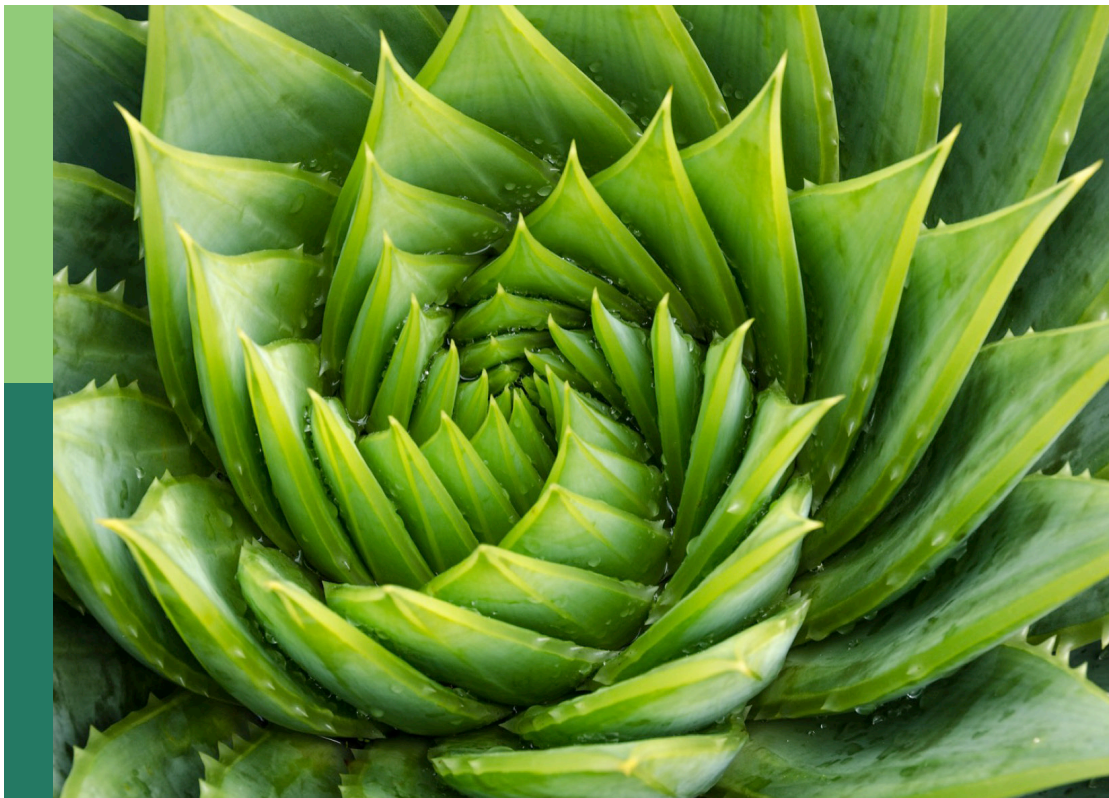
frontiers | Research Topics

Breeding forage and grain legumes for sustainable protein production

Edited by
Åshild Ergon, Leif Skot, Deyue Yu and Thomas Wood

Published in
Frontiers in Plant Science



EUCLEG

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 978-2-8325-6294-9
DOI 10.3389/978-2-8325-6294-9

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Breeding forage and grain legumes for sustainable protein production

Topic editors
Åshild Ergon — Norwegian University of Life Sciences, Norway
Leif Skot — Aberystwyth University, United Kingdom
Deyue Yu — Nanjing Agricultural University, China
Thomas Wood — National Institute of Agricultural Botany (NIAB), United Kingdom

Citation
Ergon, Å., Skot, L., Yu, D., Wood, T., eds. (2025). *Breeding forage and grain legumes for sustainable protein production*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6294-9

The EUCLEG project (Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency) was funded by Horizon 2020 - Research and Innovation Action (EU Project No. 727312).

Table of contents

05 Genome-Wide Analysis of the Glucose-6-Phosphate Dehydrogenase Family in Soybean and Functional Identification of *GmG6PDH2* Involvement in Salt Stress
Ying Zhao, Yifan Cui, Shiyu Huang, Jingyao Yu, Xinyu Wang, Dawei Xin, Xin Li, Yonghui Liu, Yuxin Dai, Zhaoming Qi and Qingshan Chen

21 QTL Mapping and Data Mining to Identify Genes Associated With the *Sinorhizobium fredii* HH103 T3SS Effector NopD in Soybean
Jinhui Wang, Jieqi Wang, Chao Ma, Ziqi Zhou, Decheng Yang, Junzan Zheng, Qi Wang, Huiwen Li, Hongyang Zhou, Zhijun Sun, Hanxi Liu, Jianyi Li, Lin Chen, Qinglin Kang, Zhaoming Qi, Hongwei Jiang, Rongsheng Zhu, Xiaoxia Wu, Chunyan Liu, Qingshan Chen and Dawei Xin

37 A Non-destructive Method to Quantify Leaf Starch Content in Red Clover
Lea Antonia Frey, Philipp Baumann, Helge Aasen, Bruno Studer and Roland Kölliker

48 The History of Domestication and Selection of Lucerne: A New Perspective From the Genetic Diversity for Seed Germination in Response to Temperature and Scarification
Wagdi Ghaleb, Lina Qadir Ahmed, Abraham J. Escobar-Gutiérrez and Bernadette Julier

59 A Genome-Wide Genetic Diversity Scan Reveals Multiple Signatures of Selection in a European Soybean Collection Compared to Chinese Collections of Wild and Cultivated Soybean Accessions
Aamir Saleem, Hilde Muylle, Jonas Aper, Tom Ruttink, Jiao Wang, Deyue Yu and Isabel Roldán-Ruiz

74 Response of a Diverse European Soybean Collection to "Short Duration" and "Long Duration" Drought Stress
Aamir Saleem, Jonas Aper, Hilde Muylle, Irene Borra-Serrano, Paul Quataert, Peter Lootens, Tom De Swaef and Isabel Roldán-Ruiz

93 Genome-wide association analysis for drought tolerance and associated traits in faba bean (*Vicia faba* L.)
Natalia Gutiérrez, Marie Pégard, Christiane Balko and Ana M. Torres

110 Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover
Michelle M. Nay, Christoph Grieder, Lea A. Frey, Helga Amdahl, Jasmina Radovic, Libor Jaluvka, Anna Palmé, Leif Skøt, Tom Ruttink and Roland Kölliker

127 A genome-wide association study of freezing tolerance in red clover (*Trifolium pratense* L.) germplasm of European origin
Stefano Zanotto, Tom Ruttink, Marie Pégard, Leif Skøt, Christoph Grieder, Roland Kölliker and Åshild Ergon

139 Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits
Marie Pégard, Philippe Barre, Sabrina Delaunay, Fabien Surault, Djura Karagić, Dragan Milić, Miroslav Zorić, Tom Ruttink and Bernadette Julier

157 Variability in drought stress response in a panel of 100 faba bean genotypes
Christiane Balko, Ana M. Torres and Natalia Gutierrez

172 Phenotypic characterization of drought responses in red clover (*Trifolium pratense* L.)
Tim Vleugels, Aamir Saleem, Reena Dubey, Hilde Muylle, Irene Borra-Serrano, Peter Lootens, Tom De Swaef and Isabel Roldán-Ruiz

188 Genome-wide association study for yield-related traits in faba bean (*Vicia faba* L.)
Natalia Gutierrez, Marie Pégard, Ignacio Solis, Dejan Sokolovic, David Lloyd, Catherine Howarth and Ana M. Torres

213 Including marker x environment interactions improves genomic prediction in red clover (*Trifolium pratense* L.)
Leif Skøt, Michelle M. Nay, Christoph Grieder, Lea A. Frey, Marie Pégard, Linda Öhlund, Helga Amdahl, Jasmina Radovic, Libor Jaluvka, Anna Palmé, Tom Ruttink, David Lloyd, Catherine J. Howarth and Roland Kölliker

231 Phenotypic characterization of soybean genetic resources at multiple locations: breeding implications for enhancing environmental resilience, yield and protein content
Tomáš Vymyslický, Oldřich Trněný, Hendrik Rietman, Christiane Balko, Vuk Đorđević, Predrag Ranđelović and Magdaléna Dybová

frontiers in Plant Science

ORIGINAL RESEARCH
published: 26 February 2020
doi: 10.3389/fpls.2020.00214

Genome-Wide Analysis of the Glucose-6-Phosphate Dehydrogenase Family in Soybean and Functional Identification of *GmG6PDH2* Involvement in Salt Stress

Ying Zhao[1†], Yifan Cui[1†], Shiyu Huang[1], Jingyao Yu[1], Xinyu Wang[1], Dawei Xin[1], Xin Li[2], Yonghui Liu[1], Yuxin Dai[1], Zhaoming Qi[1] and Qingshan Chen[1]**

[1] College of Agriculture, Northeast Agricultural University, Harbin, China, [2] Key Lab of Maize Genetics and Breeding, Heilongjiang Academy of Agricultural Sciences, Harbin, China

OPEN ACCESS

Edited by:
Thomas Henry Noel Ellis,
John Innes Centre, United Kingdom

Reviewed by:
Alberto A. Iglesias,
National University of the Littoral, Argentina
Vivekanand Tiwari,
Agricultural Research Organization (ARO), Israel

***Correspondence:**
Zhaoming Qi
qizhaoming1860@126.com
Qingshan Chen
qshchen@126.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 22 November 2019
Accepted: 12 February 2020
Published: 26 February 2020

Citation:
Zhao Y, Cui Y, Huang S, Yu J, Wang X, Xin D, Li X, Liu Y, Dai Y, Qi Z and Chen Q (2020) Genome-Wide Analysis of the Glucose-6-Phosphate Dehydrogenase Family in Soybean and Functional Identification of GmG6PDH2 Involvement in Salt Stress. Front. Plant Sci. 11:214. doi: 10.3389/fpls.2020.00214

Glucose-6-phosphate dehydrogenase (G6PDH) is known as a critical enzyme responsible for nicotinamide adenine dinucleotide phosphate (NADPH) generation in the pentose phosphate pathway (PPP), and has an essential function in modulating redox homeostasis and stress responsiveness. In the present work, we characterized the nine members of the G6PDH gene family in soybean. Phylogenic analysis and transit peptide prediction showed that these soybean G6PDHs are divided into plastidic (P) and cytosolic (Cy) isoforms. The subcellular locations of five GmG6PDHs were further verified by confocal microscopy in *Arabidopsis* mesophyll protoplasts. The respective *GmG6PDH* genes had distinct expression patterns in various soybean tissues and at different times during seed development. Among them, the *Cy-G6PDHs* were strongly expressed in roots, developing seeds and nodules, while the transcripts of *P-G6PDHs* were mainly detected in green tissues. In addition, the activities and transcripts of *GmG6PDHs* were dramatically stimulated by different stress treatments, including salt, osmotic and alkali. Notably, the expression levels of a cytosolic isoform (GmG6PDH2) were extraordinarily high under salt stress and correlated well with the G6PDH enzyme activities, possibly implying a crucial factor for soybean responses to salinity. Enzymatic assay of recombinant *GmG6PDH2* proteins expressed in *Escherichia coli* showed that the enzyme encoded by *GmG6PDH2* had functional $NADP^+$-dependent G6PDH activity. Further analysis indicated overexpression of *GmG6PDH2* gene could significantly enhance the resistance of transgenic soybean to salt stress by coordinating with the redox states of ascorbic acid and glutathione pool to suppress reactive oxygen species generation. Together, these results indicate that *GmG6PDH2* might be the major isoform for NADPH production in PPP, which is involved in the modulation of cellular AsA-GSH cycle to prevent the oxidative damage induced by high salinity.

Keywords: glucose-6-phosphate dehydrogenase, expression, enzyme activity, salt stress, transgenic plants, soybean (*Glycine max* L.)

INTRODUCTION

The pentose phosphate pathway (PPP) is a pivotal carbohydrate metabolic pathway that acts as a key role in plant development and stress responses (Krüger et al., 2011; Caretto et al., 2015). The PPP is one of the major sources of nicotinamide adenine dinucleotide phosphate (NADPH), which is the principal reducing molecule used in many metabolic pathways, such as nitrogen assimilation and amino acids synthesis (Devi et al., 2007; Sharkey and Weise, 2016). The two dehydrogenases in the PPP pathway, 6-phosphogluconate dehydrogenase (6PGDH, EC 1.1.1.44) and glucose-6- phosphate dehydrogenase (G6PDH, EC 1.1.1.49), utilize $NADP^+$ as a cofactor to generate NADPH during the conversion of glucose-6-phosphate (G6P) to pentoses (Wood, 1986). The step catalyzed by G6PDH enzyme is known as the vital reaction in the PPP due to its strict control of the NADPH/$NADP^+$ redox balance.

Genes encoding G6PDH have been cloned and characterized from some plants including *Oryza sativa* (Zhang et al., 2013), *Populus suaveolens* (Lin et al., 2005), *Arabidopsis thaliana* (Wakao and Benning, 2005), *Solanum tuberosum* (Wendt et al., 2000; Hauschild and von Schaewen, 2003), *Hordeum vulgare* (Esposito et al., 2001; Cardi et al., 2015), and *Triticum aestivum* (Nemoto and Sasakuma, 2000), and their involvement in plant development has been reported. Cytosolic (Cy) and plastidic (P) isoforms have been certified for plant G6PDHs based on their subcellular localization (Cardi et al., 2013; Castiglia et al., 2015). In addition, the P-G6PDHs are divided into two types, P1-G6PDH and P2-G6PDH, which can be distinguished by diverse gene expression profiles as well as specific biochemical characteristics, indicating the different functions of each isoform in plant metabolism (Cardi et al., 2016). In *Arabidopsis*, there are six G6PDHs targeted to different subcellular compartments: two cytosolic $NADP^+$-dependent isoforms encoded by *AtG6PDH5* and *AtG6PDH6* genes respectively, and four plastidic $NADP^+$-dependent isoforms encoded by *AtG6PDH1*, *AtG6PDH2*, *AtG6PDH3,* and *AtG6PDH4* genes respectively (Wakao and Benning, 2005). Previous studies have shown that knocking down the cytosolic G6PDH in *Arabidopsis* may inhibit the seed oil accumulation, suggesting that Cy-G6PDH is crucial for regulating the oil biosynthesis during seed development (Wakao et al., 2008). Furthermore, the plastidic isoforms are proved to be essential in providing reducing power (NADPH) for enzymes involved in ammonium assimilation and nitrate reduction (Esposito et al., 2004; Esposito, 2016).

In addition to their pivotal role in developmental processes, the key functions of plant G6PDHs in responses to different types of environmental stresses have been widely proven, including salinity (Wang et al., 2008), cold (Lin et al., 2013), drought (Landi et al., 2016), and heat (Gong et al., 2012). It has been demonstrated that the cytosolic G6PDH is the major contributor to the total cellular G6PDH activity in plants (Castiglia et al., 2015), which seems to be a significant factor for the outcome of abiotic stress responses (Esposito et al., 2001; Honjoh et al., 2007). As a main example, overexpression of a kinetically engineered G6PDH in cytosol enhanced both biotic (defense reactions) and abiotic (drought) stress tolerance of transgenic tobacco through inhibiting NADPH oxidases induced reactive oxygen species (ROS) by improving NADPH provision during early oxidative bursts (Scharte et al., 2009). Besides, overexpression of a cytosolic *PsG6PDH* gene from *Populus suaveolens* confers an increased cold tolerance in transgenic tobacco by elevating the activity of antioxidative enzymes, such as peroxidase and superoxide dismutase, and decreasing the level of membrane lipid peroxidation (Lin et al., 2005, 2013). Also, the enhanced cytosolic G6PDH activities would contribute to the improvement of drought tolerance in soybean roots, with the involvement of ABA-dependent signaling pathway (Liu et al., 2013; Wang et al., 2016).

The characteristics of *G6PDHs* with respect to salt resistance have been validated by several researchers (Liu et al., 2007; Wang et al., 2008; Sang et al., 2018). It has been shown that the oxidative burst is counteracted, more or less, by the activities and expression of G6PDH isoforms upon salt stress (Valderrama et al., 2006; Liu et al., 2007). The Cy-G6PDH isoforms in *Arabidopsis* are essential in the provision of NADPH to maintain the cellular redox homeostasis via the phosphorylation of Thr467 by glycogen synthase kinase 3 under high salinity condition (Dal Santo et al., 2012); and this process is identified as associated with a sugar-signaling molecule (Lejay et al., 2008). Northern hybridization revealed a noticeable increase of *TaG6PDH* transcripts within 12 h of exposure of wheat seedlings to salinity treatment (150 mM NaCl), and the high transcript level was maintained for several hours (Nemoto and Sasakuma, 2000, 2002). In addition, the G6PDH enzyme in rice is the key in sustaining ROS homeostasis by regulating the coordination state of G6PDH activity and NAPDH oxidase under salt stress, however, the molecular metabolism have not been investigated (Zhang et al., 2013).

Although the biological functions of G6PDH in stress responses have been described in several model plants, few information are known about soybean. Here, we characterized nine members of the *G6PDH* gene family in soybean. The cellular locations of *GmG6PDHs* were initially predicted by transit peptide analysis and subsequently verified by transient expression of GFP-tagged GmG6PDH fusion proteins in *Arabidopsis* protoplasts. We also determined the transcriptional profiles of *GmG6PDHs* in distinct organs and under various abiotic stress using qRT-PCR and high-throughput sequencing data analyses. Most notably, one cytosolic isoform (*GmG6PDH2*) had apparent transcriptional response to salt stress and did well correlate with the activity of G6PDH enzyme, possibly implying a prominent role for this isoform in response to salinity. The prokaryotic expression of *GmG6PDH2* in *Escherichia coli* demonstrated that this gene encoded an active G6PDH enzyme. In addition, overexpression of *GmG6PDH2* in soybean hair roots increased the salt tolerance in transgenic soybean seedlings, with higher AsA/DHA (ascorbic acid/dehydroascorbate), GSH/GSSG (reduced/oxidized glutathione) ratios, lower levels of ROS and lipid peroxidation. These findings indicate that the soybean *G6PDHs* participate in plant growth and stress responses, of which the cytosolic *GmG6PDH2* is the main isoform in regulating the cellular redox pool and defending against oxidative stress.

MATERIALS AND METHODS

Identification of *G6PDH* Gene Family in Soybean

To obtain all *G6PDHs* from the soybean genome, a systematic BLASTP search was carried out against the soybean genetics and genomics database (SoyBase[1]) using the published *A. thaliana* G6PDHs as queries. The protein sequences of putative soybean G6PDH family members with an E-value of $<10^{-10}$ and a sequence identity threshold > 90% were downloaded. The candidate genes were further verified by SMART[2] and Pfam[3] databases to confirm the presence of a C-terminal NADP-dependent G6PD domain (PF02781) and an N-terminal $NADP^+$-binding domain (PF00479). Information about the genetic characteristics of *GmG6PDHs*, including coding sequence lengths, chromosome locations and protein lengths were collected from the SoyBase. The isoelectric point and molecular mass were determined on ExPASy server[4]. Transit peptides and subcellular localization were predicted using TargetP 2.0[5] and CELLO 2.5[6] (Yu et al., 2004).

Evolutionary, Gene Structure, and Synteny Analyses of GmG6PDHs

The full-length proteins of G6PDHs from *Glycine max* (GmG6PDHs), *Zea mays* (ZmG6PDHs), *Oryza sativa* (OsG6PDHs), *Phaseolus vulgaris* (PvG6PDHs), *Medicago truncatula* (MtG6PDHs), *Sorghum bicolor* (SbG6PDHs), *Brachypodium distachyon* (BdG6PDHs), and *A. thaliana* (AtG6PDHs) were used for building a phylogenetic tree using MEGA 5.0 software based on neighbor-joining method with the default parameter values (Tamura et al., 2011). The gene structures of *G6PDHs* were affirmed using the GSDS database[7] by aligning the coding regions with their corresponding genomic regions. The genomic sequences and coding sequences of *G6PDH* genes in *G. max* and *A. thaliana* were obtained from the soybean genetics and genomics database and NCBI database. The syntenic blocks among *G. max*, *Z. mays*, *A. thaliana*, *O. sativa*, *P. vulgaris*, *M. truncatula*, *B. distachyon,* and *S. bicolor G6PDHs* were identified from the plant genome duplication database (PGDD[8]) (Lee et al., 2012). The gene ID and other information of the *G6PDHs* used in this study were available in **Supplementary Table S1**.

Promoter Analysis of *GmG6PDHs*

To investigate the critical *cis*-acting elements in the promoter of *GmG6PDH* genes, 2.0 kb upstream of the position of the ATG codon in these genes were obtained from the soybean genetics and genomics database[9]. The *cis*-acting regulatory DNA elements were predicted from the PlantCARE database[10] and presented by the IBS 2.0 (Liu et al., 2015).

Subcellular Localization

The entire coding region of five *GmG6PDH* genes were amplified from the seeds of soybean cultivar "SN14" (provided by the Soybean Breeding Research Center of Northeast Agricultural University, Haerbin, China) by reverse transcription- polymerase chain reaction (RT-PCR) with the high-fidelity KOD-Plus-DNA polymerase (TOYOBO, Osaka, Japan). These genes were further constructed into pBI121 vector, which both contain a CaMV35S promoter and green fluorescent protein (GFP) tag. The gene-specific primers used for cloning the putative *G6PDH* genes were shown in **Supplementary Table S2**. The fusion proteins *pBI121- GmG6PDHs:GFP* or positive control (empty vector) were temporarily expressed in *Arabidopsis* mesophyll protoplasts, which were isolated from the leaves of 14 days-old seedlings grown under a weak light condition to minimize the chloroplast autofluorescence. The subcellular location of GmG6PDH-GFP proteins was monitored 14 h after polyethylene glycol (PEG)-mediated protoplast transfection protocol (Yoo et al., 2007). Confocal laser-scanning microscopy (LSM 710, Carl Zeiss, Jena, Germany) with a 488-nm argon ion laser (for GFP excitation) was used to visualize and localize GFP-tagged proteins. The excitation/emission wavelength were as follows for GFP (488 nm/507 to 535 nm) and chlorophyll autofluorescence (610 nm/650 to 750 nm).

Expression Analysis of *GmG6PDHs*

The transcriptional patterns of *GmG6PDHs* in multiple tissues via high-throughput sequencing data from Phytozome database[11], including leaves, root, root hairs, shoot apical meristem, nodules, stem, seed, pod, and flower tissues. The results are shown as heat maps with hierarchical clustering using the software TBtools 0.665 (Chen et al., 2018) and the values were log2-transformed with normalization. To analyze the transcriptional profiles of the *GmG6PDHs* in different stages of seed development, total RNA was extracted from soybean seeds at 4, 7, 14, 30, 50, 80, 110, and 120 days after flowering (DAF). The expression level of *GmG6PDHs* in developing seeds at 4DAF was used as a calibrator. To examine the transcriptional profiling of *GmG6PDHs* under various abiotic stresses, soybean seedlings at the second trifoliolate stage were subjected to salt stress induced by 150 mM NaCl, alkali stress induced by 100 mM $NaHCO_3$, and osmotic stress induced by 20% (w/v) PEG (with a molecular weight of 6000 g/M) or 200 mM mannitol solutions. The osmotic potential of 20% PEG6000 and 200 mM mannitol was −0.53 and −0.50 MPa, respectively. Total RNA was extracted from leaf samples at 0, 6, and 12 h after the above treatments. The transcripts of *GmG6PDHs* in soybean leaf under normal environment condition were used as a calibrator. *GmGAPDH* and *GmACTIN* were used as internal reference. Each quantitative

[1] https://www.soybase.org/sbt
[2] http://smart.embl.de/
[3] http://pfam.sanger.ac.uk/
[4] http://expasy.org/
[5] http://www.cbs.dtu.dk/services/TargetP/
[6] http://cello.life.nctu.edu.tw/
[7] http://gsds.cbi.pku.edu.cn/index.php
[8] http://chibba.agtec.uga.edu/duplication
[9] https://www.soybase.org/sbt
[10] http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
[11] https://phytozome.jgi.doe.gov/pz/portal.html

real time-polymerase chain reaction (qRT-PCR) reaction was performed in triplicate (technical replicates) on three biological replicates and the transcriptional level of *GmG6PDHs* was calculated based on the $2^{-\Delta\Delta ct}$ method. All the primers used for qRT-PCR were available in **Supplementary Table S2**.

Recombinant Protein Expression and Enzyme Kinetic Property Assay

The CDS of *GmG6PDH2* with the *Xho*I and *Nco*I sites was inserted into the prokaryotic expression vector pET32a (+). The recombinant plasmid *pET32a- GmG6PDH2* was transformed into *E. coli* Rosetta strain to produce the putative recombinants. The positive clone was sequenced and cultivated in liquid LB medium supplemented with 1 mM IPTG at 37°C for 4 h to induce the expression of *GmG6PDH2*. The recombinant proteins were wall-broken by ultrasonic wave with a power output of 250 W for 10 min and then harvested by centrifuging at 10,000 *g* for 15 min. The supernatant was detected by 12% SDS-PAGE and collected for enzymatic activities assay. The total protein concentration was monitored by Bradford Protein Assay Kit purchased from Solarbio Science and Technology (Beijing, China). The kinetic parameters of GmG6PDH2 recombinant proteins with regard to the glucose-6-phosphate (G6P) and $NAPD^+$ was determined using Eadie–Hofstee plot (Wakao and Benning, 2005).

G6PDH Activity Assays

G6PDH activity assays were performed as described by Wakao and Benning (2005), with slight modifications. The soybean roots (0.2 g) were extracted in 5 mL extraction buffer containing 50 mM 2-[4-(2-Hydroxyethyl)-1-piperazinyl]ethanesulfonic acid -Tris(hydroxymethyl)aminomethane (Hepes-Tris) buffer (pH 7.8), 1 mM EDTA, 3 mM $MgCl_2$ and 1 mM phenylmethylsulfonyl fluoride. The G6PDH activities were measured with regard to the oxidation of G6P by $NADP^+$. The total reaction mixture was reduced to 1 mL with 0.5 mM $NADPNa_2$, 0.5 mM D-glucose-6-phosphate disodium salt, 3.3 mM $MgCl_2$, 50 mM Hepes-Tris (pH 7.8) and an appropriate amount of enzyme extracts (Wakao and Benning, 2005). Hepes-Tris buffer was made as follows: 0.5M Hepes was titrated to pH 7.8 with about 1M Tris, and then diluted 10-fold to give 50 mM Hepes.

Agrobacterium-Mediated Over-Expression of *GmG6PDH2* in Soybean Hairy Roots

The plasmid of *pBI121-GmG6PDH2:GFP* was transformed by electroporation into *Agrobacterium rhizogenes* strain K599, which was used to transform soybean hypocotyls. Soybean transformation in "SN14" hypocotyls and hairy root induction were performed as reported previously (Tóth et al., 2016). Soybean plants infected with the *A. rhizogenes* strain K599 were considered as control hairy roots. The transgenic lines were screened by PCR amplification and enzyme activity assay as described (Pan et al., 2016), and the non-transgenic hair roots were removed from the seedlings. Transgenic lines with similar-length hairy roots were selected and treated with 1/2 Hoagland solution containing 0 or 100 mM NaCl for 5 days, respectively. The root fresh weight and maximum root length of transgenic soybean plants were researched after 5 day of salt treatment. More than 10 independent hairy root lines were analyzed in this work to check the effects of *GmG6PDH2* over expression on salinity stress responses.

