediction accuracy of a low heritability trait with strong and well-estimated genetic correlations, as is the case of DBH and WD in *P. radiata* does not benefit from any additional marker selection. In contrast, the prediction accuracy of a low heritability trait with only moderate/lower and less precisely estimated genetic correlations, as in the *E. nitens* population, can benefit from the marker selection strategy proposed in this study (MVGBLUP2). Finding markers associated with the underlying genetic correlation structure can therefore potentially further improve the precision of genetic correlation estimates and thus the prediction accuracy of involved traits. However, it is worth noting that the scenarios showing the highest prediction accuracy for low heritability DBH were not supported by the model fit patterns (DIC) in either the *E. nitens* or *P. radiata* populations tested. Therefore, model fit is not a good indicator for selecting the best model in this case.

These findings indicate that the traits used in multivariate genomic analyses should, ideally, belong to the same variational module (set of traits that vary together and are independent of other traits) and show low to moderate genetic correlations in order to benefit from this approach. On the other hand, there is no further benefit of the proposed method when the estimated genetic correlation between the traits is high, such as the genetic correlation between WD and DBH in *P. radiata*. Traits in the same biological module usually show a high level of phenotypic integration, with pleiotropy likely contributing to this (Armbruster et al., 2014). Wagner et al. (2008) showed that most pleiotropic QTLs only affect a small number of traits and

their effect increases with the number of traits affected. Including a large number of traits that show different genetic correlations and precision levels to their estimates can increase the efficiency of this method (PLS-CA); pleiotropic QTLs are detected through weak or negative relationships between modules which increase the precision of genetic correlation estimates and thus accuracy in the prediction of low heritability traits as shown for *E. nitens*. It is worth mentioning, however, that pleiotropic QTLs can be present even when no marker-based genetic correlations are detected between traits (Gianola et al., 2015).

The efficient implementation of genomic selection in forestry requires the consideration of at least three groups of factors: (1) the genetic architecture of measured traits, (2) the structure of the training population, and (3) the quality of the phenotypic and marker data. A trait's genetic architecture is measured through factors, such as heritability, mode of inheritance (following Fisher's infinitesimal model vs. a mixed type of inheritance with a few large effect QTLs and many small effect ones) and the effective number of chromosomal fragments (Hayes et al., 2009), which depends on the distribution of QTLs across the genome and the intensity of LD decay.

The structure of the training population [the level of shared genealogy (relatedness), co-segregation and linkage disequilibrium between markers and QTLs (Habier et al., 2013)], will determine its suitability for genomic selection. The relative contribution of each of these to success depends on the composition of the training population itself. In our study, we tested two populations with different structures. While the *E. nitens* population shows two clusters due to contributions from two seed orchards with different selection strategies (Suontama et al., 2019), *P. radiata* shows no population structure but does show family clusters (**Supplementary Figure 1**). Additionally, while *E. nitens* included open-pollinated progenies with recovered full-sibs and self-sibs (Klápště et al., 2017), the *P. radiata* population contained full-sib families from 24 parents (**Supplementary Figure 1**). Genetic connectedness is vital, and good connections among parents, families or clones are important, as is the case in any quantitative analysis (Li et al., 2018). The production and testing of large full-sib families also gives the ability to dissect additive from non-additive genetic components and examine Mendelian segregation, something which is often confounded in pedigree-based analyses (Visscher et al., 2006). The size and decay rates of linkage disequilibrium between markers and QTLs, however, plays the most important role in training when mostly unrelated or only weakly related individuals are included (Meuwissen, 2009). Since the precise estimate of genetic correlations, the most critical genetic parameter considered for this approach, requires broad genetic diversity as well as familial structure in the training population (Bijma and Bastiaansen, 2014), the optimization of structure in training populations should be carefully considered.

Additionally, both populations represent advanced generations of breeding populations which underwent several generations of selection. Such conditions might introduce decreases in the accuracy of breeding values (in terms of correlation between true breeding values and estimated breeding values), depending on selection intensity and reduction in additive genetic variance (Bijma, 2012). The reduction is more pronounced in pedigree-based analyses compared to the marker-based counterpart due to the fact that the pedigree-based scenario can predict only parental averages (which explains only a small fraction of genetic variation and true breeding values of the offspring due to selection) compared to the marker-based equivalent; predicting both parental averages and Mendelian sampling (Gorjanc et al., 2015). However, the impact of selection on accuracy of breeding values depends on the data used in the analysis. While old data from previous generations pronounces the reduction in accuracy of breeding values, new data from the current selected population minimizes the impact of selection on the accuracy of breeding values (Bijma, 2012).

4.2. Genomic Data Quality, Quantity, and Selection

The quality of marker data impacts directly on the ability of these markers to capture and adequately describe the genetic control and architecture of quantitative traits. The usefulness of a genomic resource is therefore a function of the number of markers, their distribution across the genome and the accuracy of the genotype calls. The platforms available for genotyping forest tree species are often driven by the nature of their genomes. In this study, the relatively small genome length of many *Eucalyptus* species (∼0.56 Gb) has allowed the rapid and cost-effective development of the multi-species *Eucalyptus* SNP chip, based on SNP discovery from whole genome sequencing data (Silva-Junior et al., 2015). In contrast, the extensive size of the *Pinus radiata* genome (∼25 Gb) and large amount of repetitive sequences required a different SNP discovery and genotyping approach based on reduced representation sequencing of the genome (Elshire et al., 2011; Neves et al., 2013; Telfer et al., 2018). Such approaches, or other similar techniques, such as exome capture have already been successfully implemented in other conifer species (Gamal El-Dien et al., 2015; Ratcliffe et al., 2015; Bartholomé et al., 2016; Isik et al., 2016; Lenz et al., 2017; Chen et al., 2018).

The large amount of genomic data obtained in genomic selection studies can contain some level of redundancy, which can negatively affect the accuracy of breeding values (Habier et al., 2013) and might necessitate variable selection approaches. Ballesta et al. (2018) found an advantage to dimensionality reduction and variable selection, improving prediction accuracy of low-to-moderate heritability traits in a single-trait evaluation in a *Eucalyptus globulus* population. Our strategy resulted in the highest prediction accuracy for the primary trait when ∼ 66% (considering only positive loadings) and ∼ 35% (considering both positive and negative loadings) of markers were included in the marker-based relationship matrix in *E. nitens*, and ∼94% and 99% markers in *P. radiata* (**Table 4**, **Supplementary Tables 1**, **4**). Lippert et al. (2013) found that the pre-selection of QTL-related markers, or at least increasing the proportion of such markers over uninformative ones was an advantage and increased the accuracy of predicted genomic breeding values. Several other approaches have been examined, using marker weights developed using either Bayesian inference (Kemper et al., 2018) or results from previous QTL mapping or association studies (Fragomeni et al., 2017). The proportion of markers selected reflects the genetic complexity of the trait under study. For example, Müller

et al. (2017) found that 5,000–10,000 markers (representing ∼40–60% of full marker data) were sufficient to capture the major proportion of trait heritability and reach the same prediction accuracy compared to using the all marker dataset. Similarly, Resende et al. (2012) found an advantage to using reduced numbers of markers in traits, such as wood specific gravity (∼5% of total marker data) and resistance to Fusiform rust [gall volume (∼2% of total marker data) and presence or absence of rust (∼7% of the total marker data)]. Additionally, Chen et al. (2018) found that the structure of the training population (full-sib vs. half-sib families) defines the number of selected markers needed to reach prediction accuracies equivalent to using full marker data. While for a full-sib structure 4,000–8,000 markers was found to be sufficient, a half-sib structure required all 100,000 markers to reach the maximum achievable prediction accuracy. However, the selection of informative markers was performed using only single trait approaches and different standards of genomic resources.

Similar to our approach, several proposed strategies have been developed within a Bayesian multivariate framework (Cheng et al., 2018; Karaman et al., 2018). Karaman et al. (2018) found there was benefit to assigning specific weight to blocks of fixed numbers of markers rather than to each marker individually. Our approach allows for the selection of markers associated with genomic regions related to shared underlying genetic components across investigated traits, without any prior definition of the block length. Since our approach associates the markers with underlying structure rather than with each trait involved in the study, it shows benefits even in the case of sparse marker arrays as used in this study. However, the presence of full phenotypic data is required to perform marker selection through PLS-CA, and thus the investigated traits have to be screened at an operational scale.