Analysis of Cellular ROS Levels and Antioxidant Contents

The metabolites contents of AsA-GSH cycle, including AsA and GSH, and their oxidized forms, DHA and GSSG were determined using Ascorbic Acid or Glutathione Colorimetric Assay Kit purchased from Solarbio Science and Technology (Beijing, China), as per the manufacturer's protocol. The assays of hydrogen peroxide (H_2O_2) content were conducted according to the method published previously (Velikova et al., 2000). The membrane damage was determined with regard to thiobarbituric acid- reactive substances (TBARS) content, a product of lipid peroxidation (Hodges et al., 1999). Briefly, the soybean root samples (0.5 g) were extracted with 10 mL 0.1% (w/v) trichloroacetic acid (TCA), and then the homogenate was centrifuged at 10,000 *g* for 10 min at 4°C. The supernatant was used for the determination of H_2O_2 and TBARS contents. The total reaction volume of H_2O_2 assay was 2 mL containing 0.5 mL of the supernatant, 1 mL 1M potassium iodide and 0.5 mL 10 mM potassium phosphate buffer (pH 7.0). The intensity was measured at 390 nm. The total reaction volume of TBARS assay was 2 mL containing 0.5 mL of the supernatant and 1.5 mL 0.5% (w/v) thiobarbital acid in 15% TCA. The absorbancy of supernatant was read at both 532 and 600 nm.

Statistical Analysis

All experiments were performed with at least three biological replicates. Values are presented as mean ± SD. The significance of the data was evaluated using Student's *t*-test with SPSS statistics 22.0 software. The significance level was set at $P < 0.05$.

RESULTS

Identification and Classification of *G6PDH* Gene Family in Soybean

In this study, full-length proteins and conserved domains of six glucose-6-phosphate dehydrogenases (G6PDHs) in *A. thaliana* were used as BLAST queries against the soybean genetics and genomics database[12]. A total of nine *G6PDH* genes were originally obtained in the soybean genome, which were designated as *GmG6PDH1-9* (**Table 1**). Full-length coding sequences of *GmG6PDH 1-9* ranged from 1560 to 1839 bp, and encoded nine putative proteins with 518 to 612 amino acid residues. The protein isoelectric points and molecular mass of the nine GmG6PDHs varied from 5.80/59.3 to 8.76/68.8 kDa, respectively (**Table 1**). Online software, CELLO 2.5 and TargetP 1.1, were used to assess the existence of a predicted N-terminal transit peptide (TP), which indicated the localization of GmG6PDH2, 4 and 6 in the cytosol, and others in the plastid (**Table 1**).

[12]https://www.soybase.org/sbt

TABLE 1 | Basic information of the nine soybean *G6PDH* genes (*GmG6PDHs*).

Gene name	Gene ID[1]	Gene location	ORF length (bp)	Protein length	Isoelectric point	Molecular mass (KDa)	Subcellular localization
GmG6PDH1	Glyma.03G229400.1	Gm3 43144328-43150674	1839	612	6.58	68.8	Plastidic
GmG6PDH2	Glyma.19G082300.1	Gm19 29813147-29821693	1557	518	6.32	59.3	Cytoplasmic
GmG6PDH3	Glyma.08G199000.1	Gm8 16078525-16083556	1767	588	8.76	66.3	Plastidic
GmG6PDH4	Glyma.16G063200.1	Gm16 6210393-6217815	1557	518	5.80	59.3	Cytoplasmic
GmG6PDH5	Glyma.02G096800.1	Gm2 8700742-8705334	1809	602	8.28	68.2	Plastidic
GmG6PDH6	Glyma.19G077300.1	Gm19 27787739-27797138	1560	519	6.32	59.7	Cytoplasmic
GmG6PDH7	Glyma.18G284600.1	Gm18 56525770-56534088	1806	601	7.63	67.9	Plastidic
GmG6PDH8	Glyma.07G013800.1	Gm7 1073133-1082652	1767	588	8.03	66.5	Plastidic
GmG6PDH9	Glyma.19G226700.1	Gm19 47838450-47844119	1815	604	6.37	68.1	Plastidic

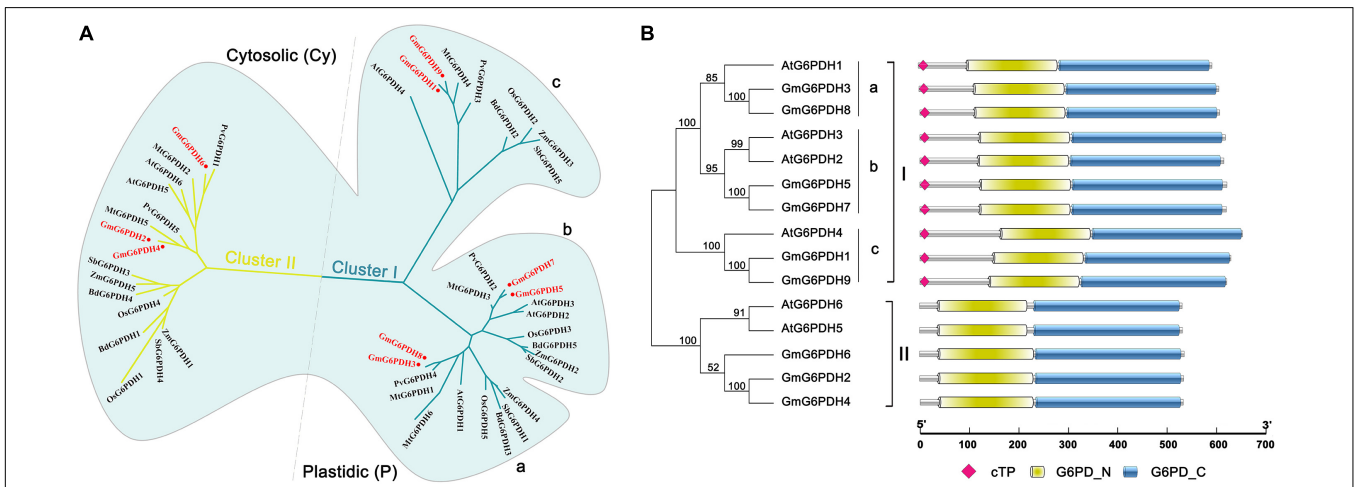


FIGURE 1 | Phylogenic and protein domain analysis of GmG6PDHs. **(A)** Phylogenetic tree of G6PDH proteins from *Glycine max* (red circles), *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, *Phaseolus vulgaris*, *Medicago truncatula*, *Brachypodium Distachyon,* and *Sorghum bicolor*. **(B)** The bi-domain structure of the soybean and *Arabidopsis* G6PDH proteins. Predicted chloroplast transit peptide (cTP) is depicted as colored diamonds.

To examine the classification and evolutionary history of soybean G6PDHs, the full-length protein sequences of GmG6PDHs were aligned with homologous G6PDHs from *S. bicolor* (StG6PDH1-4), *A. thaliana* (AtG6PDH1-6), *O. sativa* (OsG6PDH1-5), *Z. mays* (ZmG6PDH1-6), *B. distachyon* (BdG6PDH1-5), *M. truncatula* (MtG6PDH1-7), and *P. vulgaris* (PvG6PDH1-5), and a phylogenetic tree was constructed. As shown in **Figure 1A**, the phylogenic analysis suggested that the different plant G6PDHs could be clearly classified into two major clades (I and II). Clade I corresponded to plastidic (P) isoforms containing four *Arabidopsis* P-G6PDHs (AtG6PDH1-4) (Wakao and Benning, 2005; Née et al., 2009). Clade I was further segmented into three classes (a, b, and c), in which GmG6PDH3 and GmG6PDH8 were subdivided into class a, along with an *Arabidopsis* P1 isoform (AtG6PDH1); GmG6PDH5, 7 and two *Arabidopsis* P2 isoforms (AtG6PDH2, 3) fall into class b; GmG6PDH1, 9 and an inactive-G6PDH isoform (AtG6PDH4) were clustered within class c. Moreover, clade II corresponded to the cytosolic (Cy) isoforms, including GmG6PDH2, 4 and 6, together with two *Arabidopsis* Cy-G6PDHs (AtG6PDH5, 6) (Wakao et al., 2008). The phylogenetic clades were in accordance with the *in silico* prediction of GmG6PDHs. As expected, the GmG6PDH isoforms were more closely related to its homolog from *P. vulgaris* in each cluster, which all belonged to the leguminous family (**Figure 1A**).

Structural analysis of translated proteins for *GmG6PDH1-9* demonstrated that the GmG6PDHs exhibited a bi-domain protein structure similar to the reported *A. thaliana* G6PDH

proteins, consisting of an N-terminal $NADP^+$-binding domain (PF00479) and a C-terminal G6PD domain (PF02781) (**Figure 1B**). Two motifs, $NADP^+$-binding motif (NEFVIRLQP) and substrate-binding motif (RIDHYLGKE), were highly conserved in all the GmG6PDH proteins (**Supplementary Figure S1**). These protein sequences also contained a conserved Rossman fold (GXXGDLA) domain (**Supplementary Figure S1**). Based on the subcellular localization prediction, phylogenetic and protein structural analyses, GmG6PDHs were separated into two types: three cytosolic $NADP^+$-G6PDH isoforms (GmG6PDH2, 4 and 6), and six plastidic $NADP^+$-G6PDH isoforms (GmG6PDH1, 3, 5, 7, 8, and 9). The specific localization feature as represented by soybean *G6PDH* genes revealed the relatedness of distinct function for G6PDH isoforms in each clade with their evolutionary process.

Syntenic Relationship and Gene Structure Analysis of *GmG6PDHs*

A synteny analysis among *G6PDHs* from *G. max*, *P. vulgaris*, *Z. mays*, *O. sativa*, *A. thaliana*, *M. truncatula*, *B. distachyon,* and *S. bicolor* was performed in the present study to gain some insight into the potential function of *GmG6PDHs*. As shown in **Figure 2A**, the nine *GmG6PDHs* were scattered along seven out of twenty soybean chromosomes and each of the seven chromosomes comprised one to three *GmG6PDHs*. Moreover, a total of 30 orthologous pairs of *G6PDHs* were found in the above eight species (**Supplementary Figure S2** and **Supplementary Table S3**). The *GmG6PDHs* had syntenic relationship only with *PvG6PDHs*, *OsG6PDHs*, *SbG6PDHs* and *BdG6PDHs*, including four orthologous gene pairs between *G. max* and *O. sativa* or *S. bicolor*, five orthologous gene pairs between *G. max* and *B. distachyon*, and 12 orthologous gene pairs between *G. max* and *P. vulgaris* (**Figure 2A**). Besides, five paralogous *G6PDH* gene pairs were confirmed in soybean genome, and the paralogous gene pairs were apt to be found among the members in the same subfamily (**Figure 2A**). To obtain further details about the structural diversity of *GmG6PDH* genes, we subsequently compared the localization and size of exon/intron among *GmG6PDHs* and *AtG6PDHs*. As shown in **Figure 2B**, *G6PDH* genes belonged to the same cluster shared the similar exon–intron structures, particularly in relation to exon numbers. For instance, *G6PDH* genes in cluster I possess 8–12 exons, and *G6PDH* genes in cluster II exhibited an equal number of exons (15) and nearly identical exons length. These results indicated that the *GmG6PDH* genes were highly conserved in gene sequence and exon–intron organization within each phylogenic group.

Regulatory Elements in the *GmG6PDH* Promoters

To identify putative *cis*-elements involved in *GmG6PDHs* transcriptional regulation, a 2.0 kb promoter region upstream from the ATG translation start codon of each *GmG6PDH* gene was analyzed. As shown in **Figure 3**, most *GmG6PDH* genes displayed the existence of some stress responsive *cis*-elements, such as ARE, a *cis*-regulatory element involved in anoxic-inducibility, was found in all *GmG6PDH* genes, except for *GmG6PDH3*. Beyond that, MBS, a *cis*-element involved in drought responsiveness, was found in *GmG6PDH2*, *GmG6PDH4*, *GmGPDH6,* and *GmGPDH7* genes; TC-rich repeats, a *cis*-element associated to stress and defense responsiveness, was observed in *GmG6PDH8*; and LTR (cold-responsive element) was seen in *GmG6PDH5* and *GmG6PDH9* genes (**Figure 3**). Meanwhile, all *GmG6PDH* promoter sequences contained one or more *cis*-elements involved in response to multiple hormones, such as auxin-responsive element (TGA),

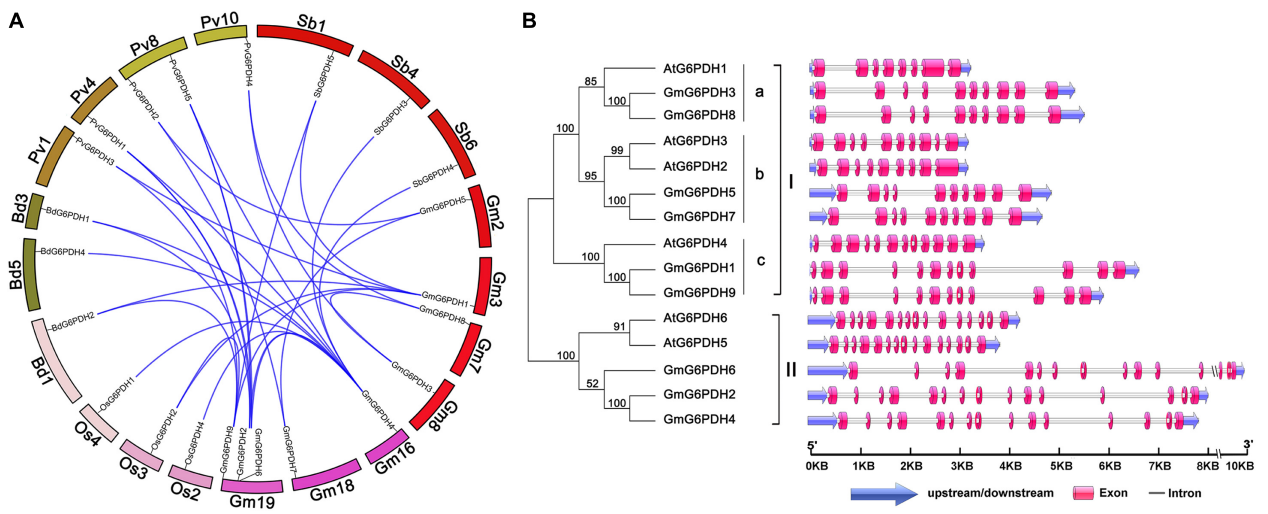


FIGURE 2 | Syntenic and exon–intron structural analysis of *GmG6PDH* family genes. **(A)** Syntenic analysis of soybean *G6PDHs* with the corresponding genes in *B. distachyon*, *P. vulgaris*, *S. bicolor,* and *Oryza sativa*. The chromosome of the above species is shown as a circle. The colored curves indicate the collinear region of the G6PDHs. **(B)** The exon–intron organization of *AtG6PDHs* and *GmG6PDHs*. Untranslated regions (UTRs) are indicated by blue arrow. Exons and introns are visualized as colored boxes and gray lines respectively.

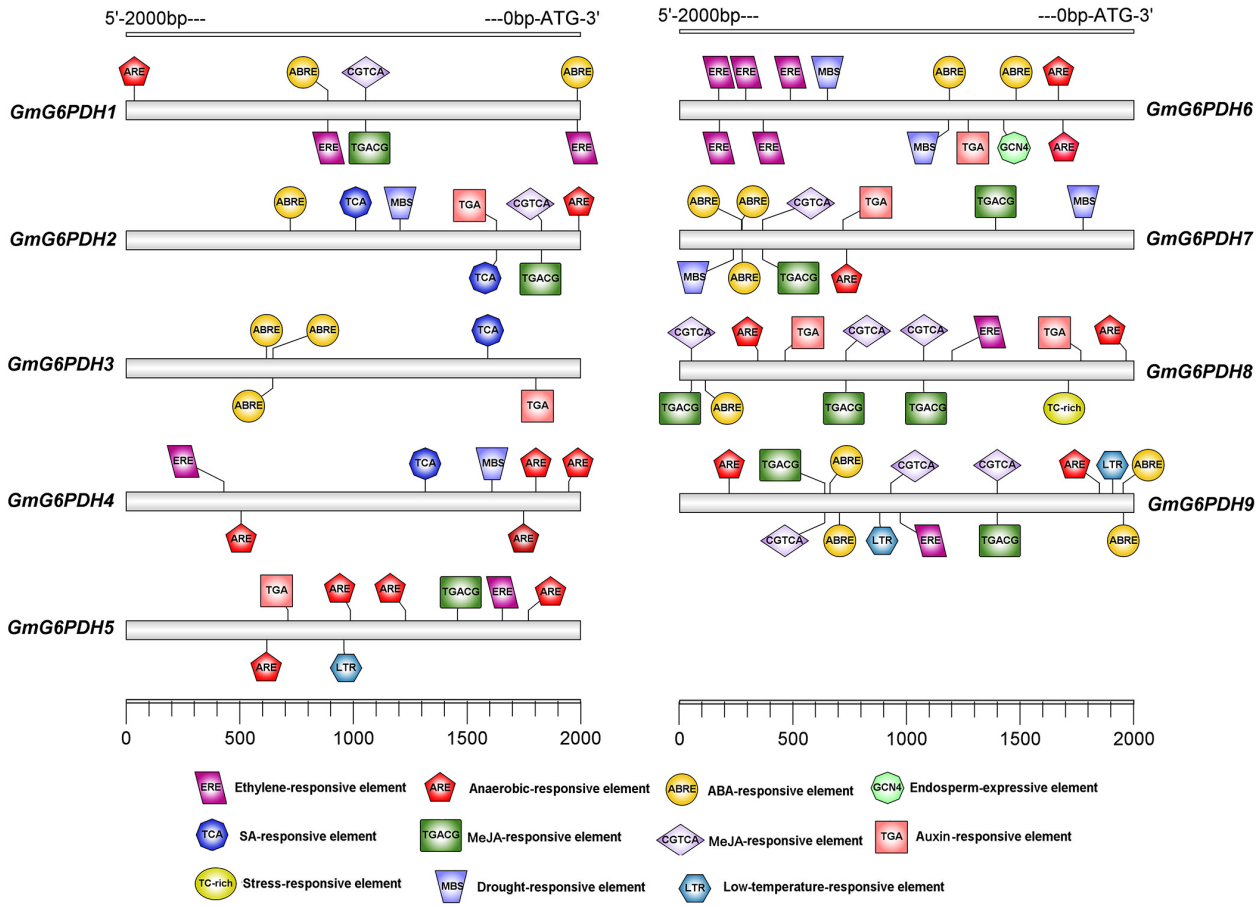


FIGURE 3 | *Cis*-elements prediction in the 2.0 kb promoter region upstream from the start codon of *GmG6PDHs*. The relative positions of *cis*-elements in each *GmG6PDH* gene are marked by different-colored boxes.

ABA-responsive element (ABRE), and SA-responsive element (TCA). Remarkably, GCN4, a *cis*-acting element required for endosperm expression was observed in *GmG6PDH6*. Bioinformatics analyses of *cis*-elements suggest *GmG6PDHs* may be pivotal in mediating stress responses as well as plant growth.

Subcellular Localization of *GmG6PDHs*

The cellular location of G6PDH proteins is closely linked to and indicates their functions. To certify the subcellular localization of GmG6PDHs, the coding sequences of five *GmG6PDHs* (*GmG6PDH1*, *2*, *5*, *8,* and *9*) were successfully cloned, verified and submitted to GenBank with the following accession numbers: *GmG6PDH1* (MN339553), *GmG6PDH2* (MN339554), *GmG6PDH5* (MN339555), *GmG6PDH8* (MN339556), and *GmG6PDH9* (MN339557). We then fused the coding region of five *GmG6PDHs* in-frame with the N-terminus of the GFP coding region. The positive control (35S: GFP) and GFP-tagged GmG6PDH proteins were transiently expressed in *Arabidopsis* mesophyll protoplasts. The free GFP protein was evenly distributed throughout the cell except for chloroplast and vacuole (**Figure 4**), while the GmG6PDH1, 5, 8 and 9 were specifically targeted to the chloroplast. Moreover, the GmG6PDH2 fusion protein was merely detected in the cytosol, suggesting that it encoded a cytoplasmic protein (**Figure 4**). These results were in line with the previous online prediction.

Expression Profiles of *GmG6PDHs* in Different Tissues and Developmental Phases

In succession, we evaluated the transcriptional patterns of *GmG6PDHs* in multiple tissues *via* high-throughput sequencing data from Phytozome database, including leaves, root, root hairs, shoot apical meristem, nodules, stem, seed, pod, and flower tissues. The transcripts of nine *GmG6PDHs* could be observed in all the tissues tested, but the transcriptional profiles were different between the cytosolic (*GmG6PDH2*, *4,* and *6*) and plastidial (*GmG6PDH1*, *3*, *5*, *7*, *8,* and *9*) isoforms (**Figure 5A**). The cytosolic *GmG6PDHs* were strongly expressed in roots, pod and nodules. Nonetheless, the transcripts of plastidial *GmGPDH3*, *5,* and *9* genes were primarily seen in leaves, and *GmG6PDH7* gene was mainly expressed in pod. The transcriptional levels of *GmG6PDH1* and *8* genes were comparatively low in all nine tissues compared with other genes. Furthermore, the *GmG6PDHs*

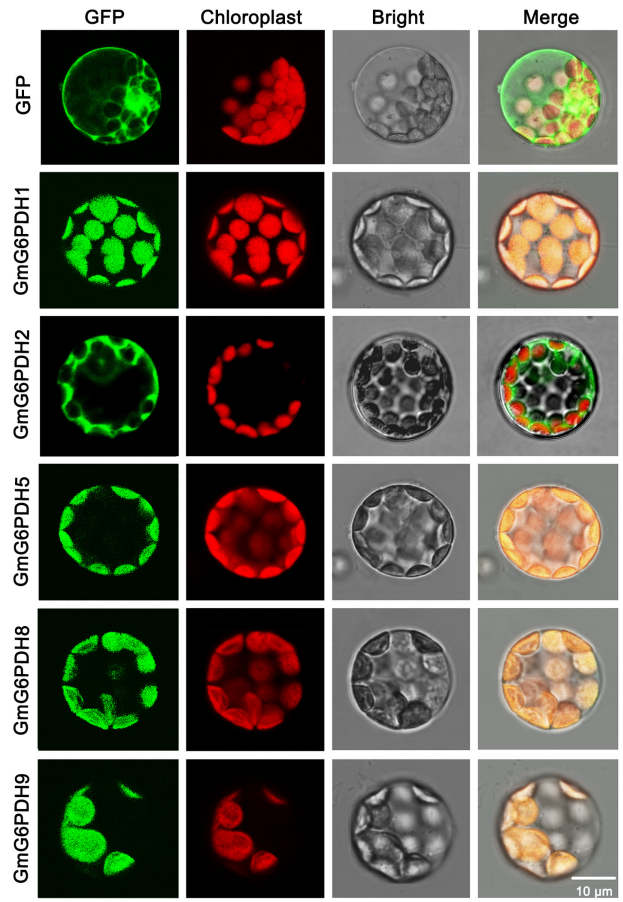


FIGURE 4 | Subcellular localization analysis of GmG6PDHs by transiently expressing five GmG6PDH-GFP fusion proteins in *Arabidopsis* mesophyll protoplasts. Confocal micrographs showing the subcellular localization of GmG6PDH1, GmG6PDH2, GmG6PDH5, GmG6PDH6, and GmG6PDH9 proteins. Merged pictures of the GFP fluorescence (first panels), the chloroplast autofluorescence (second panels) and the corresponding bright field (third panels) are shown in the fourth panels.

were classed into two major groups (I and II) based upon their expression patterns across all nine tissues, which is correlated with the phylogenetic clades for GmG6PDHs (**Figure 5A**). Within each subgroup, most of these genes exhibited similar expression profiles. Notably, the average relative mRNA levels of cytosolic *GmG6PDHs* in group II were much higher than that of plastidic *GmG6PDHs* in group I. The tissue-specific expression characteristics of *GmG6PDHs* reflected their versatile functions in multiple aspects of soybean growth and development.

In addition, it has been reported that G6PDH enzymes are required for the accumulation of lipid and starch in developing seeds (Wakao et al., 2008). Therefore, to explore the potential function of *GmG6PDHs* during soybean seed development, the transcriptional patterns of *GmG6PDHs* in developing seeds at different stages was further affirmed by qRT-PCR. All the tested genes remained extremely high expression levels during early-middle seed development (14, 30, and 50 DAF) and low expression levels during the late maturation stage of seed development (110 and 120 DAF) (**Figure 5B**). The maximum level of *GmG6PDH2*, *4*, *5*, *7*, and *9* genes expression occurred in developing seeds at 50 DAF, while the expression level of *GmG6PDH6* was peaked in developing seeds at 30 DAF. Among the tested genes, the transcript abundance of cytosolic *GmG6PDH2* gene was the highest at several stages of seed development (30, 50, 80, 110, and 120 DAF).

Enzyme Activity and Transcript Level of *GmG6PDHs* Under Abiotic Stress

It is well-known that G6PDH genes are important for stress adaptation in several model plants (Wakao and Benning, 2005; Long et al., 2016). Also, the promoter analysis of soybean *G6PDHs* revealed a number of potential *cis*-elements and transcription binding motifs (MBS, ABRE, ARE, and TCA-elements) involved in responsiveness to abiotic stresses, such as salt and drought stresses frequently encountered in our area (northeast China). Hence, to further understand how *GmG6PDHs* respond to abiotic stresses, we analyzed the transcriptional profiles of *GmG6PDHs* under alkali (100 mM $NaHCO_3$), salt (150 mM NaCl), and osmotic (20% PEG or 200 mM mannitol) stresses. As shown in **Figure 6**, there were distinct differences among cytosolic and plastidial *GmG6PDH* isoforms based on expression patterns under various stress conditions. The expression levels of cytosolic *GmG6PDHs* (*GmG6PDH2, 4, and 6*) were extraordinarily high upon salt stress, of which the transcript expression of *GmG6PDH2* was highly inducible (reaching upto 100-fold) under the salinity stress treatment for 6 h. The plastidic forms *GmG6PDH1, 3, 7, and 8* were also upregulated under salinity stress of 6 h (about 30–40-fold), while the average mRNA levels of *P-G6PDHs* was much lower than that of *Cy-G6PDHs*. Also, the cytosolic *GmG6PDHs* were dramatically induced at the early stage of PEG treatment, and the transcripts of *GmG6PDH6* at 6 h of treatment were much higher than that of other genes. Furthermore, most of *GmG6PDH* genes were apparently up-regulated under alkaline treatment, particularly the plastidial *GmG6PDHs,* which had obviously high level of transcription at 6 h. In contrast, the cytosolic *GmG6PDHs* showed the maximum transcriptional level at 12 h of the same treatment. Likewise, the plastidial *GmG6PDHs* were originally stimulated after 6 h of mannitol treatment, maintaining comparatively high abundant transcripts throughout the entire treatment period, while the cytosolic *GmG6PDHs* were markedly up-regulated after 12 h (**Figure 6**).

Similarly, the enzymatic assay revealed that the soybean G6PDH activity was significantly increased under NaCl, $NaHCO_3$, PEG and mannitol treatments (**Figure 7A**). The G6PDH activities under the above treatments were approximately 1.5- to 4.2 -fold higher than that under the normal condition. What's more, salt treatment exhibited an even higher stimulatory influence on G6PDH activity than other treatments, which causes the rapid increases of the enzyme activity within 6 h after salt stress (**Figure 7A**). During drought treatment, the G6PDH activity first peaked at 6 h and was raised again during the 12–24 h time period. The level of the G6PDH activity increased and then decreased during alkali treatment, and

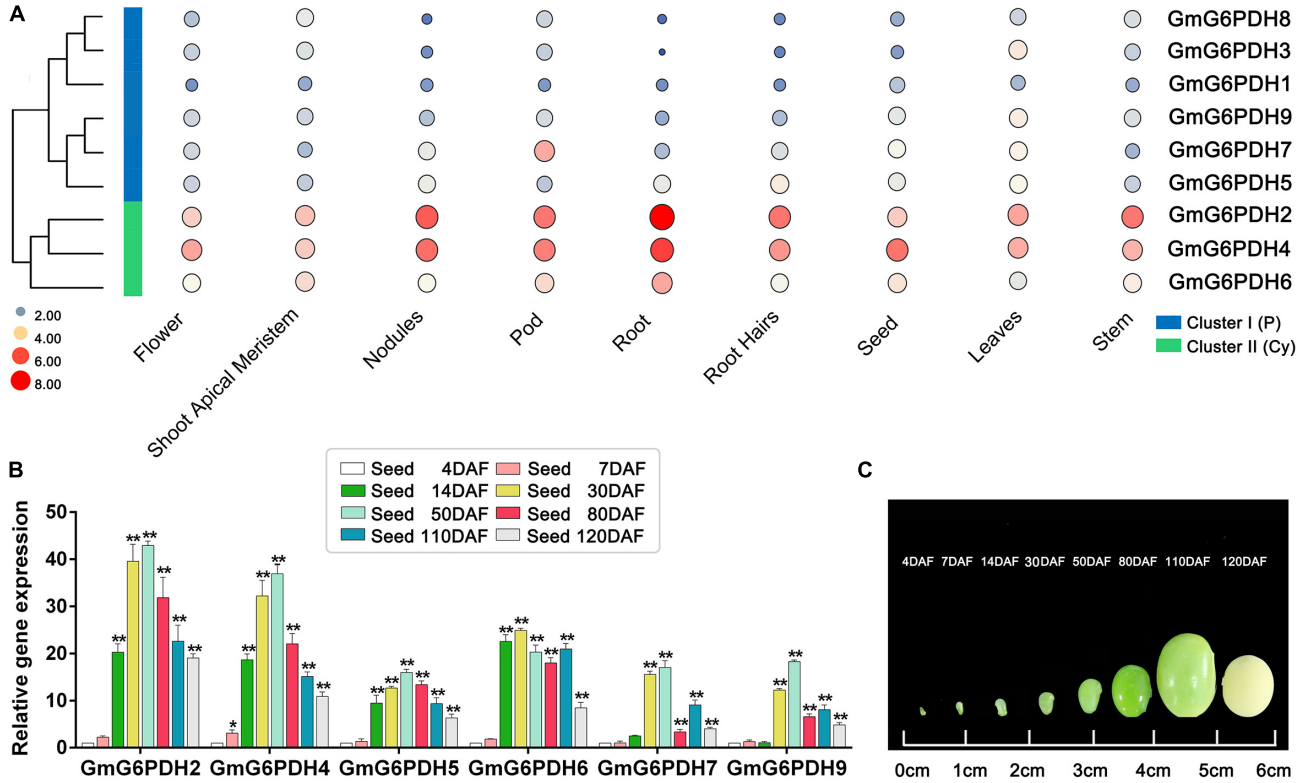


FIGURE 5 | Expression profiles of *GmG6PDHs* in multiple tissues and developmental stages. **(A)** Expression patterns cluster analysis of *GmG6PDHs* involved in tissue development. The transcripts of *GmG6PDH* genes in various tissues are investigated from Phytozome database. The results are shown as heat maps. The color scale represents log2 expression values, with red denoting high-level transcription and blue denoting low-level transcription. The size of the circle also indicates the transcription level, with a larger circle indicating high-abundance transcripts. **(B)** The transcript profiling and **(C)** physiological phenotype of *GmG6PDHs* in developing seeds at 4, 7, 14, 30, 50, 80, 110, and 120 days after flowering (DAF). The transcripts of *GmG6PDHs* in developing seeds at 4 DAF were served as the internal reference. Three biological replicates for each tissue were collected. Asterisks above bars denote a statistically significant difference by Student's *t*-test (*$P < 0.05$, **$P < 0.01$).

peaked at 12 h, while the G6PDH activity exhibited a continuous increase during osmotic treatment. Meanwhile, correlation analysis revealed the trend in G6PDH activity under abiotic stresses was well-consistent with gene expression of cytosolic *GmG6PDH2*, which indicated that it to encode the major G6PDH isoform involved in response to abiotic stresses (**Figure 7B**).

Overexpression of *GmG6PDH2* Increases Salt Tolerance

To deeply comprehend the resistance functions of *GmG6PDH2* gene, its catalytic characteristics were first determined by extracting the crude proteins from *E. coli* cells expression plasmid of His-tagged *GmG6PDH2*. An expected molecular mass of GmG6PDH2-His fusion protein (consisting of histidine marker and target gene) was identified by SDS/PAGE (**Figure 8A**). The recombinant GmG6PDH2 protein was analyzed for its enzyme kinetic with regard to both substrates: glucose-6-phosphate (G6P) and $NADP^+$. The kinetic parameters were calculated by using Eadie–Hofstee data plots, in which the K_m of G6P and $NADP^+$ was estimated as 2.14 and 0.033 mM respectively, and the V_{max} of G6P and $NADP^+$ was estimated as 0.67 and 0.62 umol $\cdot$ min^{-1} $\cdot$ mg^{-1} protein respectively (**Figures 8B,C**). These results demonstrated that the protein encoded by *GmG6PDH2* had the functional $NADP^+$-dependent G6PDH enzyme activity.

Next, to better elucidate how *GmG6PDH2* gene contribute to the response to high salinity, *GmG6PDH2*-overexpressing soybean hairy roots (OHRs) were generated by *Agrobacterium rhizogenes*-mediated gene transformation, of which 10 positive transgenic plants were verified by PCR detection (**Supplementary Figure S3**). The physiological activity test signified that the G6PDH enzyme activity in *GmG6PDH2*-OHR lines were 1.7- to 2.5-fold higher than that in the control hair roots (CHR), which suggested that *GmG6PDH2* gene was expressed successfully and functioned with G6PDH activities in OHR lines (**Supplementary Figure S3**). To look into the effects of *GmG6PDH2* overexpression on salt tolerance at soybean seedling stage, 4-week-old CHR and OHR plants were transferred to 1/2 Hoagland solution supplemented with 0 or 100 mM NaCl for 5 days. As demonstrated in **Figure 8D**, the CHR seedlings exhibited a severe growth-inhibitory phenotype, with a partial leaf shriveling phenomenon, whereas *GmG6PDH2*-OHR plants displayed significantly improved resistance to salinity stress, as reflected by a higher root fresh weight and root length (**Figures 8E,F**).

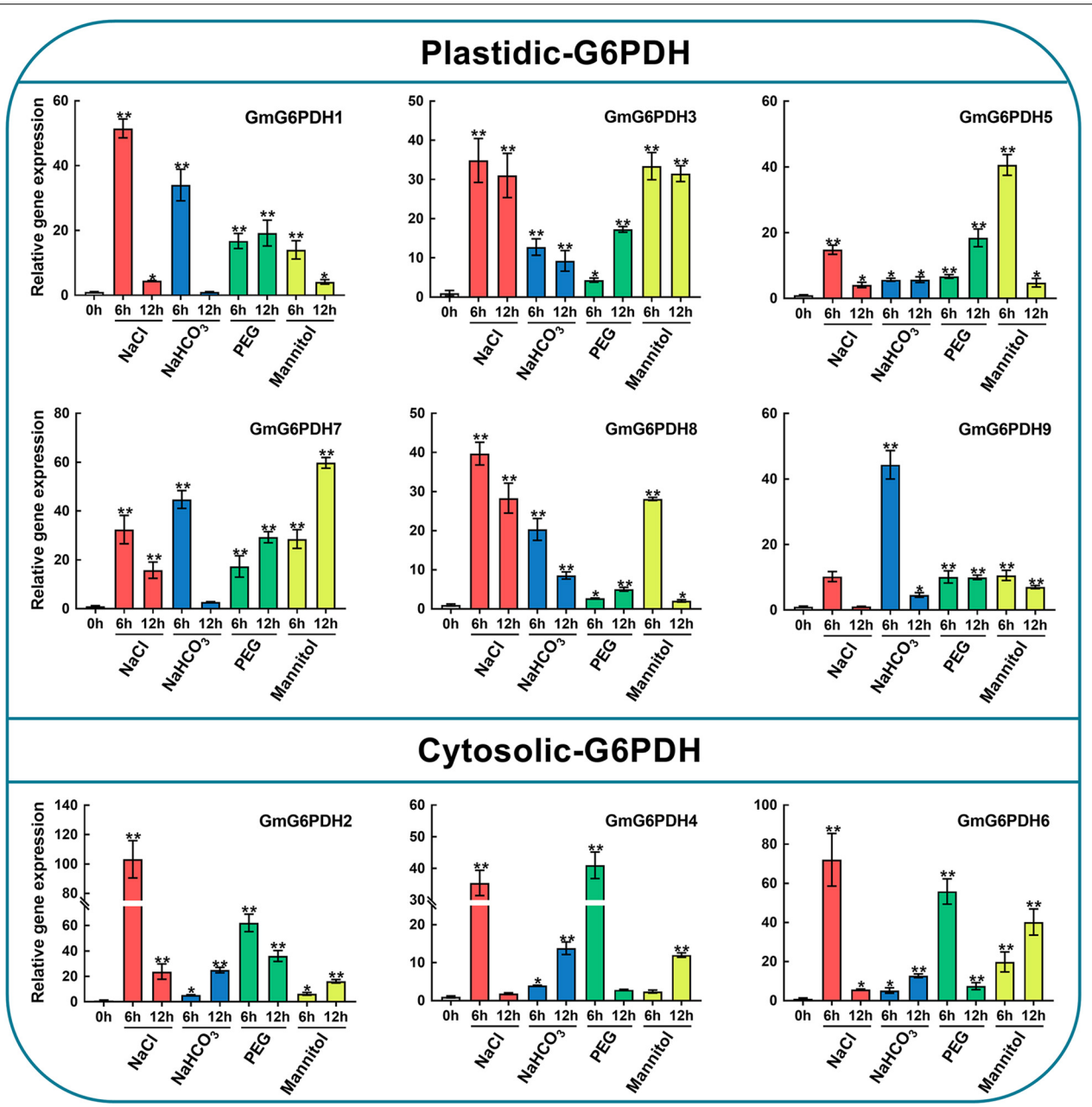


FIGURE 6 | Expression profiles of *GmG6PDHs* in leaves of soybean plants subjected to 150 mM NaCl, 100 mM $NaHCO_3$, 20% PEG, 200 mM mannitol or water (control) for 0, 6, and 12 h, respectively. The expression of *GmG6PDHs* in non-stress conditions was used as a calibrator. Three independent biological replicates were carried out and the results of qRT-PCR were analyzed using the $2^{-\Delta\Delta ct}$ method. Asterisks above bars denote a statistically significant difference by Student's *t*-test (*$P < 0.05$, **$P < 0.01$).

GmG6PDH2 Is Essential for AsA and GSH Biosynthesis Under Salt Stress

It has been reported that the plant G6PDHs perform significant functions in maintaining the cellular redox balance (Krüger et al., 2011; Cardi et al., 2016). To validate whether overexpression of cytosolic *GmG6PDH2* gene could affect the redox fluctuation, the redox states of AsA and GSH were monitored in soybean transgenic hair roots. The levels of reduced and oxidized AsA and GSH did not change among the CHR and OHR plants under standard growing conditions (**Figures 9A,B**). Nevertheless, the *GmG6PDH2*-OHR plants accumulated higher contents of AsA and GSH, and substantially lower GSH and GSSG contents than the CHR plants, thereby resulting in much higher ratios of AsA/DHA and GSH/GSSG. These results illustrated that the function of *GmG6PDH2* gene in salt tolerance could partly ascribe to the improved regeneration of the reduced form of AsA and GSH.

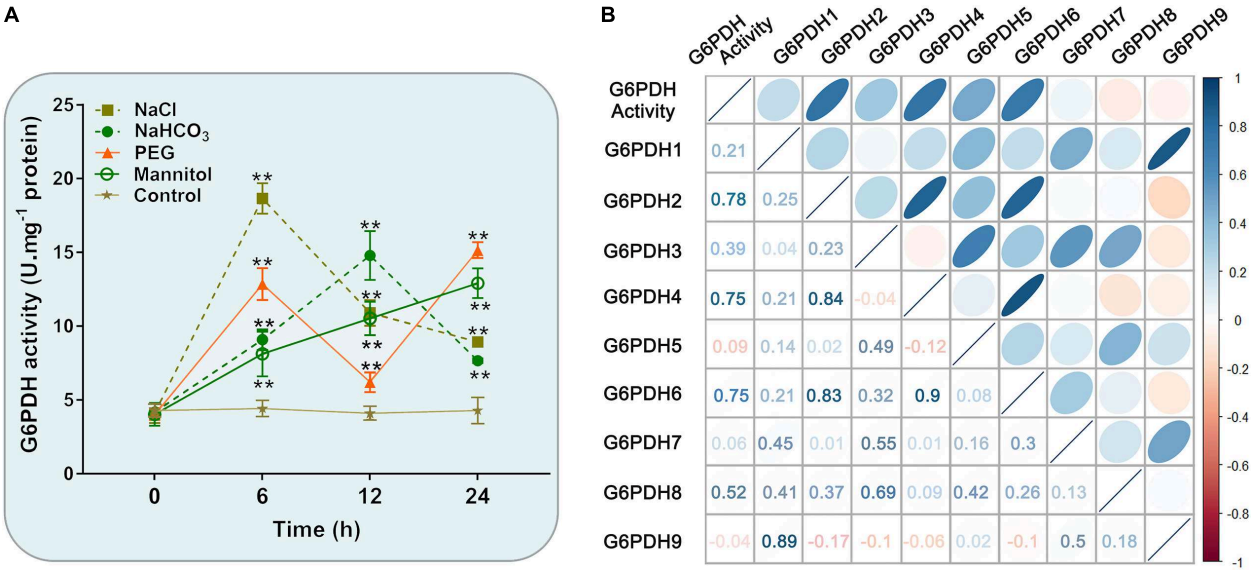


FIGURE 7 | Enzyme activity profiles of G6PDH in soybean under different abiotic stress treatments. **(A)** Analysis of G6PDH activity in leaves of soybean plants subjected to 150 mM NaCl, 100 mM $NaHCO_3$, 20% PEG, 200 mM mannitol or water (control) for 0, 6, 12, and 24 h, respectively. Data represent the means ± SD of three biological replicates. Asterisks above bars denote a statistically significant difference from the control group by Student's *t*-test (*$P < 0.05$; **$P < 0.01$). **(B)** Correlation coefficient between G6PDH enzymatic activity and the expression levels of *GmG6PDHs*. Each ellipse chart represents the correlation coefficient between any two traits. The color and slope of the ellipse represent magnitude of correlation. The ellipses of negative correlations are displayed in red and positive correlations in blue.

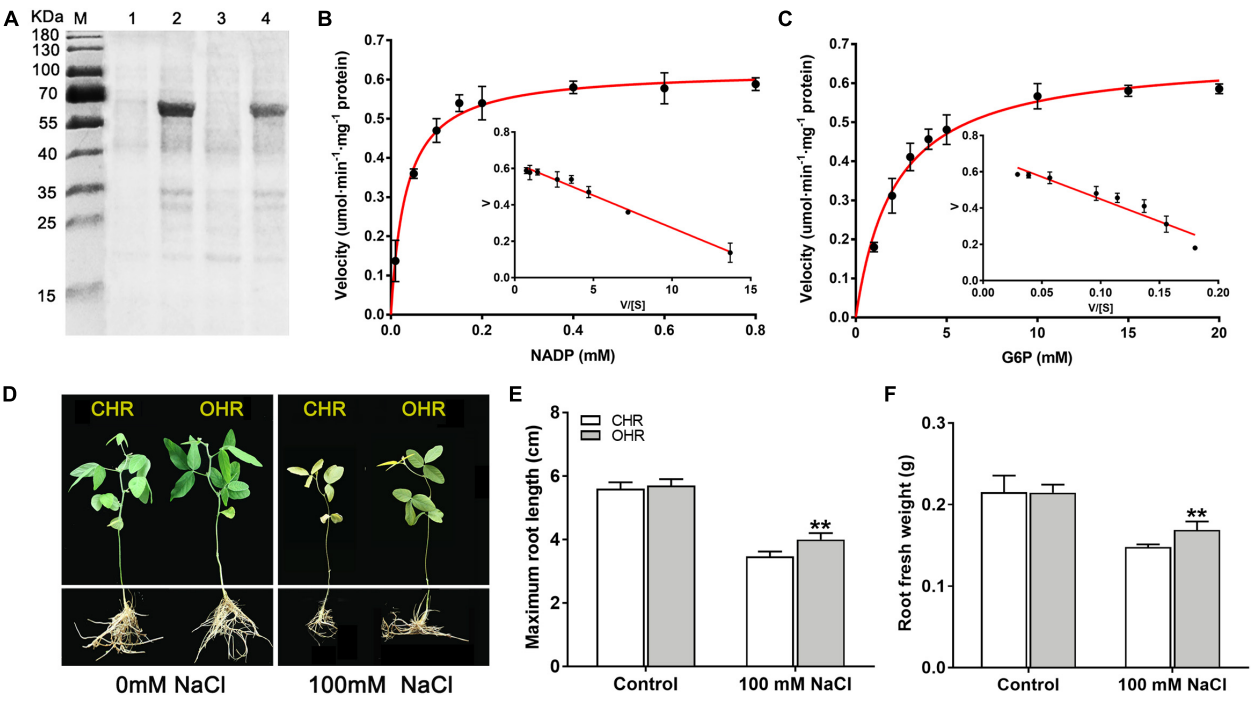


FIGURE 8 | Overexpression of cytosolic $NADP^+$-dependent isoform (GmG6PDH2) confers salt tolerance in transgenic soybean. **(A)** Coomassie-stained 12% SDS-PAGE of crude extracts of *Escherichia coli* cells harboring pET32a-GmG6PDH2 without (Lanes 1, 3) and with IPTG induction (Lanes 2, 4). The kinetic properties of *GmG6PDH2* with regard to **(B)** $NADP^+$ and **(C)** glucose-6-phosphate by Eadie–Hofstee plot. **(D)** Performance of soybean plants harboring *GmG6PDH2*-overexpressing hairy roots (OHR) and control hairy roots (CHR) treated with 0 or 100 mM NaCl for 5 days. CHR plants are generated after the infection of *Agrobacterium rhizogenes*. Comparisons of **(E)** maximum root length and **(F)** root fresh weight of CHR and OHR plants subjected to 0 or 100 mM NaCl for 5 days. More than 10 hair root lines were investigated and shown as means ± SD. Asterisks above bars denote a statistically significant difference from the CHR plants by Student's *t*-test (*$P < 0.05$, **$P < 0.01$).

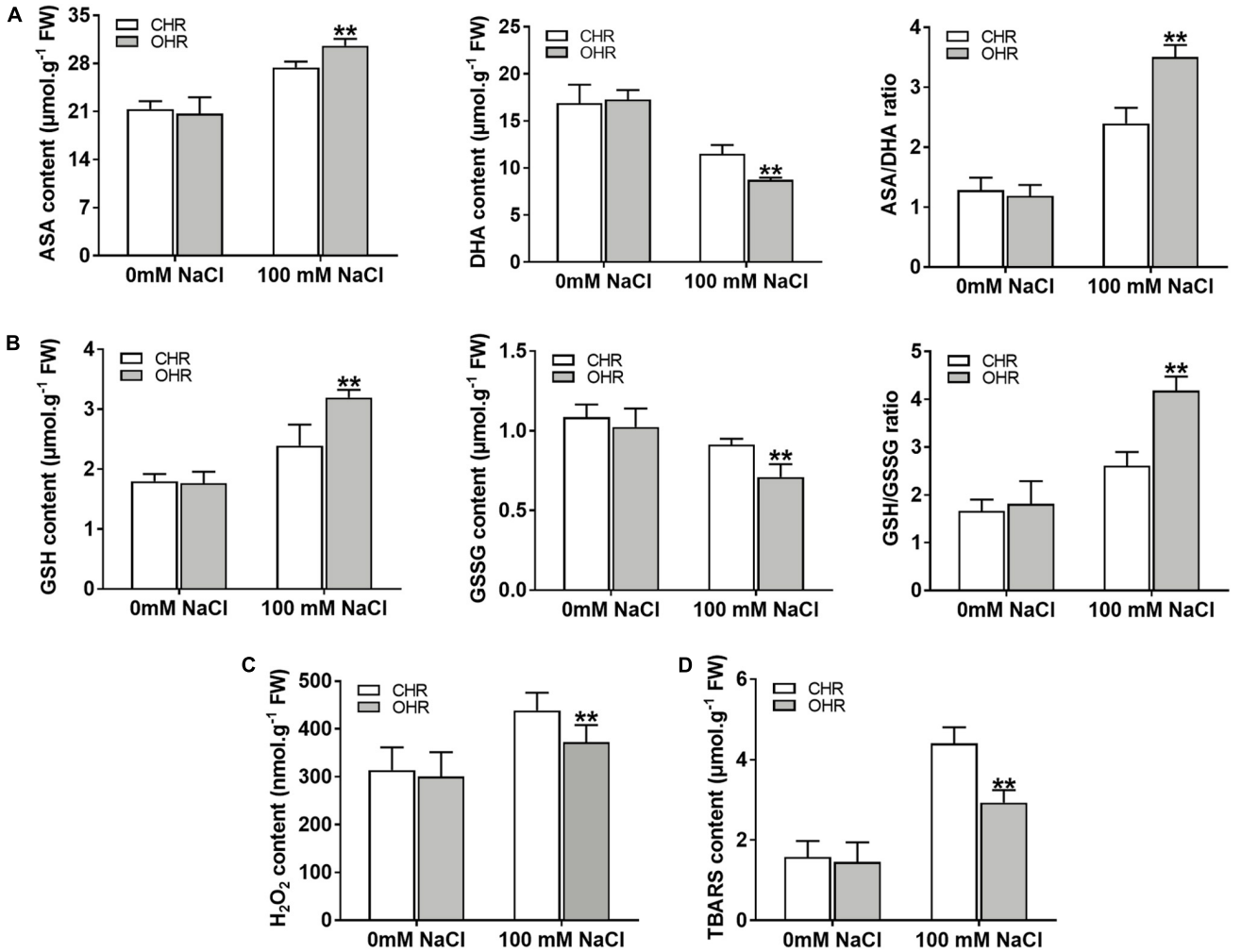


FIGURE 9 | Overexpression of *GmG6PDH2* improves the ROS-scavenging capacity under high salinity condition. **(A)** The contents of AsA and DHA, and the ratio of AsA/DHA; **(B)** The contents of GSH and GSSG, and the ratio of GSH/GSSG; **(C)** H_2O_2 and **(D)** TBARS levels in CHR and OHR plants subjected to 0 or 100 mM NaCl for 5 days. More than 10 independent transgenic hairy root lines were analyzed and shown as means ± SD. Asterisks above bars denote a statistically significant difference from the CHR plants by Student's *t*-test (*$P < 0.05$, **$P < 0.01$).

In plant stress reactions, the AsA-GSH cycle has been considered as a powerful ROS scavenging pathway, exclusively eliminating the cellular hydrogen peroxide (H_2O_2) through utilizing the AsA and GSH (Fotopoulos et al., 2010; Li et al., 2010). Thus, the remarkable changes in reduced and oxidized forms of ASA and GSH in *GmG6PDH2* transgenic lines reminded us to examine the H_2O_2 contents upon salt stress. As expected, the contents of H_2O_2 in *GmG6PDH2*-OHR lines were remarkably lower than that in the CHR plants, supporting an important role for *GmG6PDH2* in ROS detoxification under salt stress (**Figure 9C**). Moreover, the TBARS levels, as an indicator of lipid peroxidation, were evidently lower in OHR plants than that in the CHR plants under salt stress (**Figure 9D**). In conclusion, these results suggested that overexpression of cytosolic *GmG6PDH2* gene alleviated the stress-induced ROS accumulation through the regulation of the redox status of AsA and GSH, and consequently minimize the cell membrane damages under salt stress.

DISCUSSION

Soybean as the important economic food crop is one of the major sources of plant protein and oil for mankind, and its growth is largely affected by different abiotic stresses (Ghosh and Islam, 2016). A great number of researches have shown that *G6PDHs* play a critical role in plant growth and abiotic stress responses (Esposito, 2016; Hýsková et al., 2017). Although *G6PDH* genes have been cloned from several model organisms such tobacco (Scharte et al., 2009; Silva et al., 2018), *Arabidopsis* (Wakao and Benning, 2005), barley (Cardi et al., 2013; Caretto et al., 2015), wheat (Nemoto and Sasakuma, 2000) and tomato

(Landi et al., 2016), there is scarce information about their biological functions in soybean. In this study, we identified nine *G6PDH* gene family members from the soybean genome, named as *GmG6PDH1-9* (**Table 1**). Similar to other reported G6PDHs, all the GmG6PDH proteins contained the typical and necessary protein domains (PF00479, PF02781) (**Figure 1B**). Meanwhile, *GmG6PDH* genes have two highly conserved motifs in their translated protein sequences: $NADP^+$-binding motif (NEFVIRLQP) and substrate- binding motif (RIDHYLGKE) (**Supplementary Figure S1**), illustrating the proteins encoded by *GmG6PDH1-9* have enzyme catalytic activities (Hauschild and von Schaewen, 2003). According to the putative subcellular location, GmG6PDHs were segmented into two types: three cytosolic (Cy) isoforms (GmG6PDH2, 4, and 6), six plastidic (P) isoforms (GmG6PDH1, 3, 5, 7, 8, and 9), which was in accord with previous reports (Esposito et al., 2001; Wakao and Benning, 2005). Phylogeny analyses also resulted in the same classification of GmG6PDHs by dividing them into two clades (I and II), with clade I corresponding to P-G6PDHs and clade II corresponding to Cy-G6PDHs (**Figure 1A**). These results indicated that the divergence of G6PDHs into Cy and P types may be to ensure the plant adaptations to changing conditions.