The strategy proposed in this study does not attempt to improve the accuracy of all traits involved in the analysis but only those with low heritabilities, taking advantage of the genetic covariances common across all investigated traits. The latent variables created through PLS-CA analysis (Tenenhaus, 1998; Sanchez and Sanchez, 2012) tend to extract the common part of variances in both the trait and marker data by maximization of covariance between latent variables. Ideally, the algorithm searches for bridges between variational modules (group of traits that vary together) and functional modules (group of genes/proteins that are coordinated to perform semi-autonomous functions) (Kliebenstein, 2011). However, the efficiency of finding such bridges depends on adequate representation of the genome through marker data. The investigation of marker loadings associated with the latent variables can identify those markers important for explaining the variance captured by each latent variable. Additionally, this investigation will also indirectly identify the markers which most likely explain variance explaining the behavior of the corresponding latent variable derived from phenotypic data. As mentioned above, the efficiency of the proposed strategy depends on the level of integration and modularity between the traits under study. Therefore, the selection of traits included in the analysis should take into consideration their biological connection and their heritabilities.

In general, the magnitude of genetic correlations between traits has an impact on the accuracy of breeding values (Jia and Jannink, 2012). However, the method proposed in this study benefited from improvement in the precision of genetic correlation structure through marker selection only when pairwise correlations were low or moderate. In contrast, no additional benefit beyond the commonly used model (MVGBLUP) was found in a population with well-estimated strong genetic correlation between primary (DBH) and other (WD) traits. Pleiotropic QTLs, however, can be included in the underlying genetic structures used in the analysis, even where no genetic correlations between traits are detected using genetic markers. Therefore, marker-based genetic correlations can be misleading to provide inference about their causes when knowledge about LD between markers and QTLs is poor or non-existing (Gianola et al., 2015).

5. CONCLUSIONS

The approach proposed in this study selects markers aligned to the underlying dimensions extracted from a trait's covariance structure rather than investigating associations between markers with each trait, which allows for improvements even with sparse marker arrays. This method is suitable for improving the accuracy of low heritability traits where genetic correlations between traits are low/moderate in magnitude and low accuracy. In contrast, when the population shows a strong genetic correlation between the primary trait (DBH in this study) and other moderately heritable traits, this approach does not show benefit beyond that observed with the multivariate model using all genetic markers. One drawback is that this approach requires all individuals in the training population to be phenotyped for all traits included in the analysis to perform the marker selection procedure (PLS-CA).

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the ZENODO data repository: doi: 10.5281/zenodo.4040042.

AUTHOR CONTRIBUTIONS

JK performed the analyses and drafted the manuscript. MS, HD, ET, NG, and YL designed the study, assisted with drafting the manuscript, and secured the funding. RM developed the phenotyping protocols for wood quality attributes. All authors significantly contributed to the current study.

FUNDING

The study was funded by Radiata Pine Breeding Company Ltd. and NZ Ministry of Business, Innovation and Employment (MBIE) joint project RPBC1301, Specialty Wood Products Research Partnership Program (SWP) Contract No. C04X1104 and MBIE Strategic Science Investment Fund Contract No. C04X1703.

ACKNOWLEDGMENTS

We would like to thank the Radiata Pine Breeding Company Ltd. for access to field experiments and data collection and Lucy Macdonald for bioinformatics support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.499094/full#supplementary-material

REFERENCES

Armbruster, W. S., Pélabon, C., Bolstad, G. H., and Hansen, T. F. (2014). Integrated phenotypes: understanding trait covariation in plants and animals. *Philos. Trans. R. Soc. B* 369:20130245. doi: 10.1098/rstb.2013.0245

Ballesta, P., Serra, N., Guerra, F. P., Hasbún, R., and Mora, F. (2018). Genomic prediction of growth and stem quality traits in *Eucalyptus globulus* labill. at its southernmost distribution limit in Chile. *Forests* 9:779. doi: 10.3390/f9120779

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., et al. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics* 17:604. doi: 10.1186/s12864-016-2879-8

Berger, S., Pérez-Rodríguez, P., Veturi, Y., Simianer, H., and de los Campos, G. (2015). Effectiveness of shrinkage and variable selection methods for the prediction of complex human traits using data from distantly related individuals. *Ann. Hum. Genet.* 79, 122–135. doi: 10.1111/ahg.12099

Bijma, P. (2012). Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.* 129, 345–358. doi: 10.1111/j.1439-0388.2012.00991.x

Bijma, P., and Bastiaansen, J. W. (2014). Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? *Genet. Sel. Evol.* 46, 1–6. doi: 10.1186/s12711-014-0079-z

Blows, M. W., and Hoffmann, A. A. (2005). A reassessment of genetic limits to evolutionary change. *Ecology* 86, 1371–1384. doi: 10.1890/04-1209

Bouvet, J.-M., Vigneron, P., Gouma, R., and Saya, A. (2003). Trends in variances and heritabilities with age for growth traits in Eucalyptus spacing experiments. *Silv. Genet.* 52, 121–132. Available online at: http://agritrop.cirad.fr/519509/

Butler, D., Cullis, B. R., Gilmour, A., and Gogel, B. (2009). *ASReml-R Reference Manual*. Brisbane, QLD: The State of Queensland, Department of Primary Industries and Fisheries.

Calus, M. P., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43:26. doi: 10.1186/1297-9686-43-26

Carson, M. J. (1986). *Control-Pollinated Seed Orchards of Best General Combiners: A New Strategy for Radiata Pine Improvement*. Rotorua: New Zealand Forest Service.

Carson, S. (1991). Genotype x environment interaction and optimal number of progeny test sites for improving *Pinus radiata* in New Zealand. *N. Zeal. J. Forest Sci.* 21, 32–49.

Carter, P., Chauhan, S., and Walker, J. (2007). Sorting logs and lumber for stiffness using director HM200. *Wood Fiber Sci.* 38, 49–54. Available online at: https://wfs.swst.org/index.php/wfs/article/view/1650/1650

Chandler, C. H., Chari, S., Kowalski, A., Choi, L., Tack, D., DeNieu, M., et al. (2017). How well do you know your mutation? Complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. *PLoS Genet.* 13, e1007075. doi: 10.1371/journal.pgen.1007075

Chen, Z.-Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., et al. (2018). Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in norway spruce. *BMC Genomics* 19:946. doi: 10.1186/s12864-018-5256-y

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., and Fernando, R. (2018). Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. *Genetics* 209, 89–103. doi: 10.1534/genetics.118.300650

Clark, A. G. (1987). "Genetic correlations: the quantitative genetics of evolutionary constraints," in *Genetic Constraints on Adaptive Evolution*, ed V. Loeschcke (Heidelberg: Springer), 25–45. doi: 10.1007/978-3-642-72770-2_3

Cole, J., VanRaden, P., O'Connell, J., Van Tassell, C., Sonstegard, T., Schnabel, R., et al. (2009). Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.* 92, 2931–2946. doi: 10.3168/jds.2008-1762

de los Campos, G., and Grüneberg, A. (2016). *MTM (Multiple-Trait Model) Package*. Available online at: http://quantgen.github.io/MTM/vignette.html (accessed May 10, 2018).

de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? *PLoS Genet.* 11:e1005048. doi: 10.1371/journal.pgen.1005048

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Falconer, D., and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Essex: Longman Group.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Earth Env. Sci. Trans. R. Soc.* 52, 399–433. doi: 10.1017/S0080456800012163

Fragomeni, B. O., Lourenco, D. A., Masuda, Y., Legarra, A., and Misztal, I. (2017). Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Sel. Evol.* 49:59. doi: 10.1186/s12711-017-0335-0

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16:370. doi: 10.1186/s12864-015-1597-y

Gianola, D., de los Campos, G., Toro, M. A., Naya, H., Schön, C.-C., and Sorensen, D. (2015). Do molecular markers inform about pleiotropy? *Genetics* 201, 23–29. doi: 10.1534/genetics.115.179978