All the *GmG6PDHs* were found having a close relationship with the corresponding genes in common bean, which is in agreement with their genetic relationships, suggesting possible functional conservation (**Figure 1B**). Additionally, we found that *GmG6PDHs* had collinear relationships with *S. bicolor*, *O. sativa*, *P. vulgaris* and *B. distachyon G6PDHs*, illustrating that *GmG6PDHs* might have occurred before the divergence of sorghum, rice, common bean and purple false brome grass lineages (**Figure 2**). The predicted of subcellular localization of plant G6PDHs are basically assigned dependent on the presence of signal peptide (Yu et al., 2004). In the present work, we assessed the subcellular localization of five soybean G6PDHs (GmG6PDH1, 2, 5, 8, and 9) by the transient expression of GFP-tagged recombinant proteins in *Arabidopsis* mesophyll protoplasts (**Figure 4**). The merged image of chloroplast autofluorescence and GFP indicated that the GmG6PDH1, 5, 8, and 9 fusion proteins were clearly located in the chloroplast, while the GmG6PDH2 protein was specially targeted to cytosol (**Figure 4**). The results correlated with the *in silico* prediction and phylogenetic clades for GmG6PDHs. The earlier identified AtG6PDH proteins were predicted to be cytosol or plastid-localized isoforms because of the absence or presence of pronounced targeting signal and transmembrane domains (Wakao and Benning, 2005), but a clear experimental evidence was lacking.

Using quantitative real-time qRT-PCR and high-throughput sequencing data analyses, we examined the transcriptional profiling of *GmG6PDHs* in different soybean tissues and at different times during seed development. Our data showed differences in the expression level of *GmG6PDHs* in distinct tissues. To be specific, the plastidic *GmG6PDHs* exhibited a tissue-specific transcript profiling with the high mRNA levels in green tissue, such as leaves and pods (**Figure 5A**), consistent with the expression profiles of their orthologous genes from *Hevea brasiliensis*, *A. thaliana,* and *Nicotiana tabacum* (Wakao and Benning, 2005; Gong et al., 2012; Long et al., 2016). By contrast, the expression analysis of cytosolic *GmG6PDH* genes revealed relatively high-abundant transcripts in developing seeds, roots and nodules (**Figures 5A,B**). In *Arabidopsis*, the Cy-G6PDH isoforms are also proved to be crucial in the metabolism of developing seeds (Wakao et al., 2008), suggesting that the cytosolic *G6PDH* genes mainly function in sorts of sink tissues. The tissue-specific expression characteristics indicate that *GmG6PDHs* may play versatile physiological roles in soybean development (**Figures 5A–C**).

In addition to regulating plant growth, the key functions of plant G6PDHs in stress- response mechanisms have also been widely proven (Esposito, 2016). The promoter regions of the *GmG6PDHs* contained lots of *cis*-acting elements that possibly participated in plant responses to drought and salt stresses, such as MBS, ABRE, ARE, and TCA-elements (**Figure 3**) (Liu et al., 2017). By qRT-PCR and enzyme activity assays, we further examined the physiological and transcriptional responses of GmG6PDHs to different stress conditions, including salt, alkali and osmotic, and significant induction in *GmG6PDHs* was observed under the above treatments, especially under salt stress (**Figures 6**, **7**). In addition, the different osmotic stresses induced by PEG6000 and mannitol have different effects on the transcription and expression of *GmG6PDHs*, probably due to the inhibitory effects of PEG on root oxygen availability (Guo et al., 2018). It has also been reported that the activities and transcripts of *AtG6PDHs* (Wakao and Benning, 2005), *HbG6PDHs* (Long et al., 2016), *ScG6PDHs* (Yang et al., 2014), and *PsG6PDHs* (Lin et al., 2005) are markedly stimulated by adverse environmental conditions, like oxygen availability, salinity, cold, and dehydration. Furthermore, the subcellular location of *GmG6PDHs* seemed to have certain impacts on their stress responses, as represented by the higher average mRNA expression levels of cytosolic *GmGPDH2*, *4* and *6* in comparison with plastidic *GmG6PDHs* under various stresses. Meanwhile, it is worthwhile to note that a cytosolic isoform (*GmG6PDH2*) respond faster and more vigorously to salinity stress than other genes, reaching its maximum mRNA level (about 100-fold) within 6 h of salinity treatment. The average mRNA abundances of *GmGPDH2* under salt stress was much higher than that of other genes, and its expression patterns correlated well with the activity of G6PDH enzyme (**Figure 7**), implying a major role in response to salinity. Same results have been available earlier in other plants: a cytosolic G6PDH encoded by *WESR5* are proved to be important for early salt responding in wheat (Nemoto and Sasakuma, 2000, 2002); and a cytosolic *ScG6PDH* also plays a positive role in salt-induced responses in sugarcane, but the corresponding functional verification is lacking (Yang et al., 2014).

In this study, the underlying molecular metabolism of the cytosolic *GmG6PDH2* in mediating salinity adaption was further studied. Enzymatic properties analysis of recombinant GmG6PDH2 proteins expressed in *E. coli* (**Figure 8A**) showed that the protein encoded by *GmG6PDH2* gene had substrate affinity (K_m^{G6P} of 2.138 mM) (**Figure 8C**), and a good agreement

was achieved compared to the enzymatic characteristics available in previously published G6PDHs (Hauschild and von Schaewen, 2003; Scharte et al., 2009). In addition, overexpression of *GmG6PDH2* dramatically enhanced the salt tolerance in transgenic soybean plants, as reflected by a noticeable increase in root length and root fresh weight (**Figures 8E,F**). Cytosolic G6PDHs have been previously proved to participate in the modulation of cellular redox states by supply for NADPH (Xu et al., 2003; Valderrama et al., 2006). In the present work, an apparent elevation in the levels of antioxidants (AsA and GSH) as well as the ratios of AsA/DHA and GSH/GSSG were seen in *GmG6PDH2*-OHR plants related to the CHR plants (**Figures 9A,B**). These data validate that *GmG6PDH2* plays a pivotal role in modulating the changes of metabolite contents and the redox state of AsA and GSH pool under salinity condition, which was consistent with previous results (Krüger et al., 2011; Wang et al., 2016). It is known that the AsA-GSH cycle has a main function of eliminating H_2O_2, a potentially harmful ROS (Fotopoulos et al., 2010; Li et al., 2010). In agreement with this, we found that the levels of H_2O_2 were obviously lower in *GmG6PDH2* OHRs than in control plants (**Figure 9C**). In addition, the cellular membrane damage in *GmG6PDH2*-OHR plants was less severe as manifested by the reduction in TBARS content under salt stress (**Figure 9D**), suggesting that overexpression of *GmG6PDH2* alleviate the salt-induced ROS accumulation and consequently minimize the membrane lipid peroxidation. In summary, the cytosolic *GmG6PDH2* gene has profound effects on salt tolerance by increasing components of the AsA-GSH cycle which subsequently may act to prevent oxidative stress caused by salinity.

CONCLUSION

Nine soybean *G6PDH* genes were characterized in the soybean genome. Based on the subcellular localization and phylogenetic analysis, *GmG6PDHs* were divided into plastidic (P) and cytosolic (Cy) isoforms. The *GmG6PDH* genes had distinct expression patterns under various abiotic stresses, reflected the potential functional distinction of each isoform. Of the nine *GmG6PDH* genes, the cytosol-localized GPDH (*GmG6PDH2*) gene had an apparent transcriptional response to salinity, and the expression of *GmG6PDH2* showed a high correlation with the G6PDH activity, suggesting a principal participant in response to salinity. Further study indicated that *GmG6PDH2* gene encode active G6PDH enzyme, which was coordinated with AsA-GSH cycle to maintain the redox state of AsA and GSH and consequently minimize the NaCl-induced oxidative damages.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YZ and YC designed and conceived the experiments. YZ, YC, SH, and JY performed the experiments. XW, DX, XL, YL, and YD analyzed the data and interpreted the results. YZ prepared the manuscript. ZQ and QC conceived the experiments and revised the manuscript. All authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved and approved the final version to be published.

FUNDING

This study was financially supported by EUCLEG (727312; 2017YFE0111000), National Key R&D Program of China (2016YFD0100500, 2016YFD0100300, and 2016YFD0100201-21), Post-Doctoral Project of Northeast Agricultural University (NEAUBH-19002), National Natural Science Foundation of China (31701449, 31971968, 31971899, and 31501332), Natural Science Foundation of Heilongjiang (QC2017013), Special Financial Aid to Post-doctor Research Fellow in Heilongjiang (LBH-TZ1714), Heilongjiang Academy of Agricultural Sciences Funds (2019YYYF 019), International Post-doctoral Exchange Fellowship Program of China Post-doctoral Council (20180004), China Post-Doctoral Project (2015M581419), Heilongjiang Funds for Distinguished Young Scientists (JC2016004 and JC2017006), Dongnongxue zhe Project (to QC), Backbone of Young Talent scholar Project (to ZQ, 18XG01) of Northeast Agricultural University.

ACKNOWLEDGMENTS

We would like to thank Dr. Jinhui Wang (Northeast Agricultural University, China) for providing us soybean seeds.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00214/full#supplementary-material

REFERENCES

Cardi, M., Castiglia, D., Ferrara, M., Guerriero, G., Chiurazzi, M., and Esposito, S. (2015). The effects of salt stress cause a diversion of basal metabolism in barley roots: possible different roles for glucose-6-phosphate dehydrogenase isoforms. *Plant Physiol. Biochem.* 86, 44–54. doi: 10.1016/j.plaphy.2014.11.001

Cardi, M., Chibani, K., Castiglia, D., Cafasso, D., Pizzo, E., Rouhier, N., et al. (2013). Overexpression, purification and enzymatic characterization of a recombinant plastidial glucose-6-phosphate dehydrogenase from barley (*Hordeum vulgare* cv. Nure) roots. *Plant Physiol. Biochem.* 73, 266–273. doi: 10.1016/j.plaphy.2013.10.008

Cardi, M., Zaffagnini, M., De Lillo, A., Castiglia, D., Chibani, K., Gualberto, J. M., et al. (2016). Plastidic P2 glucose-6P dehydrogenase from poplar is modulated by thioredoxin m-type: distinct roles of cysteine residues in redox regulation and NADPH inhibition. *Plant Sci.* 252, 257–266. doi: 10.1016/j.plantsci.2016.08.003

Caretto, S., Linsalata, V., Colella, G., Mita, G., and Lattanzio, V. (2015). Carbon fluxes between primary metabolism and phenolic pathway in plant tissues under stress. *Int. J. Mol. Sci.* 16, 26378–26394. doi: 10.3390/ijms161125967

Castiglia, D., Cardi, M., Landi, S., Cafasso, D., and Esposito, S. (2015). Expression and characterization of a cytosolic glucose-6-phosphate dehydrogenase isoform from barley (*Hordeum vulgare*) roots. *Protein Expr. Purif.* 112, 8–14. doi: 10.1016/j.pep.2015.03.016

Chen, C., Chen, H., He, Y., and Xia, R. (2018). TBtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface. *BioRxiv* [Preprint]. doi: 10.1101/289660

Dal Santo, S., Stampfl, H., Krasensky, J., Kempa, S., Gibon, Y., Petutschnig, E., et al. (2012). Stress-induced GSK3 regulates the redox stress response by phosphorylating glucose-6- phosphate dehydrogenase in *Arabidopsis*. *Plant Cell* 24, 3380–3392. doi: 10.1105/tpc.112.101279

Devi, R., Munjral, N., Gupta, A. K., and Kaur, N. (2007). Cadmium induced changes in carbohydrate status and enzymes of carbohydrate metabolism, glycolysis and pentose phosphate pathway in pea. *Environ. Exp. Bot.* 61, 167–174. doi: 10.1016/j.envexpbot.2007.05.006

Esposito, S. (2016). Nitrogen assimilation, abiotic stress and glucose 6-phosphate dehydrogenase: the full circle of reductants. *Plants* 5:E24. doi: 10.3390/plants5020024

Esposito, S., Carfagna, S., Massaro, G., Vona, V., and Rigano, V. D. M. (2001). Glucose-6-phosphate dehydrogenase in barley roots: kinetic properties and localisation of the isoforms. *Planta* 212, 627–634. doi: 10.1007/s004250000443

Esposito, S., Guerriero, G., Vona, V., Di Martino Rigano, V., Carfagna, S., and Rigano, C. (2004). Glutamate synthase activities and protein changes in relation to nitrogen nutrition in barley: the dependence on different plastidic glucose-6P dehydrogenase isoforms. *J. . Bot.* 56, 55–64.

Fotopoulos, V., Ziogas, V., Tanou, G., and Molassiotis, A. (2010). "Involvement of AsA/DHA and GSH/GSSG ratios in gene and protein expression and in the activation of defence mechanisms under abiotic stress conditions," in *Ascorbate-Glutathione Pathway and Stress Tolerance in Plants* (Berlin: Springer), 265–302. doi: 10.1007/978-90-481-9404-9_10

Ghosh, A., and Islam, T. (2016). Genome-wide analysis and expression profiling of glyoxalase gene families in soybean (*Glycine max*) indicate their development and abiotic stress specific response. *BMC Plant Biol.* 16:87. doi: 10.1186/s12870-016-0773-9

Gong, H., Chen, G., Li, F., Wang, X., Hu, Y., and Bi, Y. (2012). Involvement of G6PDH in heat stress tolerance in the calli from *Przewalskia tangutica* and *Nicotiana tabacum*. *Biol. Plant.* 56, 422–430. doi: 10.1007/s10535-012-0072-8

Guo, Y. Y., Tian, S. S., Liu, S. S., Wang, W. Q., and Sui, N. (2018). Energy dissipation and antioxidant enzyme system protect photosystem II of sweet sorghum under drought stress. *Photosynthetica* 56, 861–872. doi: 10.1007/s11099-017-0741-0

Hauschild, R., and von Schaewen, A. (2003). Differential regulation of glucose-6-phosphate dehydrogenase isoenzyme activities in potato. *Plant Physiol.* 133, 47–62. doi: 10.1104/pp.103.025676

Hodges, D. M., DeLong, J. M., Forney, C. F., and Prange, R. K. (1999). Improving the thiobarbituric acid-reactive-substances assay for estimating lipid peroxidation in plant tissues containing anthocyanin and other interfering compounds. *Planta* 207, 604–611. doi: 10.1007/s004250050524

Honjoh, K., Machida, T., Hagisako, T., Suga, K., Yonekura, M., Shimizu, H., et al. (2007). Molecular cloning and characterization of a cDNA for low-temperature inducible cytosolic glucose 6-phosphate dehydrogenase gene from *Chlorella vulgaris* and expression of the gene in *Saccharomyces cerevisiae*. *Plant Sci.* 172, 649–658. doi: 10.1016/j.plantsci.2006.12.004

Hýsková, V., Pliskova, V., Červený, V., and Ryšlavá, H. (2017). NADP-dependent enzymes are involved in response to salt and hypoosmotic stress in cucumber plants. *Gen. Physiol Biophys.* 36, 247–258. doi: 10.4149/gpb_2016053

Krüger, A., Grüning, N. M., Wamelink, M. M., Kerick, M., Kirpy, A., Parkhomchuk, D., et al. (2011). The pentose phosphate pathway is a metabolic redox sensor and regulates transcription during the antioxidant response. *Antioxid. Redox Signal.* 15, 311–324. doi: 10.1089/ars.2010.3797

Landi, S., Nurcato, R., De Lillo, A., Lentini, M., Grillo, S., and Esposito, S. (2016). Glucose-6- phosphate dehydrogenase plays a central role in the response of tomato (*Solanum lycopersicum*) plants to short and long-term drought. *Plant Physiol. Biochemi.* 105, 79–89. doi: 10.1016/j.plaphy.2016.04.013

Lee, T. H., Tang, H., Wang, X., and Paterson, A. H. (2012). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res.* 41, D1152–D1158. doi: 10.1093/nar/gks1104

Lejay, L., Wirth, J., Pervent, M., Cross, J. M. F., Tillard, P., and Gojon, A. (2008). Oxidative pentose phosphate pathway-dependent sugar sensing as a mechanism for regulation of root ion transporters by photosynthesis. *Plant Physiol.* 146, 2036–2053. doi: 10.1104/pp.107.114710

Li, Y., Liu, Y., and Zhang, J. (2010). Advances in the research on the AsA-GSH cycle in horticultural crops. *Front. Agricu. China* 4:84. doi: 10.1007/s11703-009-0089-8

Lin, Y., Lin, S., Guo, H., Zhang, Z., and Chen, X. (2013). Functional analysis of *PsG6PDH*, a cytosolic glucose-6-phosphate dehydrogenase gene from *Populus suaveolens*, and its contribution to cold tolerance improvement in tobacco plants. *Biotechnol Lett.* 35, 1509–1518. doi: 10.1007/s10529-013-1226-2

Lin, Y. Z., Lin, S. Z., Zhang, W., Zhang, Q., Zhang, Z. Y., Guo, H., et al. (2005). Cloning and sequence analysis of a glucose-6-phosphate dehydrogenase gene *PsG6PDH* from freezing-tolerant *Populus suaveolens*. *For. Stud. China* 7, 1–6. doi: 10.1007/s11632-005-0048-2

Liu, J., Wang, X., Hu, Y., Hu, W., and Bi, Y. (2013). Glucose-6-phosphate dehydrogenase plays a pivotal role in tolerance to drought stress in soybean roots. *Plant Cell Rep.* 32, 415–429. doi: 10.1007/s00299-012-1374-1

Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., et al. (2015). IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 31, 3359–3361. doi: 10.1093/bioinformatics/btv362

Liu, Y., Wu, R., Wan, Q., Xie, G., and Bi, Y. (2007). Glucose-6-phosphate dehydrogenase plays a pivotal role in nitric oxide-involved defense against oxidative stress under salt stress in red kidney bean roots. *Plant Cell Physiol.* 48, 511–522. doi: 10.1093/pcp/pcm020

Liu, Z., Ge, X., Yang, Z., Zhang, C., Zhao, G., Chen, E., et al. (2017). Genome-wide identification and characterization of *SnRK2* gene family in cotton (*Gossypium hirsutum* L.). *BMC Genet.* 18:54. doi: 10.1186/s12863-017-0517-3

Long, X., He, B., Fang, Y., and Tang, C. (2016). Identification and characterization of the glucose-6- phosphate dehydrogenase gene family in the para rubber tree. Hevea brasiliensis. *Front. Plant Sci.* 7:215. doi: 10.3389/fpls.2016.00215

Née, G., Zaffagnini, M., Trost, P., and Issakidis-Bourguet, E. (2009). Redox regulation of chloroplastic glucose-6-phosphate dehydrogenase: a new role for f-type thioredoxin. *FEBS Lett.* 583, 2827–2832. doi: 10.1016/j.febslet.2009.07.035

Nemoto, Y., and Sasakuma, T. (2000). Specific expression of glucose-6-phosphate dehydrogenase (G6PDH) gene by salt stress in wheat (*Triticum aestivum* L.). *Plant Sci.* 158, 53–60. doi: 10.1016/s0168-9452(00)00305-8

Nemoto, Y., and Sasakuma, T. (2002). Differential stress responses of early salt-stress responding genes in common wheat. *Phytochemistry* 61, 129–133. doi: 10.1016/s0031-9422(02)00228-5

Pan, W. J., Tao, J. J., Cheng, T., Bian, X. H., Wei, W., Zhang, W. K., et al. (2016). Soybean *miR172a* improves salt tolerance and can function as a long-distance signal. *Mol. Plant* 9, 1337–1340. doi: 10.1016/j.molp.2016.05.010

Sang, S., Xie, L., Cui, X., Wang, N., Gao, M., and Wang, Z. (2018). Glucose-6-phosphate dehydrogenase is required for hpa1xoo (harpin protein fragment)-mediated salt stress tolerance in transgenic *Arabidopsis thaliana*. *Pak. J. Bot.* 50, 841–845.

Scharte, J., Schön, H., Tjaden, Z., Weis, E., and von Schaewen, A. (2009). Isoenzyme replacement of glucose-6-phosphate dehydrogenase in the cytosol improves stress tolerance in plants. *Proc. Natl. Acad. S. U.S.A.* 106, 8061–8066. doi: 10.1073/pnas.0812902106

Sharkey, T. D., and Weise, S. E. (2016). The glucose 6-phosphate shunt around the Calvin-Benson cycle. *J. Exp. Bot.* 67, 4067–4077. doi: 10.1093/jxb/erv484

Silva, F. L. B., Vieira, L. G. E., Ribas, A. F., Moro, A. L., Neris, D. M., and Pacheco, A. C. (2018). Proline accumulation induces the production of total phenolics in transgenic tobacco plants under water deficit without increasing the G6PDH activity. *Theor. Exp. Plant Physiol.* 30, 251–260. doi: 10.1007/s40626-018-0119-0

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tóth, K., Batek, J., and Stacey, G. (2016). Generation of soybean (Glycine max) transient transgenic roots. *Curr. Protoc. Plant Biol.* 1, 1–13. doi: 10.1002/cppb.20017

Valderrama, R., Corpas, F. J., Carreras, A., Gómez-Rodríguez, M. V., Chaki, M., Pedrajas, J. R., et al. (2006). The dehydrogenase-mediated recycling of NADPH is a key antioxidant system against salt-induced oxidative stress in olive plants. *PlantCell Environ.* 29, 1449–1459. doi: 10.1111/j.1365-3040.2006.01530.x

Velikova, V., Yordanov, I., and Edreva, A. (2000). Oxidative stress and some antioxidant systems in acid rain-treated bean plants: protective role of exogenous polyamines. *Plant Sci.* 151, 59–66. doi: 10.1016/s0168-9452(99)00197-1

Wakao, S., Andre, C., and Benning, C. (2008). Functional analyses of cytosolic glucose-6-phosphate dehydrogenases and their contribution to seed oil accumulation in *Arabidopsis*. *Plant Physiol.* 146, 277–288. doi: 10.1104/pp.107.108423

Wakao, S., and Benning, C. (2005). Genome-wide analysis of glucose-6-phosphate dehydrogenases in *Arabidopsis*. *Plant J.* 41, 243–256. doi: 10.1111/j.1365-313x.2004.02293.x

Wang, H., Yang, L., Li, Y., Hou, J., Huang, J., and Liang, W. (2016). Involvement of ABA-and H2O2-dependent cytosolic glucose-6-phosphate dehydrogenase in maintaining redox homeostasis in soybean roots under drought stress. *Plant Physiol. Biochem.* 107, 126–136. doi: 10.1016/j.plaphy.2016.05.040

Wang, X., Ma, Y., Huang, C., Li, J., Wan, Q., and Bi, Y. (2008). Involvement of glucose-6-phosphate dehydrogenase in reduced glutathione maintenance and hydrogen peroxide signal under salt stress. *Plant Signal. Behav.* 3, 394–395. doi: 10.4161/psb.3.6.5404

Wendt, U. K., Wenderoth, I., Tegeler, A., and Von Schaewen, A. (2000). Molecular characterization of a novel glucose-6-phosphate dehydrogenase from potato (*Solanum tuberosum* L.). *Plant J.* 23, 723–733. doi: 10.1046/j.1365-313x.2000.00840.x

Wood, T. (1986). Physiological functions of the pentose phosphate pathway. *Cell Biochem. Funct.*4, 241–247. doi: 10.1002/cbf.290040403

Xu, J., Maki, D., and Stapleton, S. R. (2003). Mediation of cadmium-induced oxidative damage and glucose-6-phosphate dehydrogenase expression through glutathione depletion. *J. Biochem. Mo,. Toxicol.* 17, 67–75. doi: 10.1002/jbt.10062

Yang, Y., Fu, Z., Su, Y., Zhang, X., Li, G., Guo, J., et al. (2014). A cytosolic glucose-6-phosphate dehydrogenase gene, *ScG6PDH*, plays a positive role in response to various abiotic stresses in sugarcane. *Sci. Rep.* 4:7090. doi: 10.1038/srep07090

Yoo, S. D., Cho, Y. H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* 2:1565. doi: 10.1038/nprot.2007.199

Yu, C. S., Lin, C. J., and Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram- negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.* 13, 1402–1406. doi: 10.1110/ps.03479604

Zhang, L., Liu, J., Wang, X., and Bi, Y. (2013). Glucose-6-phosphate dehydrogenase acts as a regulator of cell redox balance in rice suspension cells under salt stress. *Plant Growth Regul.* 69, 139–148. doi: 10.1007/s10725-012-9757-4

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 Zhao, Cui, Huang, Yu, Wang, Xin, Li, Liu, Dai, Qi and Chen. 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 Plant Science

ORIGINAL RESEARCH
published: 19 May 2020
doi: 10.3389/fpls.2020.00453

QTL Mapping and Data Mining to Identify Genes Associated With the *Sinorhizobium fredii* HH103 T3SS Effector NopD in Soybean

Jinhui Wang[1†], Jieqi Wang[1†], Chao Ma[1†], Ziqi Zhou[1], Decheng Yang[1], Junzan Zheng[1], Qi Wang[1], Huiwen Li[1], Hongyang Zhou[1], Zhijun Sun[1], Hanxi Liu[1], Jianyi Li[1], Lin Chen[1], Qinglin Kang[1], Zhaoming Qi[1], Hongwei Jiang[2], Rongsheng Zhu[1], Xiaoxia Wu[1], Chunyan Liu[1], Qingshan Chen[1]* and Dawei Xin[1]**

[1] Key Laboratory of Soybean Biology of Chinese Ministry of Education, Key Laboratory of Soybean Biology and Breeding/Genetics of Chinese Agriculture Ministry, College of Agriculture, Northeast Agricultural University, Harbin, China, [2] Jilin Academy of Agricultural Sciences, Changchun, China

In some legume–rhizobium symbioses, host specificity is influenced by rhizobial type III effectors-nodulation outer proteins (Nops). However, the genes encoding host proteins that interact with Nops remain unknown. In this study, we aimed to identify candidate soybean genes associated with NopD, one of the type III effectors of *Sinorhizobium fredii* HH103. The results showed that the expression pattern of NopD was analyzed in rhizobia induced by genistein. We also found NopD can be induced by TtsI, and NopD as a toxic effector can induce tobacco leaf death. In 10 soybean germplasms, NopD played a positively effect on nodule number (NN) and nodule dry weight (NDW) in nine germplasms, but not in Kenjian28. Significant phenotype of NN and NDW were identified between Dongnong594 and Charleston, Suinong14 and ZYD00006, respectively. To map the quantitative trait locus (QTL) associated with NopD, a recombinant inbred line (RIL) population derived from the cross between Dongnong594 and Charleston, and chromosome segment substitution lines (CSSLs) derived from Suinong14 and ZYD00006 were used. Two overlapping conditional QTL associated with NopD on chromosome 19 were identified. Two candidate genes were identified in the confident region of QTL, we found that NopD could influence the expression of *Glyma.19g068600* (FBD/LRR) and expression of *Glyma.19g069200* (PP2C) after HH103 infection. Haplotype analysis showed that different types of *Glyma.19g069200* haplotypes could cause significant nodule phenotypic differences, but *Glyma.19g068600* (FBD/LRR) was not. These results suggest that NopD promotes *S. fredii* HH103 infection *via* directly or indirectly regulating *Glyma.19g068600* and *Glyma.19g069200* expression during the establishment of symbiosis between rhizobia and soybean plants.

Keywords: symbiosis, T3SS effector, NopD, quantitative trait locus (QTL), haplotype, soybean

OPEN ACCESS

Edited by:
Deyue Yu,
Nanjing Agricultural University, China

Reviewed by:
Zhihui Shan,
Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, China
Dan Zhang,
Henan Agricultural University, China

***Correspondence:**
Chunyan Liu
cyliucn@126.com
Qingshan Chen
qshchen@126.com
Dawei Xin
xdawei@163.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *13 December 2019*
Accepted: *27 March 2020*
Published: *19 May 2020*

Citation:
Wang J, Wang J, Ma C, Zhou Z, Yang D, Zheng J, Wang Q, Li H, Zhou H, Sun Z, Liu H, Li J, Chen L, Kang Q, Qi Z, Jiang H, Zhu R, Wu X, Liu C, Chen Q and Xin D (2020) QTL Mapping and Data Mining to Identify Genes Associated With the Sinorhizobium fredii HH103 T3SS Effector NopD in Soybean. Front. Plant Sci. 11:453. doi: 10.3389/fpls.2020.00453

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] is a widely grown commercial crop around the world and supplies a large amount of protein and oil for humans and animals (Xin et al., 2016). Nitrogen is an indispensable element for soybean growth and an important limiting factor in crop production. Nowadays, huge amounts of nitrogen fertilizers are applied to improve crop production, but nitrogen fertilizers can also cause negative effects, such as soil acidification, change of soil microbial diversity, soil compaction, and groundwater pollution (Vance, 2001). Biological nitrogen fixation (BNF) could sustainably supply large amounts of nitrogen for agricultural production and could reduce the application of synthetic nitrogen fertilizer (Ladha and Peoples, 2012). Legumes can recognize and accept various strains of *Rhizobium* to establish a symbiotic relationship; numerous different strains are present in different ecoregions (Zimmer et al., 2016). The recognition and acceptance of rhizobia by legumes are a complex process. Secretion of Nod factor from rhizobia induces the curling of root hair tips, rhizobial cells are wrapped by the curling tips, and then start to infect host cells. Rhizobia induce the development of infection threads in host cells by means of which rhizobia can be transported into the root cortical cells (Riely et al., 2013). When rhizobia symbiotically colonize soybean roots, the plants can fix atmospheric nitrogen (Riches et al., 2013). *Sinorhizobium fredii* strain HH103 can nodulate soybean efficiently, is a fast-growing rhizobia similar as model strain *S. fredii* NGR234. In recent years, the genome of *S. fredii* HH103 has been uncovered, and extensive analyses of its genome and transcriptome have paved a good foundation for gene functional characterization (Margaret et al., 2011; Weidner et al., 2012; Vinardell et al., 2015; López-Baena et al., 2016; Pérez-Montaño et al., 2016). Thus, it is an ideal strain for studying the molecular mechanisms symbiosis between soybean and rhizobium.

The establishment of an effective symbiotic interaction is a complex process that requires multiple signal exchanges between the legume and rhizobia (Miwa and Okazaki, 2017). Among those signals, type III effectors (T3E) play vital roles during the infection of the host with rhizobia. T3Es are secreted through the type 3 secretion system (T3SS) and are translocated into host cells. Within the host cells, the effectors change host signaling, including suppressing plant immunity systems and supplying a more favorable environment for rhizobial infection and multiplication. Similar to plant pathogen effectors, some rhizobial T3Es can also induce strong defense responses that suppress rhizobial infection after being recognized by host legume resistance proteins (Marie et al., 2003; Tampakaki, 2014; Staehelin and Krishnan, 2015; López-Baena et al., 2016). These findings support that T3Es from rhizobia could have either positive or negative influences on the establishment of symbioses. To date, 12 nodulation outer proteins (Nops), namely, NopA, NopAA (GunA), NopB, NopC, NopD, NopI, NopJ, NopL, NopM, NopP, NopT, and NopX, have been identified in *S. fredii* strain HH103 (López-Baena et al., 2016; Jiménez-Guerrero et al., 2019). Among these effectors, NopA and NopB are important components of the needle of the T3SS (Saad et al., 2005; Kim and Krishnan, 2014). NopAA (GunA) is a cellulase that is able to break down the soybean cell wall and so promote infection (Jiménez-Guerrero et al., 2019). NopAA increased *GmPR1* expression at an early stage of symbiosis (Jiménez-Guerrero et al., 2019). NopC is secreted into soybean root cells, exerting a positive function during infection (Jiménez-Guerrero et al., 2015a). NopM, a NEL-domain E3 ubiquitin ligase, appears to induce target sumoylation and may dampen the flg22-induced burst in reactive oxygen species in tobacco (Xin et al., 2012). An HH103 T3SS mutant that failed to secrete T3Es altered the expression of *GmPR1*, suggesting that the T3E might be related to the defense response (Jiménez-Guerrero et al., 2015b). NopP is a substrate for plant kinases, and its secretion by strain USDA112 was associated with host effector-triggered immunity to regulate symbiotic incompatibility with *Rj2* soybeans (Sugawara et al., 2018). Different effectors might have various functions during the establishment of symbiosis.

Rhizobial TtsI can regulate Nops expression during rhizobial infection. In a TtsI mutant, Nops expression was clearly suppressed (López Baena et al., 2008). NopD was first detected in culture supernatants of *S. fredii* strain HH103 induced by genistein (Rodrigues et al., 2007) and was regulated by TtsI (López-Baena et al., 2016). NopD showed homology to Blr1693, a putative outer protein of *Bradyrhizobium japonicum*. The C-terminal region of Blr1693 contains a domain with homology to the ubiquitin-like protease Ulp1 (Kaneko et al., 2002). XopD, one of the *Xanthomonas campestris pv. vesicatoria* T3E, belongs to the C48 cysteine peptidase family and encodes a ubiquitin-like protease 1 (Ulp1). Interestingly, bioinformatic analysis showed that the C-terminal region of *S. fredii* HH103 NopD shares sequence similarities with Blr1693 and XopD. XopD interacts with a small ubiquitin-like modifier (SUMO)-conjugated protein and removes the SUMO conjugate in plants during *X. campestris pv. vesicatoria* infection (Hotson et al., 2003), suggesting that NopD protease might similarly cleave SUMO modifications from SUMO-conjugated proteins. Besides desumoylation, XopD can play a role in host plant defense by interacting with the transcription factor MYB30 (Canonne et al., 2011). In tomato, XopD targeted the transcription factor SlERF4. This interaction influenced signaling in response to ethylene and promoted pathogen reproduction (Kim et al., 2013). *S. fredii* HH103 NopD is one of the Ulp1 proteins with similarities to XopD from *Xanthomonas*. NopD might influence the host cell signaling pathway in a similar way to XopD. However, no proteins that directly interact with NopD have yet been reported.

Numerous important traits in crops have been studied using quantitative trait locus (QTL) mapping to identify genes related to the target traits. Traits related to nodulation are controlled by various genes, and QTL mapping has been used to identify loci or genes associated with symbiosis (Hwang et al., 2014). Several loci related to nodulation have been mapped, such as *rj1, rj2, rj3, rj4, rj5, rj6, rj7,* and *rj8* (Caldwell, 1966; Vest, 1970; Vest and Caldwell, 1972; Caetano-Anollés and Gresshoff, 1991; Vuong et al., 1996). Among these loci, *rj2* and *rj4* were recently cloned (Yang et al., 2010; Tang et al., 2016). The *rj2* gene can associate with NopP to determine symbiotic specificity (Okazaki et al., 2013; Sugawara et al., 2018). *Rj4* can regulate soybean compatibility and incompatibility with rhizobia, and

interestingly, the *Rj4* gene was found to associate with some T3Es of *Bradyrhizobium elkanii* to influence establishment of the symbiosis (Faruque et al., 2015). In recent studies in soybean, *PP2C-related gene* and *RPK* were detected by QTL mapping and were shown to interact with *S. fredii* HH103 NopL to regulate the infection of soybean root cells by rhizobia (Zhang et al., 2018). *Via* QTL mapping, *S. fredii* HH103 NopP was found to induce the expression of *TLP* and *MAPK3* during rhizobium infection (Wang et al., 2018). The identification and study of T3Es and their interacting genes could enhance the understanding of the signal communication between host plants and rhizobia during the establishment of symbiosis.

In this study, we show that NopD can be secreted from *S. fredii* HH103 in the presence of genistein. The nodulation effect of NopD was analyzed on 10 soybean germplasms, including Charleston, Dongnong594, Suinong14, and ZYD00006. These varieties showed significant differences in nodulation phenotype after being inoculated with *S. fredii* HH103 (wild type), the NopD and TtsI mutants, respectively. The recombinant inbred line (RIL) population derived from Charleston × Dongnong594 was used to identify QTL loci related to NopD. The conditional QTL related to NopD were verified by the chromosome segment substitution lines (CSSLs). Finally, two genes *Glyma.19g068600* (FBD/LRR) and *Glyma.19g069200* (PP2C) located on the overlap region of QTL location were identified as candidate genes that can interact with NopD. The expression of both genes can be induced by NopD. However, the haplotype effect on nodule traits is different between *19g068600* and *Glyma.19g069200*.

MATERIALS AND METHODS

Strains, Primers, and Plasmids in This Study

Bacterial strains *S. fredii* HH103, the derived mutants HH103Ω*NopD*, HH103Ω*TtsI*, and *Escherichia coli* DH5α were used are listed in **Supplementary Table S2**. Primers for gene cloning and qRT-PCR are listed in **Supplementary Table S1**. Plasmids used for mutant construction and studies of gene function are listed in **Supplementary Table S2**.

Construction of the HH103Ω*NopD* and HH103Ω*TtsI*

The construction of insertion mutants was performed as follows: a 1.4 kb fragment containing a 550 bp fragment upstream of the *NopD* ATG codon and an 850-bp fragment downstream of the ATG codon was cloned into pGWC, yielding plasmid pGWC-*NopD*1400. A *Spe*I restriction enzyme site was constructed close to the start codon of *NopD* using the Fast Mutagenesis System (Transgene Co., Beijing, China). Primers for site-directed mutagenesis are listed in **Supplementary Table S1**. A kanamycin Ω interposon was obtained from pEASY-Blunt with *Spe*I and then ligated into pGWC-*NopD*1400 *Spe*I site, yielding pGWC-*NopD*2400Ω. The 2,400 bp construct was then cloned into the suicide vector pJQ200SK (Quandt and Hynes, 1993) using *Xba*I and *Sma*I. The triparental mating was used to the transferred pJQ-*NopD*2400Ω from *E. coli* DH5α cells into *S. fredii* HH103 in the presence of pRK2013 helper plasmid (Figurski et al., 1979). Candidate mutant recombination colonies were obtained by screening for kanamycin resistance and growth on sucrose (5% w/v). Subsequently, positive mutants were screened by antibiotics and 5% sucrose. The candidate NopD and TtsI mutants were detected by PCR, qRT-PCR, and analysis of nodulation outer proteins. All the bacterial strains, primers, and plasmids used for mutant construction are shown in **Supplementary Tables S1, S2**.

RNA Isolation of Rhizobia and qRT-PCR Analyses of NopD

S. fredii strains HH103, HH103Ω*NopD*, and HH103Ω*TtsI* were incubated with shaking at 28°C in YM medium in the presence or absence of 3.7 μM genistein. Rhizobial RNA was extracted as described (Jiménez-Guerrero et al., 2015a), gDNA was treated by gDNA remover (Transgene Co., Beijing, China) to eliminate its effects on expression, and then RNA samples were synthesized into cDNA using TransScript® One-Step cDNA Synthesis SuperMix (Transgene Co., Beijing, China). qRT-PCR was performed with TransStart® Top Green qPCR SuperMix (Transgene Co., Beijing, China) in a Roche LightCycler 480 II System. The qRT-PCR program was as follows: denaturation at 94°C for 30 s, followed by 40 cycles of 94°C for 5 s, 60°C for 15 s, and 72°C for 10 s. The 16S rRNA gene was used as a reference gene to calibrate the transcript abundance values among different cDNA samples (Crespo-Rivas et al., 2007). The threshold cycle values were analyzed by the software in the Roche LightCycler 480 II. All sample harvests were performed with three biological replicates, and the individual values for each RNA sample were analyzed by three technical replications. The primers for expression analysis are listed in **Supplementary Table S1**.

Analysis of NopD in Nodulation Outer Proteins

The wild-type strain and two mutants were each cultured in 500 ml YM medium at 28°C until OD_{600} reached 0.6. The bacteria were cultured in the presence of 3.7 mM genistein for about 40 h at 28°C. After centrifugation of the cells at 8,000 g for 30 min (4°C), the supernatant was collected for purification of outer proteins. To eliminate contamination by bacteria and increase the protein concentration, the supernatant was filtered through Millipore™ filter units (0.22 μM) (Millipore Co., Germany), then concentrated using Millipore™ Amicon™ Ultra-15 (100 kDa) centrifugal filter units (Millipore Co., Germany). Proteins were precipitated in the presence of 10% w/v trichloroacetic acid for 20 h at 4°C then collected by centrifugation at 10,000 g for 20 min (4°C). After washing twice with cold 80% acetone (v/v), the precipitated proteins were resuspended in 8 M urea. Extracellular proteins from the different strains were separated by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). For immunostaining, extracellular proteins were transferred onto nitrocellulose membranes, then the membranes were blocked with TBST pH 7.5 (per liter contains 3.03 g Tris, 8 g NaCl, and 1 ml Tween 20) containing 5% (w/v) skim milk, followed

by incubation for 1 h with anti-NopD rabbit antiserum (1,000-fold dilution). Subsequently, the membranes were incubated with goat anti-rabbit immunoglobulin AP-conjugated secondary antibody (Abmart Co., China) for 1 h in accord with the supplier's instructions, and reaction results were visualized using SuperSignal West Pico Chemiluminescent Substrate (Thermo Co., United States).

Agrobacterium tumefaciens-Mediated Transformation

Agrobacterium-mediated transformation by agroinfiltration was performed as follows: the *NopD* gene (GenBank: CEO91485.1) was cloned into the entry vector pGWC, and the entry clone was subsequently recombined into the destination vector pGWB17 using the Gateway® system (Invitrogen Co., United States). Plasmids pGWB17-*NopD*, the empty vector pGWB17-T, and pGWB17-*HopQ1* were transformed into *A. tumefaciens* EHA105 by electroporation. Four-week-old *Nicotiana benthamiana* plants were used for transient expression: the *A. tumefaciens* culture was adjusted to OD_{600} 0.5 using infiltration buffer (10 mM $MgCl_2$, 10 mM MES-KOH pH 5.6, 150 μM acetosyringone). Top leaves were used for infiltrating, then at 0–4 days after infiltration, leaves were harvested for detection of cell death. Staining of *N. benthamiana* leaves was performed with trypan blue as described by Tennant (1964). Electrolyte leakage was used to evaluate cell death in leaf tissues by measuring ion conductivity (Mackey et al., 2002).

Nodulation Tests

For nodulation tests, the wild-type strain *S. fredii* HH103 and mutants HH103Ω*NopD* and HH103Ω*TtsI* were used. The soybean germplasms of different ecoregions and RIL population used are listed in **Supplementary Table S3**. Soybean seeds were sterilized with chlorine gas for 12–14 h, and then germinated into sterilized 300 ml plastic jar units containing nitrogen-free nutrient solution in the lower vessel. Each jar was kept in a greenhouse with 16 h/8 h light/dark at 26°C until plants grew to Vc stage, then plants were inoculated with 2 ml of 10 mM $MgSO_4$ containing approximately 1×10^9 bacteria (Zhang et al., 2018). Thirty days after inoculation, all soybean roots were harvested for nodulation evaluation. The nodulation tests on each soybean germplasm were performed with five replicates and three independent experiments. The statistical significance of differences in phenotype was detected using *t*-tests.

The Conditional Quantitative Trait Locus Mapping of Nodulation-Related Traits

The experimental RIL population used in this study (n = 150) was derived from the cross of Charleston and Dongnong594 (Jiang et al., 2018). A high-density genetic map with 5,308 specific locus amplified fragment sequencing (SLAF-seq) markers had been constructed *via* this population (Qi et al., 2014). Recently, we also completed the genomic resequencing of CSSLs and their parents (unpublished). These genetic backgrounds supply a useful support to identify the candidate genes in interested QTL.

To detect the QTL underlying nodule-related traits, a composite interval mapping method was used with WinQTL Cartographer (Wang et al., 2018; Zhang et al., 2018). The detailed parameters followed the published method (Zhang et al., 2018). The nodule number (NN) and nodule dry weight (NDW) of RILs and their parents after inoculation with NopD mutant or the parental strain were used for QTL identification. The differences in phenotypic values were used to determine the location of conditional QTL. At the same time, the phenotype differences in RILs inoculated with HH103Ω*TtsI* or the parental strains were used to determine the conditional QTL's location (Wang et al., 2018). The detailed method is as follows. When the operation was running, the control marker number was set to 5 and window size was 10 cM. The forward regression method and a walk speed of 0.5 cM were selected. The proportional and additive effects of variances interpreted by each specific QTL were obtained by composite interval mapping analysis. The log of the odds (LOD) score peaks higher than 3.0 (WinQTL Cartographer default threshold) was selected to indicate the existence of conditional QTL for the nodule traits inoculated with the two type strains, respectively. About the additive-effects signals, "+" indicates increasing allelic effects from "Dongnong 594" and "−" indicates decreasing allelic effects from "Charleston." The 1,000 permutations of each genotypic marker against the phenotype in RIL population determined the experimental threshold levels for linkage. When the two values for a marker were greater than the critical value at p = 0.05, it indicated that the linkage was significant. CSSLs produced by the cross between soybean cultivar SN14 and wild soybean ZYD00006 (*G. soja* Sieb. & Zucc.) were used for verifying consensus QTL, according to the genetic map by Xin et al. (2016).

Annotation of Candidate Genes in Quantitative Trait Locus of NopD-Related

The "Williams 82. a2. v1" genome was the first published soybean genome and could provide valuable information for QTL mapping of soybean important traits (Brensha et al., 2012). Phytozome website[1] and Soybase database[2] integrated the soybean Williams 82 genome information and the latest information on the soybean genome uploaded by researchers, could provide necessary information for mapping. Our laboratory completed the construction of a high-density genetic map of the RIL population in the early stage (Qi et al., 2014). So the soybean genes in the major QTL intervals could be identified by combining the high-density genetic map information with two important database tools, and corresponding gene annotations were performed (Wang et al., 2018). Candidate genes involved in the plant immunity and signal conduction were selected from the annotation data, which were used for subsequent verification.

[1]https://phytozome.jgi.doe.gov/pz/portal.html#
[2]https://www.soybase.org/

Verification of NopD Candidate Genes by qRT-PCR

qRT-PCR in various soybean germplasm materials was performed to verify candidate genes that potentially interacted with NopD. Root samples were harvested at 0, 12, 24, 36, 48, and 60 h post-inoculation with the wild-type strain HH103, NopD mutant, and TtsI mutant, respectively. The classic TRIzol reagent (Invitrogen Co., United States) was used for the extraction of total RNA from soybean roots, and total RNA was treated by gDNA remover (Transgene Co., Beijing, China) to eliminate its effects on expression, and then RNA samples were synthesized into cDNA by the TransScript® One-Step cDNA Synthesis SuperMix (Transgene Co., Beijing, China). qRT-PCR was performed as described above, and *GmELF1b* was used as the internal control to normalize the transcript amounts in different samples (Jian et al., 2008). The gene-specific primers for qRT-PCR are listed in **Supplementary Table S1**.

Haplotype Analysis of Candidate Genes Based on Chromosome Segment Substitution Line Population

Haplotype analysis of *Glyma.19g068800* and *Glyma.19g069200* was based on the resequencing data of CSSL populations in a subset of 142 soybean accessions. The genomic regions including the gene transcript sequence and approximately 2.0-kb promoter region of two genes upstream of the start codon were used for the haplotype analysis.

RESULTS

NopD Can Be Secreted From *S. fredii* HH103 by Type 3 Secretion System

The *NopD* coding sequence was located in the plasmid pSfHH103d of *S. fredii* HH103. The promoter region contains a *tts* box (−257 bp), supporting the idea that *NopD* can be induced by TtsI (**Figure 1A**). By phylogenetic analysis, we can separate the NopD tree into two branches (I-Red and II-Green). Branch I comprises only three *Sinorhizobium* species. Branch II comprises *Mesorhizobium* and *Bradyrhizobium* species without *Sinorhizobium* species (**Figure 1A**).

NopD's expression was studied by qRT-PCR in strains HH103, HH103Ω*NopD*, and HH103Ω*TtsI* induced or not by genistein. The qRT-PCR results showed that genistein promoted the expression of *NopD* significantly in the wild strain (**Figure 1B**), but *NopD* expression was not detected in the NopD and TtsI mutant either with or without genistein. This result was consistent with the previous reports that *NopD*'s expression was downregulated in a flavonoid mutant, NodD1 mutant, and TtsI mutant (Pérez-Montaño et al., 2016). An antibody against NopD was used to detect the NopD protein in supernatants from HH103, HH103Ω*NopD*, and HH103Ω*TtsI*. Western blot results showed a band corresponding to NopD (about 160 kDa) in samples from *S. fredii* HH103 induced with genistein, but not in NopD or TtsI mutants (**Figure 1B**).

Nodulation Tests

To elucidate the role of NopD in establishing symbiosis, we collected 10 soybean varieties with differing genetic backgrounds from various ecoregions in China and United States. In most soybean germplasms, there were significant differences in NN and NDW after inoculation with the NopD mutant or the wild-type HH103 (**Figure 1C**). Only in Kenjian28, NopD had a negative effect on the NN and NDW, and there was no difference in NN of Charleston when inoculated with the wild strain or HH103Ω*NopD*. However, NopD had a positive effect on NN and NDW in eight of the 10 soybean germplasms, except for the NDW of ZYD00006 (**Figure 1C**). Thus, NopD exerting either positive or negative effects on the various soybean germplasms might be due to the different genetic backgrounds of germplasms. On the other hand, these results support NopD mainly played a positive effect on soybean nodulation.

Nodulation tests show that Dongnnong594 and Charleston, as the parents of RIL population, had a significant difference in NN and NDW when inoculated with wild strain and NopD mutant, the same situation also occurs in the CSSL populations with Suinong14 and ZYD00006 as their parents, so RIL and CSSL populations could be used to map important QTL interactions with NopD.

NopD Induces Leaf Death in *Nicotiana benthamiana*

Agrobacterium-mediated transient transformation was used to study NopD's effects within tobacco leaves. pGWB17, a binary vector containing the cauliflower mosaic virus 35S promoter and the NopD coding sequence, was introduced into *A. tumefaciens* EHA105, which was then infiltrated into leaves of 4-week-old tobacco plants. Four days after infiltration, a clearly necrotic leaf zone (hypersensitive response with induced cell death) was observed in tissue transformed with NopD as well as in leaves expressing HopQ1 of *Pseudomonas syringae*, which was used as a positive control (**Figure 2**). These results indicated that NopD functions like an avirulence effector in tobacco. Trypan blue staining and electrolyte leakage were used to detect the effect of NopD expression in leaves (**Figure 2**).

Phenotype of Nodulation Analysis in Recombinant Inbred Lines

Nodulation tests on the soybean germplasms showed that there were significant differences in NN and NDW between Charleston and Dongnong594 when inoculated with the wild-type strain HH103 or NopD and TtsI mutants (**Figure 1C** and **Table 1**). The RIL population was derived from the cross between Charleston and Dongnong594. From the nodulation tests, the NopD mutant gave a higher average NN and NDW in the whole RIL population compared with inoculation with the wild-type strain or TtsI mutant. There were no significant differences in average NN or NDW of the whole RIL population after inoculation with the wild-type strain or the TtsI mutant. The RIL population has a more complex genetic background than the individual parents Charleston and Dongnong594, the genetic information between individuals were quite different,

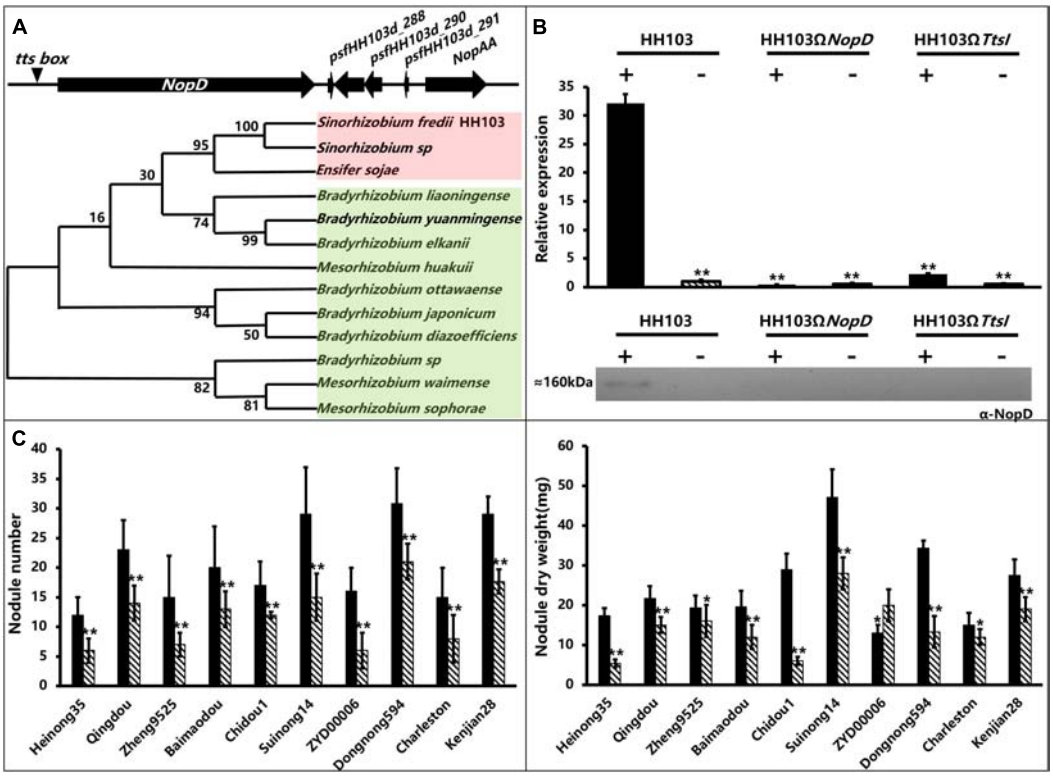


FIGURE 1 | The information of NopD gene, detection of NopD protein induction by genistein and nodule traits of soybean plants inoculated with *S. fredii* wild-type HH103 or mutants HH103Ω*NopD* and HH103Ω*TtsI*. **(A)** Position of the NopD gene and phylogenetic tree of NopD from various rhizobial strains by neighbor-joining method using MEGA6.0 software. **(B)** Expression analysis of NopD in *S. fredii* wild-type strain HH103 and mutants HH103Ω*NopD* and HH103Ω*TtsI* with (+) or without (–) genistein (3.7 μM). Final expression was calculated relative to the expression of the wild-type strain HH103 in the absence of genistein. All sample harvests were performed with three biological replicates, and the individual values for each RNA sample were analyzed by three technical replications. The sample of the wild-type strain *S. fredii* HH103 in the absence of genistein was used as the control. Asterisks indicate significant differences at the level $\alpha = 1\%$ ($P < 0.01$). Immunoblot analysis of NopD in extracellular protein extracts of the wild-type strain *S. fredii* HH103 and indicated mutant derivatives induced with or without genistein (3.7 μM). Immunoblots were performed with anti-NopD antibodies. **(C)** The analysis of phenotype was performed three times; significant differences were determined by *t*-tests; * indicates $0.01 \leq P \leq 0.05$ and **indicates $P \leq 0.01$. Soybean varieties (with origins): Heinong35 (Heilongjiang), Qingdou (Shanxi), Zheng9525 (Henan), Baimaodou (Zhejiang), Chidou1 (Inner Mongolia), Suinong14 (Heilongjiang), ZYD00006 (Heilongjiang), Charleston (America), Dongnong594 (Heilongjiang), Kenjian28 (Heilongjiang).

these individuals should have different responses to the wild strain, NopD mutant, or TtsI mutant, and this might have caused the observed differences compared with 10 soybean varieties in phenotype. So the nodule traits of the whole RIL population are not representative after inoculation with the wild strain, NopD mutant, and TtsI mutant. The more complex genetic background of the RIL population should facilitate the mining of QTL associated with the phenotype.