Gorjanc, G., Bijma, P., and Hickey, J. M. (2015). Reliability of pedigree-based and genomic evaluations in selected populations. *Genet. Sel. Evol.* 47:65. doi: 10.1186/s12711-015-0145-1

Grattapaglia, D., and Resende, M. D. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi: 10.1007/s11295-010-0328-4

Grattapaglia, D., Silva-Junior, O. B., Resende, R. T., Cappa, E. P., Müller, B. S., Tan, B., et al. (2018). Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front. Plant Sci.* 9:1693. doi: 10.3389/fpls.2018.01693

Guillaume, F., and Otto, S. P. (2012). Gene functional trade-offs and the evolution of pleiotropy. *Genetics* 192, 1389–1409. doi: 10.1534/genetics.112.143214

Guo, G., Lund, M. S., Zhang, Y., and Su, G. (2010). Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. *J. Anim. Breed. Genet.* 127, 423–432. doi: 10.1111/j.1439-0388.2010.00878.x

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15:30. doi: 10.1186/1471-2156-15-30

Habier, D., Fernando, R., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic-BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194, 597–607. doi: 10.1534/genetics.113.152207

Hayatgheibi, H., Fries, A., Kroon, J., and Wu, H. X. (2019). Estimation of genetic parameters, provenance performances, and genotype by environment interactions for growth and stiffness in lodgepole pine (*Pinus contorta*). *Scand. J. Forest Res.* 34, 1–11. doi: 10.1080/02827581.2018.1542025

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/S0016672308009981

Isik, F. (2014). Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forest.* 45, 379–401. doi: 10.1007/s11056-014-9422-z

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., et al. (2016). Genomic selection in maritime pine. *Plant Sci*. 242, 108–119. doi: 10.1016/j.plantsci.2015.08.006

Jia, Y., and Jannink, J.-L. (2012). Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192, 1513–1522. doi: 10.1534/genetics.112.144246

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 11:94. doi: 10.1186/1471-2156-11-94

Karaman, E., Lund, M. S., Anche, M. T., Janss, L., and Su, G. (2018). Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. *Genes Genom. Genet.* 8, 3549–3558. doi: 10.1534/g3.118.200673

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genet. Sel. Evol*. 50:10. doi: 10.1186/s12711-018-0377-y

King, D. A. (1990). The adaptive significance of tree height. *Am. Nat*. 135, 809–828. doi: 10.1086/285075

Klápště, J., Suontama, M., Dungey, H. S., Telfer, E. J., and Stovold, G. T. (2019). Modelling of population structure through contemporary groups in genetic evaluation. *BMC Genet*. 20:81. doi: 10.1186/s12863-019-0778-0

Klápště, J., Suontama, M., Telfer, E. J., Graham, N. J., Low, C., Stovold, T., et al. (2017). Exploration of genetic architecture through sib-ship reconstruction in advanced breeding population of *Eucalyptus nitens*. *PLoS ONE* 12:e0185137. doi: 10.1371/journal.pone.0185137

Kliebenstein, D. (2011). Genetic and functional modularity: how does an organism solve a nearly infinite genetic/environmental problem space? *Heredity* 106:909. doi: 10.1038/hdy.2010.136

Latta, R. G. (2010). Natural selection, variation, adaptation, and evolution: a primer of interrelated concepts. *Int. J. Plant Sci*. 171, 930–944. doi: 10.1086/656220

Lenz, P. R., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., and Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics* 18:335. doi: 10.1186/s12864-017-3715-5

Li, Y., Dungey, H. S., Carson, M., and Carson, S. (2018). Genotype by environment interaction for growth and Dothistroma resistance and clonal connectivity between environments in radiata pine in New Zealand and Australia. *PLoS ONE* 13:e0205402. doi: 10.1371/journal.pone.0205402

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., and Heckerman, D. (2013). The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep*. 3:1815. doi: 10.1038/srep01815

Marchal, A., Legarra, A., Tisné, S., Carasco-Lacombe, C., Manez, A., Suryana, E., et al. (2016). Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Mol. Breed.* 36, 1–13. doi: 10.1007/s11032-015-0423-1