Conditional Quantitative Trait Locus Mapping for Nodule Number and Nodule Dry Weight and Validation of Consensus Quantitative Trait Locus

Seven conditional QTL for NN and four conditional QTL for NDW were identified, after inoculation with *S. fredii* HH103 or the NopD and TtsI mutants (**Table 2** and **Figure 3**). Seven conditional QTL associated with NN were located on chromosomes Gm02 ($n = 2$), Gm17 ($n = 1$), Gm18 ($n = 1$), and Gm19 ($n = 3$), and four conditional QTL for NDW were located on Gm08 ($n = 1$), Gm9 ($n = 1$), Gm11 ($n = 1$), and Gm19 ($n = 1$).

Overlapping conditional QTL (118.4 cM) for NN and NDW were identified on Chromosome 19, and we found two further conditional QTL (111.0 and 133.2 cM) close to these overlapping QTL. An earlier investigation of soybean revealed two QTL (99.7 and 108 cM) on Gm19 associated with nodule weight, and two QTL (97.5 and 108.2 cM) associated with nodule size (Hwang et al., 2014). This previously identified QTL support that in the region of QTL we identified is confident. So we selected this region as a candidate region that might contain gene response to NopD. To test our hypothesis, specific lines of CSSLs were identified and used for further work.

Nodulation tests showed significant differences in NN and NDW between Suinong14 and ZYD00006 after inoculation with wild-type strain HH103, NopD mutant, and TtsI mutant (**Figure 1C**). A CSSL population derived from the cross between Suinong14 and ZYD00006 was used to verify whether the overlapping QTL might be related to NN or NDW in

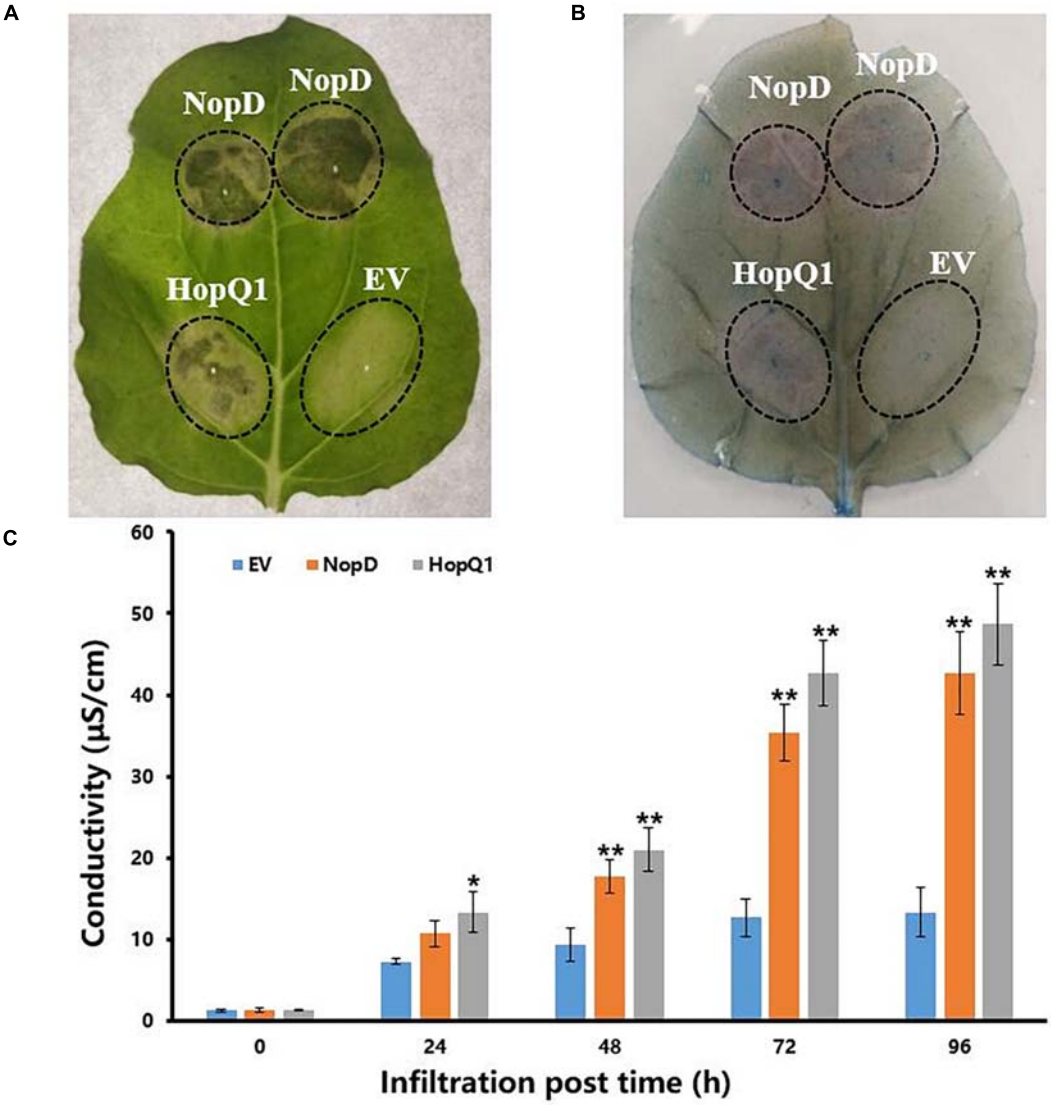


FIGURE 2 | Effects of NopD on *N. benthamina* cells. Infiltrated zones are marked by circles (EV, infiltration zone with *A. tumefaciens* carrying the empty vector pGWB17-T). For comparison, *A. tumefaciens* carrying plasmid pGWB17-HopQ1-myc was also used for infiltration (expression of HopQ1 of *P. syringae* pv. tomato DC3000). **(A)** Leaves were photographed 4 days post infiltration. **(B)** Trypan blue-based cell death staining of a representative *N. benthamiana* leaf transiently expressing NopD and HopQ1 4 days post *Agrobacterium infiltration*. **(C)** Electrolyte leakage levels in leaves. "*" indicated the significant differences ($p \leq 0.05$) and "**" indicated the significant differences ($p \leq 0.01$).

TABLE 1 | Nodule traits of the recombinant inbred line (RIL) population inoculated with the wild-type strain and NopD and TtsI mutants.

		RILs (n = 150)		Parents (average)	
	Traits	**Average**	**Standard deviation**	**Charleston**	**Dongnong594**
HH103 Rif^R	Nodule number	8.1	10.9	20.6 ± 4.7**	30.8 ± 5.8
	Nodule dry weight (mg)	7.5	14.9	18.0 ± 3.6**	24.3 ± 3.3
HH103 Rif^RΩ*NopD*	Nodule number	11.2*	7.0	13.4 ± 3.5**	17.8 ± 4.1
	Nodule dry weight (mg)	9.9*	18.4	12.0 ± 2.9**	21.3 ± 4.6
HH103 Rif^RΩ*TtsI*	Nodule number	7.0	4.4	12.2 ± 3.8**	18.4 ± 4.7
	Nodule dry weight (mg)	7.3	7.2	13.0 ± 3.6**	19.3 ± 5.8

**Indicates significant differences with different inoculations, $p \leq 0.05$, **indicates $p \leq 0.01$.*

TABLE 2 | Distribution of conditional quantitative trait locus (QTL) for nodule number (NN) and nodule dry weight (NDW) among linkage groups and chromosomes.

Strain	Trait	LG/QTL	Chrom.	Start position (cM)	End position (cM)	LOD[a]	R^2 (%)[b]	ADD[c]
HH103Ω*NopD*	NDW	QK/NDW01	9	68.81	71.81	3.00	5.78	0.49×10^{-3}
		QB1/NDW02	11	50.31	53.31	3.70	9.36	0.64×10^{-3}
	NN	QG/NN01	18	42.91	45.91	3.30	2.08	−1.21
		QL/NN02	19	131.73	134.73	3.80	0.62	−1.32
		QL/NN03	19	116.92	119.92	3.20	1.83	−1.10
HH103Ω*TtsI*	NDW	QA2/NDW03	8	61.41	64.41	3.50	1.31	0.95×10^{-3}
		QL/NDW04	19	116.92	119.92	3.90	5.38	1.78×10^{-3}
	NN	QD1b/NN04	2	28.11	31.11	3.90	6.47	−1.71
		QD1b/NN05	2	124.32	127.32	4.80	3.70	1.73
		QD2/NN06	17	31.81	34.8	3.60	7.12	−1.59
		QL/NN07	19	109.52	112.52	4.00	2.53	0.95

[a]LOD, log of odds. [b]R^2 (%), the contribution rate of the QTL. [c]ADD, the additive effects contributed by the QTL. LG, linkage group; QD1b, chromosome 2; QA2, chromosome 8; QK, chromosome 9; QB1, chromosome 11; QD2, chromosome 17; QG, chromosome 18; QL, chromosome 19.

other soybean populations (**Supplementary Table S4**). We selected six soybean lines from the CSSL, three of which (CSSL-600, CSSL-603, and CSSL-612) had higher NN and NDW than the parent Suinong14 after inoculation with NopD mutant and TtsI mutant, and another three of which (CSSL-519, CSSL-593, and CSSL-648) had lower NN and NDW than ZYD00006 after inoculation with NopD mutant and TtsI mutant. SSR markers located the conditional QTL on chromosome 19 when the RIL population were screened. Substituted chromosome fragments were also detected in related regions: we found an overlap region between the conditional QTL and the fragment of a CSSL (Sat_134–Satt398) on chromosome 19 (97.5–133.2 cM), these lines including CSSL-600, CSSL-603, and CSSL-612. No corresponding region was found in lines CSSL-519, CSSL-593, CSSL-648, or the parent ZYD00006 (**Figure 4**).

Furthermore, our candidate QTL were close to those QTL in the previous reports and identified by the CSSL population. No QTL or genes interacting with NopD had been identified in previous reports, so it was interesting to determine whether the QTL identified in our work might interact with NopD.

Expression Analysis of Candidate Genes Associated With Nodulation Phenotype

In the confident QTL region, seven genes (*Glyma.19g065800, Glyma.19g066800, Glyma.19g067200, Glyma.19g068300, Glyma.19g068600, Glyma.19g068800*, and *Glyma.19g069200*) on Chromosome 19 associated with pathogen resistance, signal exchange, and symbiosis were selected for further analyses (**Supplementary Table S5**). To identify whether these genes have interaction with NopD, soybean Suinong14 was inoculated with the wild-type HH103, HH103Ω*NopD*, and HH103Ω*TtsI*. The expression pattern of these genes was detected by qRT-PCR (**Figure 5**). Among the seven genes, *Glyma.19g067200* showed no expression in Suinong14 inoculated with any of the three strains because of no expression signal detected by qRT-PCR. At 12 h post-inoculation, the *Glyma.19g065800* expression levels were significantly different comparing TtsI mutant with the wild strain and NopD mutant, at 24 and 60 h post-inoculation, the wild strain could induce the higher expression. It showed that NopD and TtsI mutant could induce a different expression pattern, indicating that *Glyma.19g065800* had no obvious interaction with NopD. At 24 h post-inoculation with the wild-type strain HH103 and NopD mutant, the expression level of *Glyma.19g066800* reached a maximum, but TtsI mutant did not have a similar trend. *Glyma.19g068300* had the similar expression trend in soybean inoculated with three strains, and expression levels all changed around 1.0. At 12 h post-inoculation, the *Glyma.19g068300* expression levels was significantly different inoculated with TtsI mutant compared to the wild strain and NopD mutant. These results showed that *Glyma.19g068300* and *Glyma.19g066800* had no obvious interaction with NopD similar with *Glyma.19g065800*. The expression of *Glyma.19g068800* had a strange pattern; NopD mutant could induce the gene to reach two peaks at 24 and 60 h post-inoculation, its expression level was higher compared with inoculation with the wild strain and TtsI mutant which had a similar induction pattern, so we could not infer that *Glyma.19g068800* could interact with NopD. *Glyma.19g068600* and *Glyma.19g069200* showed significantly different expression patterns in Suinong14 after inoculation with *S. fredii* HH103 and the two mutant strains. When Suinong14 was inoculated with *S. fredii* HH103, the expression patterns of *Glyma.19g068600* and *Glyma.19g069200* were similar. At 36 h post-inoculation with the wild-type strain HH103, the expression level of *Glyma.19g068600* and *Glyma.19g069200* reached a minimum, and then the expression level of these two genes increased. But in a clear difference from the wild-type strain, infection with the mutants HH103Ω*NopD* and HH103Ω*TtsI* showed no evident effects on the expression patterns of these two genes. At 36 h post-inoculation with HH103Ω*NopD* and HH103Ω*TtsI*, the expression level of *Glyma.19g068600* was 6.7 and 5.2 times higher than expression induced by inoculation with HH103. Under the same conditions, the expression level of *Glyma.19g069200* induced by HH103Ω*NopD* and HH103Ω*TtsI* was 10.7 and 12.0 times higher than that in plants induced by HH103, respectively. These qRT-PCR results supported the hypothesis that NopD interacted with

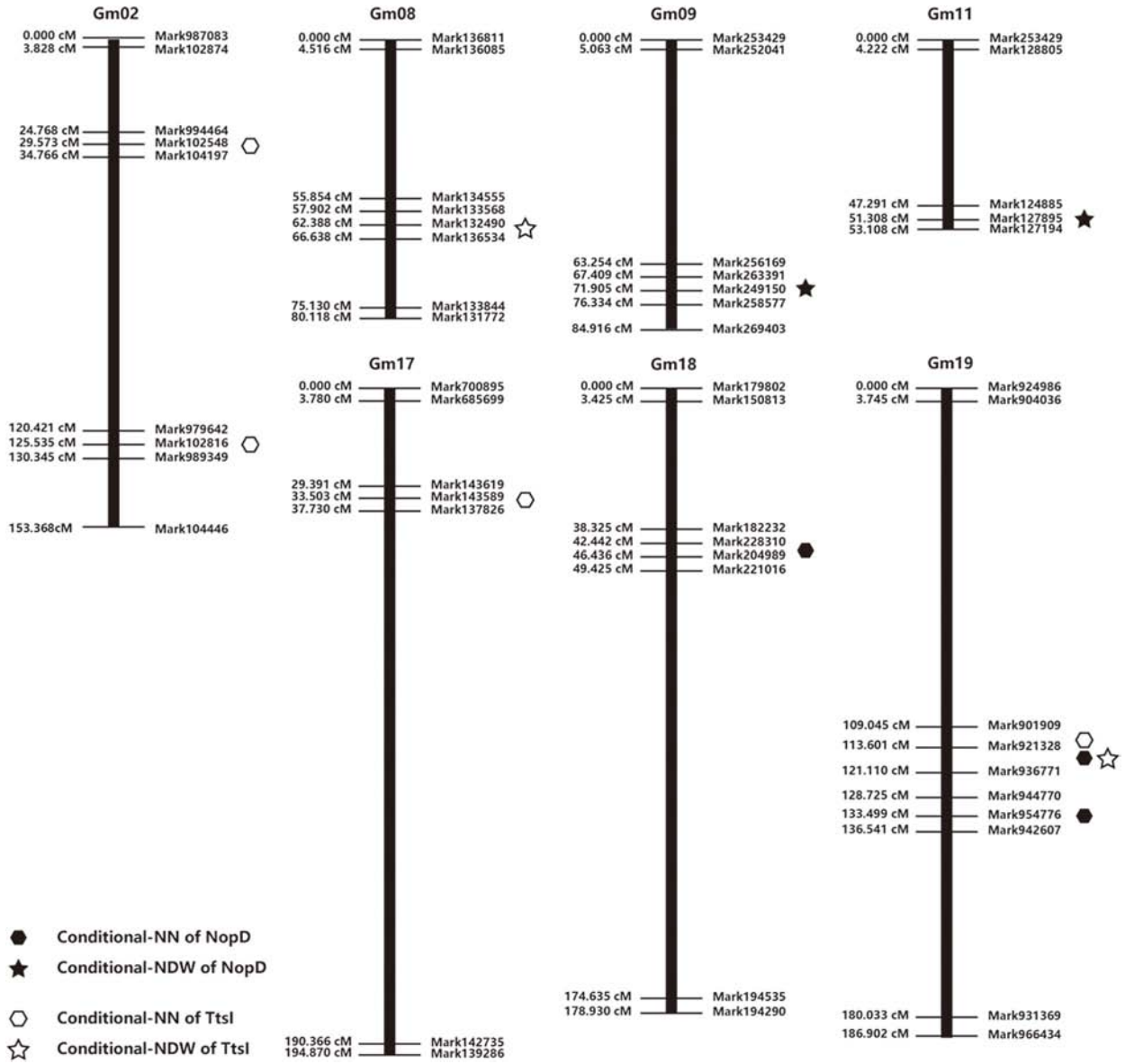


FIGURE 3 | Distribution of conditional quantitative trait locus (QTL) for nodule number and nodule dry weight among linkage groups and chromosomes.

Glyma.19g068600 and *Glyma.19g069200* in the early stage of rhizobial infection.

Haplotypes of *Glyma.19g068600* and *Glyma.19g069200* That Correlated With Nodulation Traits

Considering that *Gm19g068600* and *Gm19g069200* can be suppressed at a lower level in Suinong14 during the wild HH103 infecting (**Figure 5**) and nodule traits had the significant differences in parents of CSSL populations (**Figure 1C**). We analyzed the haplotypes of *Gm19g068600* and *Gm19g069200* in 142 accessions of CSSL populations. In total, 14 single-nucleotide polymorphisms (SNPs) and/or indels were found in the promoter and coding sequence of *Gm19g068600* from the CSSL populations. According to 14 SNPs and/or indels, 142 soybean accessions were classified into eight haplotypes (Hap1–Hap8) (**Figure 6A**). Hap1, the largest group, includes 40 soybean accessions; Hap2 including 32 soybean accessions was the second largest group. In these two types, two SNPs are located in the exon, but these differences did not result in the amino acid change. Ten SNPs and indels located in the promoter sequence. The relative expression of gene was detected during some rhizobia strains infecting in Hap1 and Hap2, the relative expression showed that *Gm19g068600* had the similar expression pattern in Hap1 accession and Hap2 accession after inoculated with the wild strain HH103, NopD mutant and TtsI mutant, respectively. The nodule traits of Hap1 and Hap2 accessions

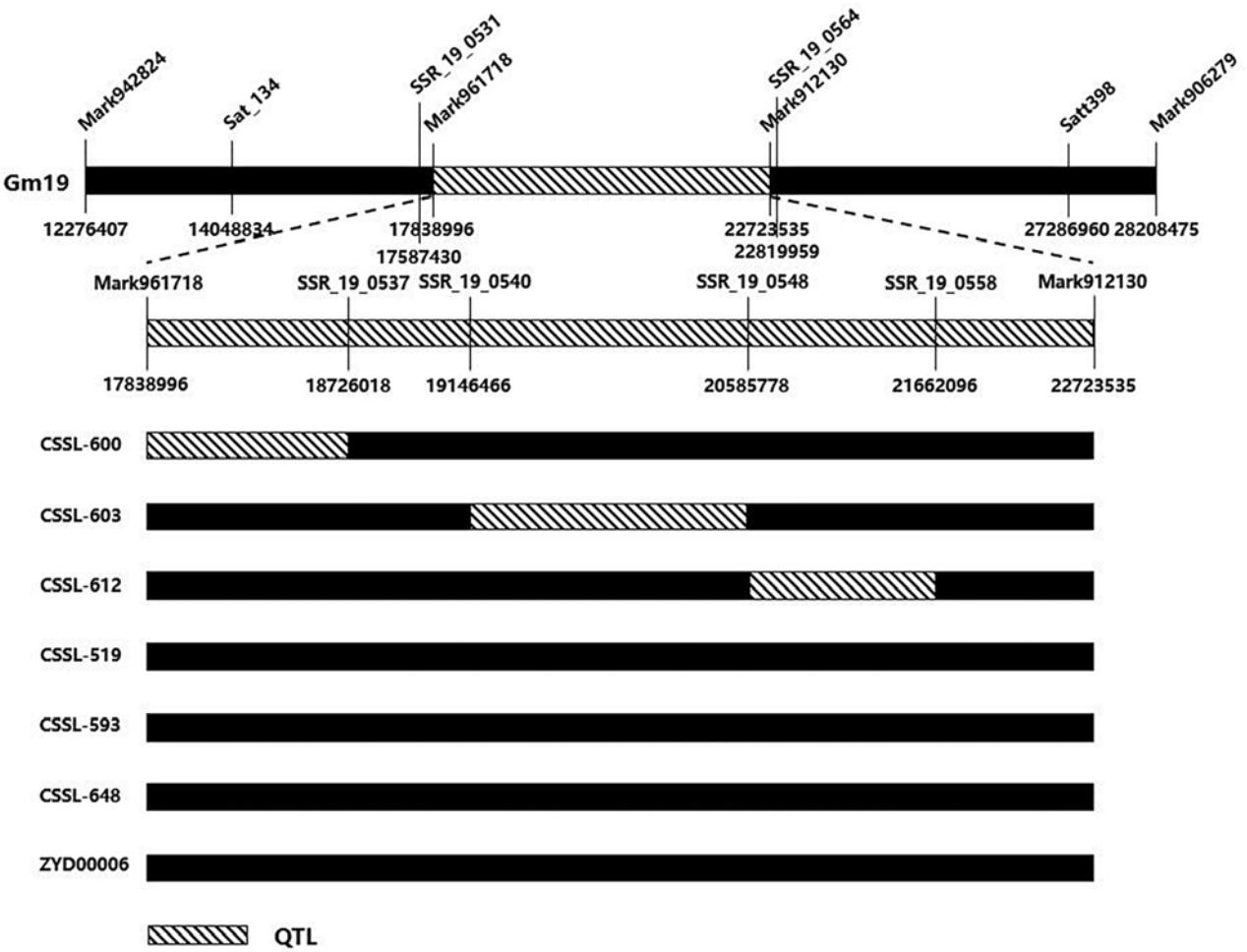


FIGURE 4 | Validation of consensus quantitative trait locus (QTL) in chromosome segment substitution lines (CSSLs). The consensus QTL associated with nodulation phenotype on chromosome Gm19 (97.5–133.2 cM) had a corresponding partial region in the substituted wild soybean chromosomal segment on Gm19 in CSSL-600, CSSL-603, and CSSL-612. On the contrary, there was no corresponding region in CSSL-519, CSSL-593, CSSL-648, or ZYD00006.

were further compared, the nodulation results show that the nodule traits of haplotypes do not have significant differences after inoculation with the wild HH103, NopD mutant, and TtsI mutant, respectively (**Figure 6B**). The relative expression and nodule trait analysis indicated that *Gm19g068600* could be related to NopD, but haplotype difference could not influence expression pattern.

The haplotype analysis of *Gm19g069200* shows that there were 15 SNPs and/or indels found in the promoter and coding sequence. Based on the 15 SNPs and/or indels, 142 soybean accessions were classified into seven haplotypes (Hap1–Hap7). Hap1, the largest group, includes 42 soybean accessions, and Hap2 was the second largest group, including 39 soybean accessions (**Figure 7A**). One SNP located in the exon, but the change did not result in the amino acid change, five SNPs or indels located in the promoter sequence. As it is known, the promoters always affects the function of genes by regulating gene expression, and the relative expression of *Gm19g069200* was studied in one Hap1 accession and one Hap2 accession, respectively. These results show that the relative expression of *Gm19g069200* had a significant difference in Hap1 and Hap2 soybean accession (**Figure 8**). *Gm19g069200* could be regulated by NopD in HH103 *via* Hap1 accession, but not in Hap2 accession. The nodule traits of Hap1 and Hap2 accessions were further compared, and the nodulation results show that haplotypes have significantly different effects on nodule trait after inoculation with the wild strain, NopD mutant, and TtsI mutant. The NN and NDW of Hap1 soybean accessions were higher than the Hap2 soybean accessions, and the difference was significant ($P < 0.01$) (**Figure 7B**). Similar to the wild strain, NopD mutant and TtsI mutant could result in higher NN and NDW in Hap1 accessions than Hap2 accessions. NopD mutant and TtsI mutant could reduce NN and NDW compared with the wild strain in both Hap1 accessions and Hap2 accessions. These results suggested that *Gm19g069200* can be associated with the rhizobia infection. Comparisons of nodule phenotypes of Hap1 accessions and Hap2 accessions inoculated with different strains show that the presence of NopD in the wild strain could cause

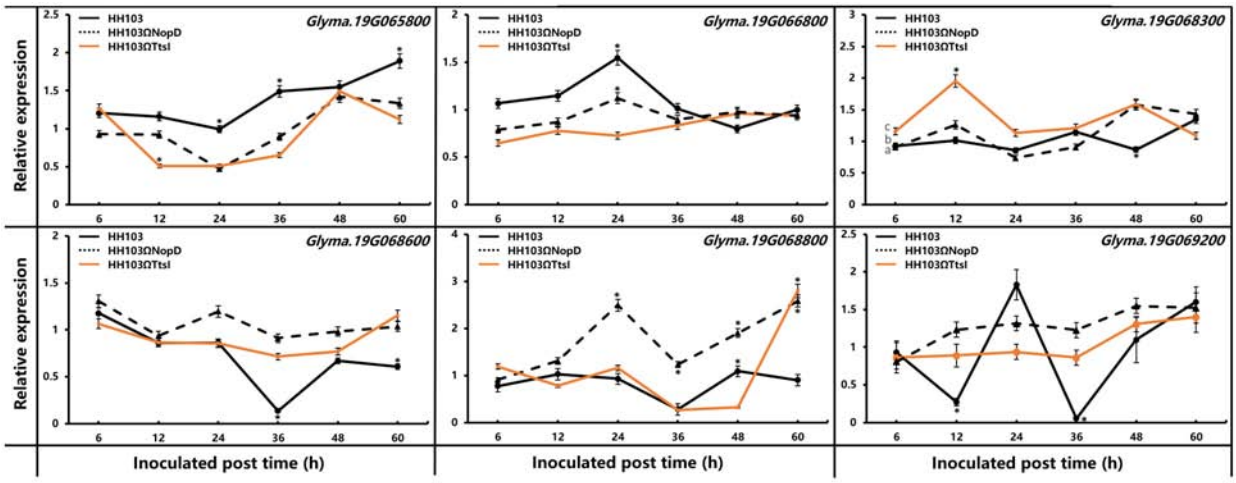
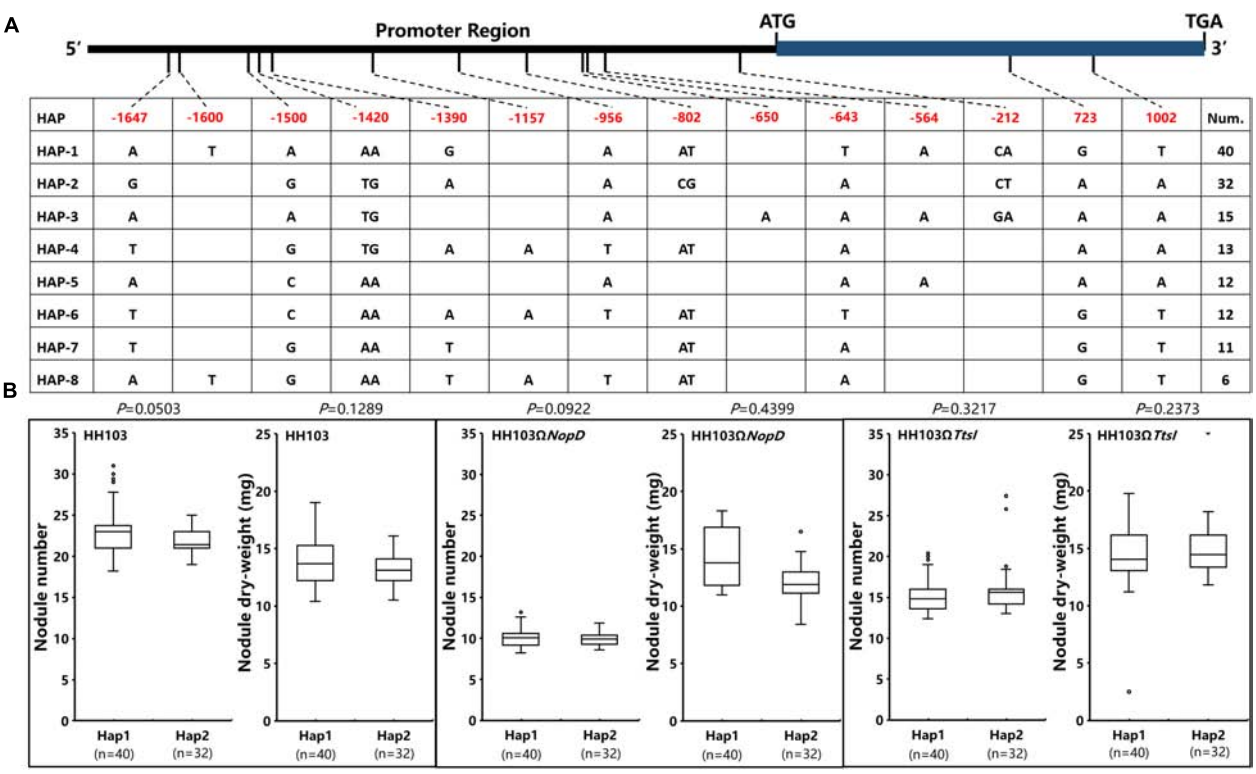
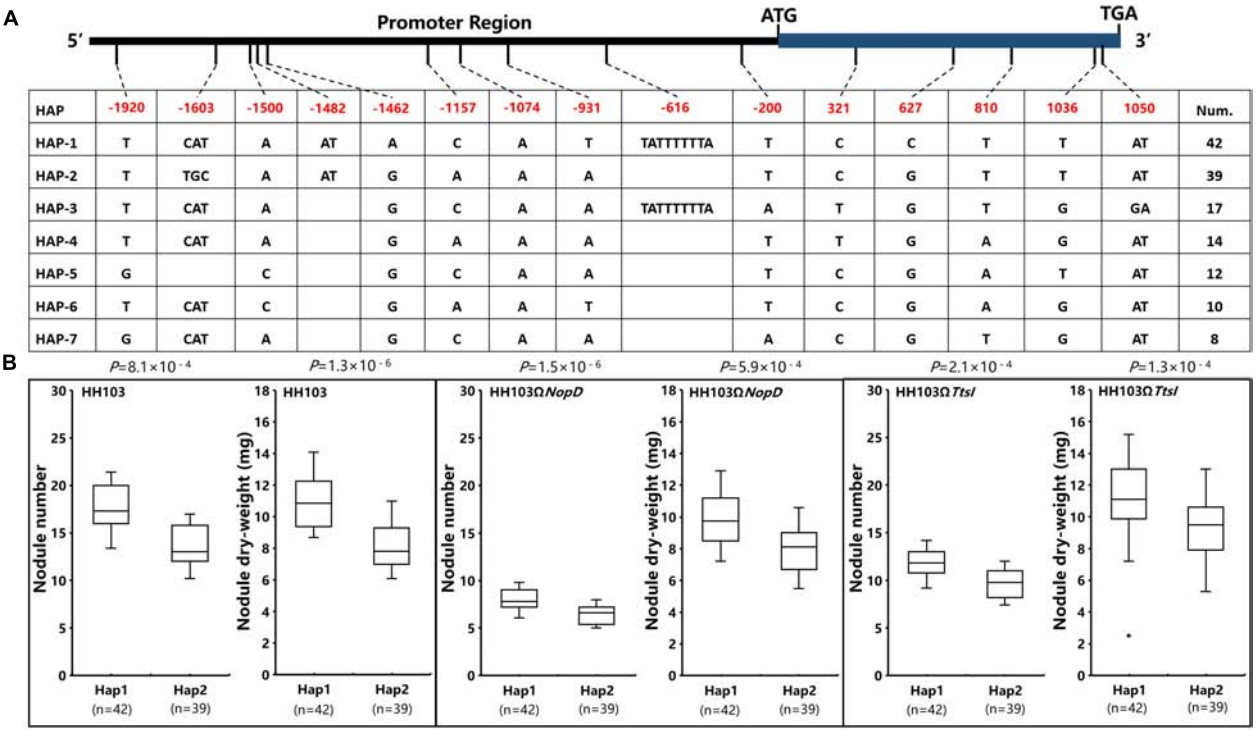


FIGURE 5 | Relative expression of *Glyma.19g065800, Glyma.19g066800, Glyma.19g068300, Glyma.19g068600, Glyma.19g068800*, and *Glyma.19g069200* was measured by qRT-PCR in RNA extracted from roots of soybean Suinong14 plants inoculated with *S. fredii* HH103 wild-type or mutants HH103Ω*NopD* and HH103Ω*TtsI*. Uninoculated Suinong14 plants were used as the control. The $2^{-\Delta\Delta Ct}$ method was used to calculate the relative expression levels of candidate genes. Error bars indicate the mean ± standard error of three replications. Significant differences were determined by *t*-tests, and "*" indicated the significant differences ($p \leq 0.05$) at the time point.



HAP	-1647	-1600	-1500	-1420	-1390	-1157	-956	-802	-650	-643	-564	-212	723	1002	Num.
HAP-1	A	T	A	AA	G		A	AT		T	A	CA	G	T	40
HAP-2	G		G	TG	A		A	CG		A		CT	A	A	32
HAP-3	A		A	TG			A		A	A	A	GA	A	A	15
HAP-4	T		G	TG	A	A	T	AT		A			A	A	13
HAP-5	A		C	AA			A			A	A		A	A	12
HAP-6	T		C	AA	A	A	T	AT		T			G	T	12
HAP-7	T		G	AA	T			AT		A			G	T	11
HAP-8	A	T	G	AA	T	A	T	AT		A			G	T	6

FIGURE 6 | The analysis of *Gm19g068600* haplotypes and nodule traits in two Hap accessions. **(A)** *Gm19g068600* gene haplotypes (eight haplotypes) in chromosome segment substitution line (CSSL) populations (n = 142). The start codon site is set at "0" position. **(B)** Comparison and analysis of the nodule phenotype between haplotypes Hap1 (n = 40) and Hap2 (n = 32). n represents the number of the two haplotypes. Two-tailed *t*-test was used to detect the significance.



HAP	-1920	-1603	-1500	-1482	-1462	-1157	-1074	-931	-616	-200	321	627	810	1036	1050	Num.
HAP-1	T	CAT	A	AT	A	C	A	T	TATTTTTTA	T	C	C	T	T	AT	42
HAP-2	T	TGC	A	AT	G	A	A	A		T	C	G	T	T	AT	39
HAP-3	T	CAT	A		G	C	A	A	TATTTTTTA	A	T	G	T	G	GA	17
HAP-4	T	CAT	A		G	A	A	A		T	T	G	A	G	AT	14
HAP-5	G		C		G	C	A	A		T	C	G	A	T	AT	12
HAP-6	T	CAT	C		G	A	A	T		T	C	G	A	G	AT	10
HAP-7	G	CAT	A		G	C	A	A		A	C	G	T	G	AT	8

FIGURE 7 | The analysis of *Gm19g069200* haplotypes and nodule traits in two Hap accessions. **(A)** *Gm19g069200* gene haplotypes (seven haplotypes) in chromosome segment substitution line (CSSL) populations (n = 142). The start codon site is set at "0" position. **(B)** Comparison and analysis of the nodule phenotype between haplotypes Hap1 (n = 42) and Hap2 (n = 39). n represents the number of the two haplotypes. Two-tailed t-test was used to detect the significance.

a greater difference than the absence of NopD, indicating that NopD could be associated with *Gm19g069200*.

DISCUSSION

The locus we identified in this study on soybean chromosome Gm08 overlaps with a previously identified QTL related to compatibility of soybean with *Bradyrhizobium* strains (Ramongolalaina et al., 2018). A QTL previously shown to be associated with nodule weight (Hwang et al., 2014) was found to be adjacent to the QTL on Gm18 in our study. In addition, a locus on Gm17 related to NopL, a T3SS effector of *S. fredii* HH103 (Zhang et al., 2018), overlaps with a locus adjacent to a QTL on Gm17 that we detected.

We identified two genes (*Glyma.19g068600* and *Glyma.19g069200*) by qRT-PCR that had a similar expression pattern induced by the wild-type strain HH103 (showing a minimum of expression at 36 h after inoculation), but this pattern was not identified when soybean was inoculated with the mutant HH103Ω*NopD* or HH103Ω*TtsI*. This expression pattern suggests that *Glyma.19g068600* and *Glyma.19g069200* might be associated with the NopD-triggered signaling pathway activated after inoculation with rhizobia. *Glyma.19g068600* encodes a protein belonging to the F-Box/LRR-repeat (FBD/LRR) protein family. FBD/LRR proteins have been identified in several plant species, for instance, in radish (Zhai et al., 2016), soybean (Chen et al., 2018), tomato (Du et al., 2014; Quintana-Camargo et al., 2015), and peanut (Agarwal et al., 2018). Some members of the FBD/LRR protein family have been found to be associated with pathogen resistance, such as tomato FBD/LRR3, where overexpression of *FBD/LRR3* increased plant resistance to *X. perforans* (Du et al., 2014). One novel *FBD/LRR* gene was identified from a high-density genetic map, and expression of this gene was induced during pathogen resistance responses in peanut (Agarwal et al., 2018). Using proteome analysis, GmFBD/LRR was shown to have a different expression pattern in response to *Phytophthora sojae* infection. The FBD/LRR proteins contain an F-box domain and LRR-repeat domain. The F-box domain is usually associated with a transcription factor that participates in the defense response induced by jasmonic acid (Xu et al., 2009). The LRR-repeat domain has been identified in a functionally and evolutionarily diverse series of proteins. In these proteins, the LRR-repeat domains are essential to protein–protein interactions or signal transduction (Tornero et al., 1996). From the nodulation tests in this study, the NopD mutant decreased NN and NDW in most soybean germplasms, suggesting that NopD could play a positive role in rhizobial infection and nodule formation. Our qRT-PCR results confirmed that NopD can depress *FBD/LRR* expression during infection with the wild-type strain HH103. The haplotype analysis classified the haplotypes into eight types, and Hap1 and Hap2 were the major two types, but Hap1 and Hap2 could not change the gene expression pattern and nodule traits inoculated

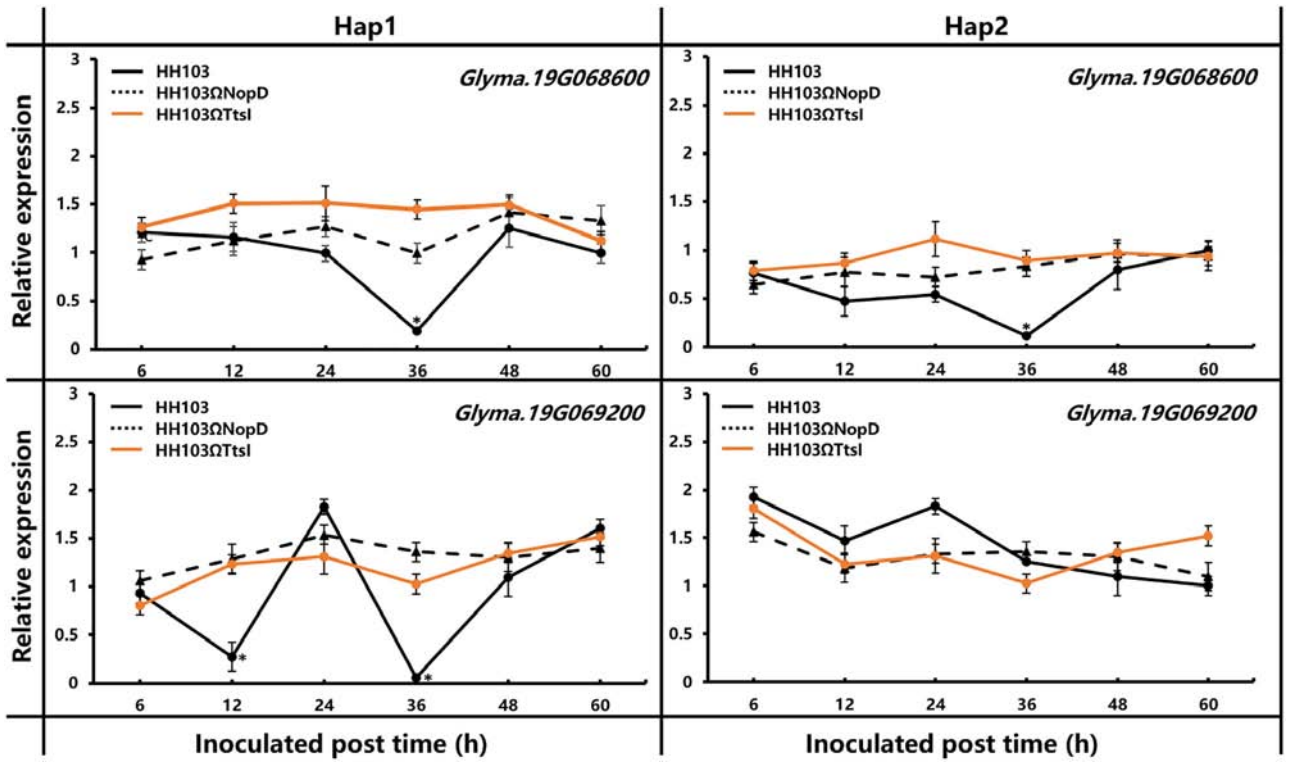


FIGURE 8 | Relative expression of *Glyma.19g066800* and *Glyma.19g069200* was measured by qRT-PCR in RNA extracted from roots of Hap1 and Hap2 accession plants inoculated with *S. fredii* HH103 wild-type or mutants HH103Ω*NopD* and HH103Ω*TtsI*. Uninoculated soybean plants were used as the control. The $2^{-\Delta\Delta Ct}$ method was used to calculate the relative expression levels of candidate genes. Error bars indicate the mean ± standard error of three replications. Significant differences were determined by *t*-tests, and '*' indicated the significant differences ($p \leq 0.05$) at the time point.

with the same rhizobia strain. The finding that *FBD/LRR* encodes an FBD/LRR protein was unexpected. Since some FBD/LRR proteins in other plants are known to be associated with resistance to pathogens, the interaction of NopD with an FBD/LRR protein is interesting, and so further research will be necessary to clarify the nature of this interaction.

Glyma.19g069200 encodes a protein phosphatase 2C (PP2C). The PP2C protein family is widely distributed in almost all plants, whether eukaryotic or prokaryotic plant. PP2C proteins interact with many signaling pathways. Overexpression of the rice *PP2C* gene *OsBIPP2C1* in transgenic tobacco was found to suppress pathogen infection and enhance some abiotic tolerance (Hu et al., 2006). GmPP2C3a, a member of the soybean PP2C protein family, was identified as an antiviral protein that was able to suppress virus infection and spread (Seo et al., 2014). Earlier studies indicated that PP2Cs negatively regulate the mitogen-activated protein kinase (MAPK) pathways in yeast and *Arabidopsis thaliana* (Meskiene et al., 2003), as well as the MAPK pathway triggered by a bacterial flagellin (Cristina et al., 2010). The MAPK pathway might be involved in establishing the rhizobial symbiosis since *MtTDY1*, one of the MAPK pathway genes, was shown to be associated with nodule formation and to regulate the development of the root tip (Schoenbeck et al., 1999). In *Lupinus albus*, bradyrhizobia can activate MAPK genes *SIMK* and *SAMK*, and mutants of SIMK and SAMK inhibited bradyrhizobial infection (Fernandez-Pascual et al., 2006). In soybean, *GMK1* has been identified as a MAPK homolog, and its expression was associated with infection by *Bradyrhizobium japonicum* USDA110 (Lee et al., 2008). T3SS is essential to rhizobial infection, and T3Es can also interact with MAPKs during establishment of symbiosis. The T3E NopP can induce MAPK3 expression at an early stage of rhizobial infection (Wang et al., 2018). In specific *Lotus japonicus* lines of a monogenic-recessive mutant carrying the symbiosis-associated locus, considerable changes to *LjNPP2C1* were observed, suggesting that *LjNPP2C1* may be functional during the early and late nodule development stages (Kapranov et al., 1999). NopL, another T3SS effector of HH103, was shown to suppress the expression of the PP2C-related protein *Glyma.07g099700* during HH103 infection of soybean (Zhang et al., 2018). In this study, expression pattern and haplotype analysis of *PP2C* gene were studied; the result showed *PPC2* to be depressed by the wild-type strain HH103 (which produces NopD) in Hap1 accession, but not in Hap2. This also explained that *PP2C* had a different expression regulation pattern because of the difference in promoter region. Nodule traits indicated that there was a significant difference between Hap1 and Hap2 inoculated with the wild strain HH103, NopD mutant, and

TtsI mutant. These results suggesting that expression of *PP2C* could regulation nodule formation. It is interesting to detect the interaction mechanism between NopD and PP2C. However, further research is essential to fully clarify this interaction.

The T3Es have been shown to be secreted into legume cells *via* rhizobial T3SSs, in a similar way to many gram-negative pathogenic bacteria. In a previous study, the function of rhizobial T3Es in legume cells has been questioned (Büttner and Bonas, 2006). However, more recent research has confirmed that T3 effectors are essential to symbiosis (Schechter et al., 2010; Wenzel et al., 2010). NopD is a conserved T3Er in most rhizobial strains, and our nodulation tests showed that its secretion can positively influence the formation of nodules. The expression of NopD also induced tobacco leaf cell death, giving us evidence that NopD was directly or indirectly recognized by a non-host plant. In our study, NopD influenced the immunity pathway but had no effect on the symbiosis pathway (**Figure 2**). This is similar to XopD, one of the *X. campestris pv. vesicatoria* T3SS effectors, which belongs to the C48 cysteine peptidase family. XopD is able to suppress the host immunity and so promote infection by the pathogens. This suggested that NopD could influence the symbiosis between soybean and rhizobia by affecting signal transduction in the host immunity system. We propose that NopD inhibits *FBD/LRR* and *PP2C* expression, thereby promoting infection by the wild-type HH103.

In this work, *FBD/LRR* and *PP2C* were identified by QTL mapping and can be used to aid further detect of signaling communication induced by NopD. Further identification and clarification of the host genes involved in interactions with rhizobial effector molecules could enhance the understanding of symbiosis establishment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

DX, QC, and CL conceived the study and designed and managed the experiments. ZQ, HJ, RZ, and XW provided soybean seeds. JinW, JieW, CM, JL, LC, and DY performed trials and collected data. JinW, QK, HZ, JieW, ZS, HLiu, ZZ, JZ, HLi, and QW completed statistical analyses of phenotypic data and wrote the manuscript. DX, QC, and JinW participated in correcting the manuscript. All authors contributed to writing the manuscript.

FUNDING

This work was supported by the Ministry of Science and Technology of the People's Republic of China Project (Grant No. 2017YFE0111000), the Europe Horizon 2020 (EUCLEG, 727312), the National Natural Science Foundation of China (Grant Nos. 31771882, 31971899, 31400074, 31471516, 31271747, and 30971809), the Natural Science Foundation of Heilongjiang Province of China (Grant No. ZD201213), the Heilongjiang Postdoctoral Science Foundation (Grant No. LBH-Q16014), the Harbin Science Technology Project (Grant Nos. 2013RFQXJ005 and 2014RFXXJ012), and the Foundation for University Key Teachers from the Education Department of Heilongjiang Province in China (Grant No. 1254G011), University Project of Young Scientist (UNPYSCT-2015011).

ACKNOWLEDGMENTS

We thank Huw Tyson, Ph.D., from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00453/full#supplementary-material

REFERENCES

Agarwal, G., Clevenger, J., Pandey, M. K., Wang, H., Shasidhar, Y., Chu, Y., et al. (2018). High-density genetic map using whole-genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut. *Plant Biotechnol. J.* 16, 1954–1967. doi: 10.1111/pbi.12930

Brensha, W., Kantartzi, S. K., Meksem, K., Grier, R. L. IV, Barakat, A., Lightfoot, D. A., et al. (2012). Genetic analysis of root and shoot traits in the 'Essex'by 'Forrest'recombinant inbred line (RIL) population of soybean [*Glycine max* (L.) Merr.]. *J. Plant Genome Sci.* 1, 1–9. doi: 10.5147/jpgs.2012.0051

Büttner, D., and Bonas, U. (2006). Who comes first? How plant pathogenic bacteria orchestrate type III secretion. *Curr. Opin. Microbiol.* 9, 193–200. doi: 10.1016/j.mib.2006.02.006

Caetano-Anollés, G., and Gresshoff, P. M. (1991). Plant genetic control of nodulation. *Annu. Rev. Microbiol.* 45, 345–382. doi: 10.1146/annurev.mi.45.100191.002021

Caldwell, B. E. (1966). Inheritance of a strain-specific ineffective nodulation in soybeans 1. *Crop Sci.* 6, 427–428.

Canonne, J., Marino, D., Jauneau, A., Pouzet, C., Brière, C., Roby, D., et al. (2011). The *Xanthomonas* type III effector XopD targets the *Arabidopsis* transcription factor MYB30 to suppress plant defense. *Plant Cell* 23, 3498–3511.

Chen, Q. S., Yu, G. L., Zou, J. N., Jing, W. A. N. G., and Qiu, H. M. (2018). GmDRR1, a dirigent protein resistant to *Phytophthora sojae* in *Glycine max* (L.) Merr. *J. Integr. Agric.* 17, 1289–1298.

Crespo-Rivas, J. C., Margaret, I., Pérez-Montaño, F., López-Baena, F. J., Vinardell, J. M., Ollero, F. J., et al. (2007). A pyrF auxotrophic mutant of *Sinorhizobium fredii* HH103 impaired in its symbiotic interaction with soybean and other legumes. *Int. Microbiol.* 10, 169–176. doi: 10.2436/20.1501.01.24

Cristina, M. S., Petersen, M., and Mundy, J. (2010). Mitogen-activated protein kinase signaling in plants. *Annu. Rev. Plant Biol.* 61, 621–649. doi: 10.1073/pnas.252641899

Du, H., Li, W., Wang, Y., and Yang, W. (2014). Identification of genes differentially expressed between resistant and susceptible tomato lines during time-course interactions with *Xanthomonas perforans* race T3. *PLoS One* 9:e93476. doi: 10.1371/journal.pone.0093476

Faruque, O. M., Miwa, H., Yasuda, M., Fujii, Y., Kaneko, T., Sato, S., et al. (2015). Identification of *Bradyrhizobium elkanii* genes involved in incompatibility with soybean plants carrying the *Rj4* allele. *Appl. Environ. Microbiol.* 81, 6710–6717. doi: 10.1128/AEM.01942-15

Fernandez-Pascual, M., Lucas, M. M., de Felipe, M. R., Boscá, L., Hirt, H., and Golvano, M. P. (2006). Involvement of mitogen-activated protein kinases in the symbiosis *Bradyrhizobium-Lupinus*. *J. Exp. Bot.* 57, 2735–2742. doi: 10.1093/jxb/erl038

Figurski, D. H., Meyer, R. J., and Helinski, D. R. (1979). Suppression of ColE1 replication properties by the Inc P-1 plasmid RK2 in hybrid plasmids constructed *in vitro*. *J. Mol. Biol.* 133, 295–318. doi: 10.1016/0022-2836(79)90395-4

Hotson, A., Chosed, R., Shu, H., Orth, K., and Mudgett, M. B. (2003). *Xanthomonas* type III effector XopD targets SUMO-conjugated proteins in planta. *Mol. Microbiol.* 50, 377–389. doi: 10.1046/j.1365-2958.2003.03730.x

Hu, X., Song, F., and Zheng, Z. (2006). Molecular characterization and expression analysis of a rice protein phosphatase 2C gene, *OsBIPP2C1*, and overexpression in transgenic tobacco conferred enhanced disease resistance and abiotic tolerance. *Physiol. Plant.* 127, 225–236. doi: 10.1111/j.1399-3054.2006.00671.x

Hwang, S., Ray, J. D., Cregan, P. B., King, C. A., Davies, M. K., and Purcell, L. C. (2014). Genetics and mapping of quantitative traits for nodule number, weight, and size in soybean (*Glycine max* L.[Merr.]). *Euphytica* 195, 419–434.