Merilä, J., and Sheldon, B. (2000). Lifetime reproductive success and heritability in nature. *Am. Nat*. 155, 301–310. doi: 10.1086/303330

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. Available online at: https://www.genetics.org/content/157/4/1819

Meuwissen, T. H. (2009). Accuracy of breeding values of'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol*. 41:35. doi: 10.1186/1297-9686-41-35

Mihai, G., and Mirancea, I. (2016). Age trends in genetic parameters for growth and quality traits in *Abies alba*. *iForest* 9:954. doi: 10.3832/ifor1766-009

Müller, B. S., Neves, L. G., de Almeida Filho, J. E., Resende, M. F., Muñoz, P. R., dos Santos, P. E., et al. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of eucalyptus. *BMC Genomics* 18:524. doi: 10.1186/s12864-017-3920-2

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of *Eucalyptus grandis*. *Nature* 510:356. doi: 10.1038/nature13308

Neale, D. B., and Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nat. Rev. Genet*. 12, 111–122. doi: 10.1038/nrg2931

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., et al. (2014). Decoding the massive genome of loblolly pine using haploid dna and novel assembly strategies. *Genome Biol*. 15:R59. doi: 10.1186/gb-2014-15-3-r59

Nejati-Javaremi, A., Smith, C., and Gibson, J. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci*. 75, 1738–1745. doi: 10.2527/1997.7571738x

Neves, L. G., Davis, J. M., Barbazuk, W. B., and Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *Plant J*. 75, 146–156. doi: 10.1111/tpj.12193

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., et al. (2013). The norway spruce genome sequence and conifer genome evolution. *Nature* 497:579. doi: 10.1038/nature12211

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

Pollak, E., Van der Werf, J., and Quaas, R. (1984). Selection bias and multiple trait evaluation. *J. Dairy Sci*. 67, 1590–1595. doi: 10.3168/jds.S0022-0302(84)81481-2

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet*. 11, 800–805. doi: 10.1038/nrg2865

Ratcliffe, B., Gamal El-Dien, O., Klápště, J., Porth, I., Chen, C., Jaquish, B., et al. (2015). A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × glauca) using unordered SNP imputation methods. *Heredity* 115, 547–555. doi: 10.1038/hdy.2015.57

Resende, M. F., Muñoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012). Accuracy of genomic selection methods in a standard dataset of loblolly pine (*Pinus taeda* L.). *Genetics* 190, 1503–1510. doi: 10.1534/genetics.111.137026

Sanchez, G., and Sanchez, M. G. (2012). *Package 'plsdepot'. Partial Least Squares (PLS) Data Analysis Methods, v. 0.1*, 17.

Scutari, M., Mackay, I., and Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet*. 12:e1006288. doi: 10.1371/journal.pgen.1006288

Sedlacek, J., Cortés, A. J., Wheeler, J., Bossdorf, O., Hoch, G., Klápště, J., et al. (2016). Evolutionary potential in the Alpine: trait heritabilities and performance variation of the dwarf willow *Salix herbacea* from different elevations and microhabitats. *Ecol. Evol*. 6, 3940–3952. doi: 10.1002/ece3.2171

Silva-Junior, O. B., Faria, D. A., and Grattapaglia, D. (2015). A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytol*. 206, 1527–1540. doi: 10.1111/nph.13322

Smith, D. M. (1954). *Maximum Moisture Content Method for Determining Specific Gravity of Small Wood Samples*. USDA Report, 2014. USDA Madison: University of Wisconsin

Stejskal, J., Lstibůrek, M., Klápště, J., Čepl, J., and El-Kassaby, Y. (2018). Effect of genomic prediction on response to selection in forest tree breeding. *Tree Genet. Genomes* 14:74. doi: 10.1007/s11295-018-1283-8

Su, G., Christensen, O. F., Janss, L., and Lund, M. S. (2014). Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *J. Dairy Sci*. 97, 6547–6559. doi: 10.3168/jds.2014-8210

Sun, X., Fernando, R., and Dekkers, J. (2016). Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genet. Sel. Evol*. 48:77. doi: 10.1186/s12711-016-0255-4