Jian, B., Liu, B., Bi, Y., Hou, W., Wu, C., and Han, T. (2008). Validation of internal control for gene expression study in soybean by quantitative real-time PCR. *BMC Mol. Biol.* 9:59. doi: 10.1186/1471-2199-9-59

Jiang, H., Ying, L. Y., Qin, H., Qi, H., Li, C., Wang, N., et al. (2018). Identification of major QTLs associated with first pod height and candidate gene mining in soybean. *Front. Plant Sci.* 9:1280. doi: 10.3389/fpls.2018.01280

Jiménez-Guerrero, I., Pérez-Montaño, F., Medina, C., Ollero, F. J., and López-Baena, F. J. (2015a). NopC is a *Rhizobium*-specific type 3 secretion system effector secreted by *Sinorhizobium (Ensifer) fredii* HH103. *PLoS One* 10:e0142866. doi: 10.1371/journal.pone.0142866

Jiménez-Guerrero, I., Pérez-Montaño, F., Monreal, J. A., Preston, G. M., Fones, H., Vioque, B., et al. (2015b). The *Sinorhizobium (Ensifer) fredii* HH103 type 3 secretion system suppresses early defense responses to effectively nodulate soybean. *Mol. Plant Microbe* 28, 790–799. doi: 10.1094/MPMI-01-15-0020-R

Jiménez-Guerrero, I., Pérez-Montaño, F., Zdyb, A., Beutler, M., Werner, G., Göttfert, M., et al. (2019). GunA of *Sinorhizobium (Ensifer) fredii* HH103 is a T3SS-secreted cellulase that differentially affects symbiosis with cowpea and soybean. *Plant Soil* 435, 15–26.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., et al. (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* 9, 189–197. doi: 10.1093/dnares/9.6.189

Kapranov, P., Jensen, T. J., Poulsen, C., De Bruijn, F. J., and Szczyglowski, K. (1999). A protein phosphatase 2C gene, *LjNPP2C1*, from *Lotus japonicus* induced during root nodule development. *Proc. Natl. Acad. Sci. U.S.A.* 96, 1738–1743. doi: 10.1073/pnas.96.4.1738

Kim, J. G., Stork, W., and Mudgett, M. B. (2013). *Xanthomonas* type III effector XopD desumoylates tomato transcription factor SlERF4 to suppress ethylene responses and promote pathogen growth. *Cell Host Microbe* 13, 143–154. doi: 10.1016/j.chom.2013.01.006

Kim, W. S., and Krishnan, H. B. (2014). A nopA deletion mutant of *Sinorhizobium fredii* USDA257, a soybean symbiont, is impaired in nodulation. *Curr. Microbiol.* 68, 239–246. doi: 10.1007/s00284-013-0469-4

Ladha, J. K., and Peoples, M. B. (eds) (2012). *Management of Biological Nitrogen Fixation for the Development of More Productive and Sustainable Agricultural Systems: Extended Versions of Papers Presented at the Symposium on Biological Nitrogen Fixation for Sustainable Agriculture at the 15th Congress of Soil Science, Acapulco, Mexico, 1994*, Vol. 65, (Dordrecht: Springer). doi: 10.1007/978-94-011-0053-3_1

Lee, H., Kim, J., Im, J. H., Kim, H. B., Oh, C. J., and An, C. S. (2008). Mitogen-activated protein kinase is involved in the symbiotic interaction between *Bradyrhizobium japonicum* USDA110 and soybean. *J. Plant Biol.* 51, 291–296. doi: 10.1007/BF03036129

López Baena, F. J., Vinardell González, J. M., Pérez Montaño, F. D. A., Crespo Rivas, J. C., Bellogín Izquierdo, R. A., Espuny Gómez, M. D. R., et al. (2008). Regulation and symbiotic significance of nodulation outer proteins secretion in *Sinorhizobium fredii* HH103. *Microbiology* 154, 1825–1836. doi: 10.1099/mic.0.2007/016337-0

López-Baena, F., Ruiz-Sainz, J., Rodríguez-Carvajal, M., and Vinardell, J. (2016). Bacterial molecular signals in the *Sinorhizobium fredii*-soybean symbiosis. *Int. J. Mol. Sci.* 17:E755. doi: 10.3390/ijms17050755

Mackey, D., Holt, B. F. III, Wiig, A., and Dangl, J. L. (2002). RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell* 108, 743–754. doi: 10.1016/S0092-8674(02)00661-X

Margaret, I., Becker, A., Blom, J., Bonilla, I., Goesmann, A., Göttfert, M., et al. (2011). Symbiotic properties and first analyses of the genomic sequence of the fast growing model strain *Sinorhizobium fredii* HH103 nodulating soybean. *J. Biotechnol.* 155, 11–19. doi: 10.1016/j.jbiotec.2011.03.016

Marie, C., Deakin, W. J., Viprey, V., Kopcińska, J., Golinowski, W., Krishnan, H. B., et al. (2003). Characterization of Nops, nodulation outer proteins, secreted via the type III secretion system of NGR234. *Mol. Plant Microbe* 16, 743–751. doi: 10.1094/MPMI.2003.16.9.743

Meskiene, I., Baudouin, E., Schweighofer, A., Liwosz, A., Jonak, C., Rodriguez, P. L., et al. (2003). Stress-induced protein phosphatase 2C is a negative regulator of a mitogen-activated protein kinase. *J. Biol. Chem.* 278, 18945–18952. doi: 10.1074/jbc.M300878200

Miwa, H., and Okazaki, S. (2017). How effectors promote beneficial interactions. *Curr. Opin. Plant Biol.* 38, 148–154. doi: 10.1016/j.pbi.2017.05.011

Okazaki, S., Kaneko, T., Sato, S., and Saeki, K. (2013). Hijacking of leguminous nodulation signaling by the rhizobial type III secretion system. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17131–17136. doi: 10.1073/pnas.1302360110

Pérez-Montaño, F., Jiménez-Guerrero, I., Acosta-Jurado, S., Navarro-Gómez, P., Ollero, F. J., Ruiz-Sainz, J. E., et al. (2016). A transcriptomic analysis of the effect of genistein on *Sinorhizobium fredii* HH103 reveals novel rhizobial genes putatively involved in symbiosis. *Sci. Rep.* 6:31592. doi: 10.1038/srep31592

Qi, Z., Huang, L., Zhu, R., Xin, D., Liu, C., Han, X., et al. (2014). A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS One* 9:e104871. doi: 10.1371/journal.pone.0104871

Quandt, J., and Hynes, M. F. (1993). Versatile suicide vectors which allow direct selection for gene replacement in gram-negative bacteria. *Gene* 127, 15–21. doi: 10.1016/0378-1119(93)90611-6

Quintana-Camargo, M., Méndez-Morán, L., Ramirez-Romero, R., Gurrola-Díaz, C. M., Carapia-Ruiz, V., Ibarra-Laclette, E., et al. (2015). Identification of genes differentially expressed in husk tomato (*Physalis philadelphica*) in response to whitefly (*Trialeurodes vaporariorum*) infestation. *Acta Physiol. Plant.* 37:29. doi: 10.1007/s11738-015-1777-z

Ramongolalaina, C., Teraishi, M., and Okumoto, Y. (2018). QTLs underlying the genetic interrelationship between efficient compatibility of *Bradyrhizobium* strains with soybean and genistein secretion by soybean roots. *PLoS One* 13:e0194671. doi: 10.1371/journal.pone.0194671

Riches, D., Porter, I. J., Oliver, D. P., Bramley, R. G. V., Rawnsley, B., Edwards, J., et al. (2013). Soil biological properties as indicators of soil quality in Australian viticulture. *Aust. J. Grape Wine Res.* 19, 311–323. doi: 10.1111/ajgw.12034

Riely, B. K., Larrainzar, E., Haney, C. H., Mun, J. H., Gil-Quintana, E., González, E. M., et al. (2013). Development of tools for the biochemical characterization of the symbiotic receptor-like kinase DMI2. *Mol. Plant Microbe* 26, 216–226. doi: 10.1094/MPMI-10-11-0276

Rodrigues, J. A., Lopez-Baena, F. J., Ollero, F. J., Vinardell, J. M., Espuny, M. D. R., Bellogín, R. A., et al. (2007). NopM and NopD are rhizobial nodulation outer proteins: identification using LC-MALDI and LC-ESI with a monolithic capillary column. *J. Proteome Res.* 6, 1029–1037. doi: 10.1021/pr060519f

Saad, M. M., Kobayashi, H., Marie, C., Brown, I. R., Mansfield, J. W., Broughton, W. J., et al. (2005). NopB, a type III secreted protein of *Rhizobium* sp. strain NGR234, is associated with pilus-like surface appendages. *J. Bacteriol.* 187, 1173–1181. doi: 10.1128/JB.187.3.1173-1181.2005

Schechter, L. M., Guenther, J., Olcay, E. A., Jang, S., and Krishnan, H. B. (2010). Translocation of NopP by *Sinorhizobium fredii* USDA257 into Vigna unguiculata root nodules. *Appl. Environ. Microbiol.* 76, 3758–3761. doi: 10.1128/AEM.03122-09

Schoenbeck, M. A., Samac, D. A., Fedorova, M., Gregerson, R. G., Gantt, J. S., and Vance, C. P. (1999). The alfalfa (*Medicago sativa*) *TDY1* gene encodes a mitogen-activated protein kinase homolog. *Mol. Plant Microbe* 12, 882–893. doi: 10.1094/MPMI.1999.12.10.882

Seo, J. K., Kwon, S. J., Cho, W. K., Choi, H. S., and Kim, K. H. (2014). Type 2C protein phosphatase is a key regulator of antiviral extreme resistance limiting virus spread. *Sci. Rep.* 4:5905. doi: 10.1038/srep05905

Staehelin, C., and Krishnan, H. B. (2015). Nodulation outer proteins: double-edged swords of symbiotic rhizobia. *Biochem. J.* 470, 263–274. doi: 10.1042/BJ20150518

Sugawara, M., Takahashi, S., Umehara, Y., Iwano, H., Tsurumaru, H., Odake, H., et al. (2018). Variation in bradyrhizobial NopP effector determines symbiotic incompatibility with *Rj2*-soybeans via effector-triggered immunity. *Nat. Commun.* 9:3139. doi: 10.1038/s41467-018-05663-x

Tampakaki, A. P. (2014). Commonalities and differences of T3SSs in rhizobia and plant pathogenic bacteria. *Front. Plant Sci.* 5:114. doi: 10.3389/fpls.2014.00114

Tang, F., Yang, S., Liu, J., and Zhu, H. (2016). *Rj4*, a gene controlling nodulation specificity in soybeans, encodes a thaumatin-like protein but not the one previously reported. *Plant Physiol.* 170, 26–32. doi: 10.1104/pp.15.01661

Tennant, J. R. (1964). Evaluation of the trypan blue technique for determination of cell viability. *Transplantation* 2, 685–694. doi: 10.1097/00007890-196411000-00001

Tornero, P., Mayda, E., Gómez, M. D., Cañas, L., Conejero, V., and Vera, P. (1996). Characterization of LRP, a leucine-rich repeat (LRR) protein from tomato plants that is processed during pathogenesis. *Plant J.* 10, 315–330. doi: 10.1046/j.1365-313X.1996.10020315.x

Vance, C. P. (2001). Symbiotic nitrogen fixation and phosphorus acquisition, Plant nutrition in a world of declining renewable resources. *Plant Physiol.* 127, 390–397. doi: 10.1104/pp.010331

Vest, G. (1970). *Rj3* A gene conditioning ineffective nodulation in soybean 1. *Crop Sci.* 10, 34–35. doi: 10.2135/cropsci1970.0011183X001000010013x

Vest, G., and Caldwell, B. E. (1972). *Rj4*-A gene conditioning ineffective nodulation in soybean1. *Crop Sci.* 12, 692–693. doi: 10.2135/cropsci1972.0011183X001200050042x

Vinardell, J. M., Acosta-Jurado, S., Zehner, S., Goettfert, M., Becker, A., Baena, I., et al. (2015). The *Sinorhizobium fredii* HH103 genome: a comparative analysis with *S. fredii* strains differing in their symbiotic behavior with soybean. *Mol. Plant Microbe* 28, 811–824. doi: 10.1094/MPMI-12-14-0397-FI

Vuong, T. D., Nickell, C. D., and Harper, J. E. (1996). Genetic and allelism analyses of hypernodulation soybean mutants from two genetic backgrounds. *Crop Sci.* 36, 1153–1158. doi: 10.2135/cropsci1996.0011183X003600050015x

Wang, J., Wang, J., Liu, C., Ma, C., Li, C., Zhang, Y., et al. (2018). Identification of soybean genes whose expression is affected by the *Ensifer fredii* HH103 effector protein NopP. *Int. J. Mol. Sci.* 19:3438. doi: 10.3390/ijms19113438

Weidner, S., Becker, A., Bonilla, I., Jaenicke, S., Lloret, J., Margaret, I., et al. (2012). Genome sequence of the soybean symbiont *Sinorhizobium fredii* HH103. *J. Bacteriol.* 194, 1617–1618.

Wenzel, M., Friedrich, L., Göttfert, M., and Zehner, S. (2010). The type III–secreted protein NopE1 affects symbiosis and exhibits a calcium-dependent autocleavage activity. *Mol. Plant Microbe* 23, 124–129. doi: 10.1094/MPMI-23-1-0124

Xin, D., Qi, Z., Jiang, H., Hu, Z., Zhu, R., Hu, J., et al. (2016). QTL location and epistatic effect analysis of 100-seed weight using wild soybean (*Glycine soja* Sieb. & Zucc.) chromosome segment substitution lines. *PLoS One* 11:e0149380. doi: 10.1371/journal.pone.0149380

Xin, D. W., Liao, S., Xie, Z. P., Hann, D. R., Steinle, L., Boller, T., et al. (2012). Functional analysis of NopM, a novel E3 ubiquitin ligase (NEL) domain effector of *Rhizobium* sp. strain NGR234. *PLoS Pathog.* 8:e1002707. doi: 10.1371/journal.ppat.1002707

Xu, G., Ma, H., Nei, M., and Kong, H. (2009). Evolution of *F-box* genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proc. Natl. Acad. Sci. U.S.A.* 106, 835–840. doi: 10.1073/pnas.0812043106

Yang, S., Tang, F., Gao, M., Krishnan, H. B., and Zhu, H. (2010). *R* gene-controlled host specificity in the legume-rhizobia symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18735–18740. doi: 10.1073/pnas.1011957107

Zhai, L., Xu, L., Wang, Y., Zhu, X., Feng, H., Li, C., et al. (2016). Transcriptional identification and characterization of differentially expressed genes associated with embryogenesis in radish (*Raphanus sativus* L.). *Sci. Rep.* 6:21652. doi: 10.1038/srep21652

Zhang, Y., Liu, X., Chen, L., Fu, Y., Li, C., Qi, Z., et al. (2018). Mining for genes encoding proteins associated with NopL of *Sinorhizobium fredii* HH103 using quantitative trait loci in soybean (*Glycine max* Merr.) recombinant inbred lines. *Plant Soil* 431, 245–255. doi: 10.1007/s11104-018-3745-z

Zimmer, S., Messmer, M., Haase, T., Piepho, H. P., Mindermann, A., Schulz, H., et al. (2016). Effects of soybean variety and *Bradyrhizobium* strains on yield, protein content and biological nitrogen fixation under cool growing conditions in Germany. *Eur. J. Agron.* 72, 38–46. doi: 10.1016/j.eja.2015.09.008

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 Wang, Wang, Ma, Zhou, Yang, Zheng, Wang, Li, Zhou, Sun, Liu, Li, Chen, Kang, Qi, Jiang, Zhu, Wu, Liu, Chen and Xin. 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 Plant Science

ORIGINAL RESEARCH
published: 15 October 2020
doi: 10.3389/fpls.2020.569948

A Non-destructive Method to Quantify Leaf Starch Content in Red Clover

Lea Antonia Frey[1†], Philipp Baumann[2†], Helge Aasen[3], Bruno Studer[1] and Roland Kölliker[1]*

[1] Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland, [2] Sustainable Agroecosystems, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland, [3] Crop Science, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland

Grassland-based ruminant livestock production provides a sustainable alternative to intensive production systems relying on concentrated feeds. However, grassland-based roughage often lacks the energy content required to meet the productivity potential of modern livestock breeds. Forage legumes, such as red clover, with increased starch content could partly replace maize and cereal supplements. However, breeding for increased starch content requires efficient phenotyping methods. This study is unique in evaluating a non-destructive hyperspectral imaging approach to estimate leaf starch content in red clover for enabling efficient development of high starch red clover genotypes. We assessed prediction performance of partial least square regression models (PLSR) using cross-validation, and validated model performance with an independent test set under controlled conditions. Starch content of the training set ranged from 0.1 to 120.3 mg g^{-1} DW. The best cross-validated PLSR model explained 56% of the measured variation and yielded a root mean square error (RMSE) of 17 mg g^{-1} DW. Model performance decreased when applying the trained model on the independent test set (RMSE = 29 mg g^{-1} DW, R^2 = 0.36). Different variable selection methods did not increase model performance. Once validated in the field, the non-destructive spectral method presented here has the potential to detect large differences in leaf starch content of red clover genotypes. Breeding material could be sampled and selected according to their starch content without destroying the plant.

Keywords: red clover, starch content, hyperspectral imaging, partial least square regression, forage quality, grassland

OPEN ACCESS

Edited by:
Deyue Yu,
Nanjing Agricultural University, China

Reviewed by:
Javaid Akhter Bhat,
Nanjing Agricultural University, China
Fernando Mauricio Ortega-Klose,
Institute of Agricultural Research, Chile

****Correspondence:***
Roland Kölliker
roland.koelliker@usys.ethz.ch

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *02 July 2020*
Accepted: *17 September 2020*
Published: *15 October 2020*

Citation:
Frey LA, Baumann P, Aasen H, Studer B and Kölliker R (2020) A Non-destructive Method to Quantify Leaf Starch Content in Red Clover. Front. Plant Sci. 11:569948. doi: 10.3389/fpls.2020.569948

INTRODUCTION

Temporary and permanent grassland account for roughly 70% of agricultural land and play a significant role in sustainable agriculture worldwide by providing roughage for ruminant livestock production. Pasture and grassland-based agroecosystems maintain carbon balances, nutrient cycles, biodiversity and water quality (Steinfeld et al., 2006). However, they were gradually replaced by

Abbreviations: DW, dry weight; ED, at the end of the day; EN, at the end of the night; MLR, multiple linear regression; PCC, top 50 starch correlated wavelengths; PLSR, partial least square regression; VIP, variable importance in the projection.

intensified production systems, where the high feed energy content required by today's livestock breeds is largely covered through starch from cereals and maize. Starch is an important form of assimilated carbohydrates in plants, which diurnally accumulates in the leaf and is nocturnally mobilized to support growth (Geiger and Servaites, 1994; Stitt et al., 2007; Stettler et al., 2009). The accumulation of starch and its linear degradation at night is thought to be crucial for stable growth and to be directly correlated to plant biomass (Graf et al., 2010; Mugford et al., 2014). However, plant biomass and leaf starch content are not always negatively correlated in species such as birdsfoot trefoil (*Lotus japonicus* L.) or red clover (*Trifolium pratense* L.; Vriet et al., 2010; Ruckle et al., 2017). Starch accumulation and degradation varies not only among plant species, but also among genotypes, seasons, and management regimes (Griggs et al., 2007; Pelletier et al., 2010; Moraes et al., 2013; Liu et al., 2018).

Red clover is one of the most important forage legumes in temperate climates (Taylor, 2008). Its high yield potential, high crude protein content and high digestibility make it an excellent feed, not only for cattle but also for other livestock and poultry (Broderick, 1995; Halling et al., 2001). Red clover has the potential to accumulate up to one third of its leaf dry mass as starch, and some genotypes degrade less than 50% of their starch during the night (Ruckle et al., 2017). Thus, selecting for red clover plants with high starch content and low degradation rates is likely to result in high starch cultivars. These could provide an alternative, high-energy feed source, which would significantly improve sustainability of ruminant livestock production.

Developing a high starch red clover requires a better understanding of the starch metabolism in red clover and an efficient method to quantify starch in leaf tissue. Starch is commonly quantified with an enzymatic method, where leaf samples are flash frozen, ground and weighted before extraction is performed (Hostettler et al., 2011). This procedure is laborious, expensive and involves destructive sampling. A non-destructive method to measure leaf starch content would enable detailed studies of starch turnover in red clover plants, and dynamic changes during plant development could be traced on the same plant throughout the entire season. Specifically, different genotypes could be investigated under different management regimes and across different environments over an extended period of time. Furthermore, this method could be applied in breeding to develop high starch red clover cultivars.

Hyperspectral imaging and near infrared spectroscopy (NIRS) are routinely used to estimate biochemical compounds such as lignin, cellulose, starch, sugars and proteins in numerous crops (Goetz et al., 1990; Hattey et al., 1994; Yoder and Pettigrew-Crosby, 1995). These two methods have largely replaced wet chemistry as the standard analytical procedure for detection and quantification of plant biochemical compounds in the food industry (Card et al., 1988; Barton, 1991). Infrared spectra result from the fundamental vibrational absorptions of photons in the mid-infrared region (500–4000 cm^{-1}, 350–25000 nm) by bonds within specific functional groups of molecules. These absorptions are mirrored to the NIR region (Card et al., 1988). Multivariate statistics, chemometrics, or machine learning methods are then used to quantify and classify specific compounds or properties (Kumar et al., 2001). NIRS or other spectral techniques are most accurate when using dried and homogenized (i.e., milled) plant material. For example, starch has been accurately quantified on dried cotton leaves or dry forage maize using NIRS ($R^2 > 0.9$; Hattey et al., 1994; Hetta et al., 2017; Lu et al., 2017). Estimating chemical compounds with spectral measurements on fresh leaf tissue is often less reliable due to masking effects of light absorption by the cuticle or the leaf water content (Curran et al., 1992; Fourty and Baret, 1998). For successful spectroscopy-based diagnostics using fresh leaf tissue, spectral pre-processing and statistical modeling are essential to at least partially correct for confounding effects (Curran et al., 1992).

The following study aimed at developing a non-destructive spectroscopic method to estimate leaf starch content in fresh leaf tissue of red clover. Although developed in the greenhouse under controlled conditions, such a method could, once validated in the field, enable to monitor starch turnover on the same genotype over a longer period, under different management regimes, and under various environmental conditions.

MATERIALS AND METHODS

Plant Material and Growth Conditions

Leaf starch was determined in two independent sets (i.e., a training set and a test set) of plants grown in two separate experiments. Plants from the two experiments were grown in a completely randomized design. The plants of the test and the training set were grown in spring and autumn 2018, respectively. All plants were clonally propagated using cuts that contained only one shoot and one root meristem to ensure comparable physiological states of all plants. These clonally propagated plants were grown in a climate chamber for 90 days before harvesting (pot size ø 16 cm, height 13 cm; 3:2:1 soil: peat: perlite substrate; photoperiod of 14:10-hour L:D; day temperature 20 ± 2°C, night temperature 15 ± 2°C; relative humidity, 60 ± 10%; fluorescent light bulbs T7 (Phillips, DE), and Grolux® fluorescent lamps (Sylvania, DE) at a ratio of 3:1; total light intensity 200 $\mu mol\ m^{-2}\ s^{-1}$). Samples were taken at the end of the night (EN; before lights were turned on), and at the end of the day (ED; before lights were turned off). Single plants were selected from cultivars from Europe and from breeding germplasm. In addition, nine plants from an advanced breeding population previously shown to have a high variation for leaf starch content, were selected (Ruckle et al., 2017; **Supplementary Table S1**). Wet laboratory measurements were taken on exactly the same material as used for the spectral measurements.

The training set included 18 genotypes, six thereof clonally duplicated (**Supplementary Table S1**). Starch was measured on 15 leaf cuts per plant, taken on the three leaflets of each leaf. The youngest, fully emerged leaf (y), the oldest leaf (o) and three mature leaves (m) were sampled (**Figure 1**), resulting in 360 measurements. The test set included six genotypes; three genotypes were harvested at ED and the same three genotypes at EN (**Supplementary Table S1**). Ten leaf cuts were taken per plant on mature leaves, resulting in 60 measurements.

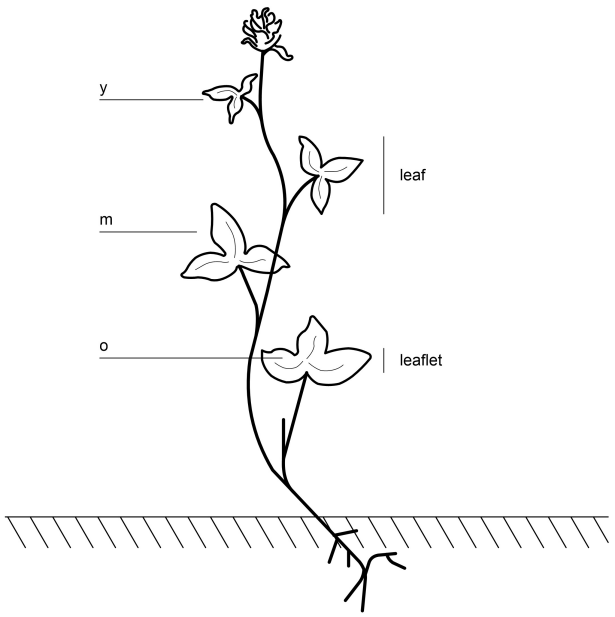


FIGURE 1 | Sampling on red clover plants was performed on the youngest leaf, where three leaflets were fully emerged (y), on the three leaflets of the oldest leaf (o), and on three different mature leaves, in total nine leaflets (m).

Leaf Spectroscopy

Leaflets were cut using a round, sharpened tube with a diameter of 12 mm to standardize leaf area (**Supplementary Figure S1**). These leaf cuts were placed on the matt black surface of the FieldSpec 4 pro device (Analytical Spectral Devices, Boulder, CO, United States; Ely et al., 2019). The device is not influenced by external light sources, potentially enabling the application in field experiments. Radiance between 350 and 2500 nm was measured. The spectrometer's contact probe was fixed on a clamp, and the sample was placed so that no light escaped through the sides. Leaf samples were referenced to a spectralon white reference every fifth recording and the radiance measurements where transformed to reflectance. Immediately after taking spectral measurements, leaf cuts were flash frozen in liquid nitrogen and freeze-dried for 48 h.

After taking spectral measurements, whole plants from the training set were cut 2 cm above ground, flash-frozen and freeze-dried for starch quantification.

Wet Lab Analysis for Starch Quantification

Starch in leaf cuts and whole plants was quantified using a protocol of Hostettler et al. (2011), which was modified and described by Ruckle et