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., et al. (2019). Efficiency of genomic prediction across two *Eucalyptus nitens* seed orchards with different selection histories. *Heredity* 122, 370–379. doi: 10.1038/s41437-018-0119-5

Telfer, E., Graham, N., Macdonald, L., Sturrock, S., Wilcox, P., and Stanbra, L. (2018). Approaches to variant discovery for conifer transcriptome sequencing. *PLoS ONE* 13:e0205835. doi: 10.1371/journal.pone.0205835

Telfer, E. J., Graham, N. J., Klápště, J., Li, Y., Resende, M. Jr., Neves, L. G., et al. (2019). A high-density exome capture genotyping-by-sequencing panel for forestry breeding *Pinus radiata*. *PLoS ONE* 14:e0222640. doi: 10.1371/journal.pone.0222640

Tenenhaus, M. (1998). *La régression PLS: Théorie et Pratique*. Paris: Editions Technip.

Treloar, C., and Lausberg, M. (1997). *Sampling and Data Handling Techniques for Wood Quality Analyses, Volume 201 of FRI Bulletin*. Rotorua: New Zealand Forest Research Institute.

Ukrainetz, N. K., Kang, K.-Y., Aitken, S. N., Stoehr, M., and Mansfield, S. D. (2008). Heritability and phenotypic and genetic correlations of coastal douglas-fir (*Pseudotsuga menziesii*) wood quality traits. *Can. J. Forest Res.* 38, 1536–1546. doi: 10.1139/X07-234

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

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., et al. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*. 2:e41. doi: 10.1371/journal.pgen.0020041

Wagner, G. P., Kenney-Hunt, J. P., Pavlicev, M., Peck, J. R., Waxman, D., and Cheverud, J. M. (2008). Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* 452, 470–472. doi: 10.1038/nature 06756

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet*. 8, 921–931. doi: 10.1038/nrg 2267

Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., et al. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet*. 51, 1339–1348. doi: 10.1038/s41588-019-0481-0

Wilcoxon, F. (1992). "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*, eds S. Kotz and N.L. Johnson (New York, NY: Springer), 196–202. doi: 10.1007/978-1-4612-4380-9_16

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat*. 56, 330–338. doi: 10.1086/279872

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648. doi: 10.1371/journal.pone.0012648

Conflict of Interest: The authors declare that this study received funding from NZ Ministry of Business, Innovation and Employment and Radiata Pine Breeding Company Ltd. The Radiata Pine Breeding Company Ltd. was involved in phenotypic data collection. The funders were not involved in the study design, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

JK was employed by the company Scion.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Klápště, Dungey, Telfer, Suontama, Graham, Li and McKinley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

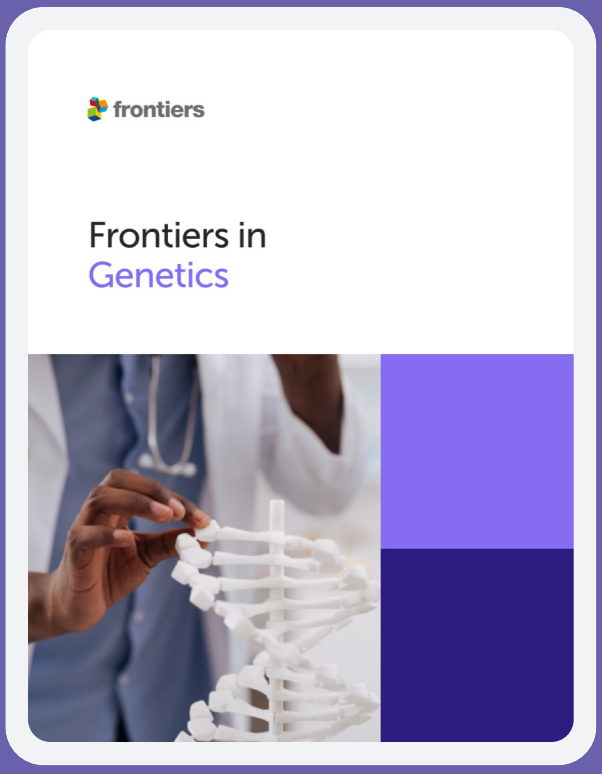
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact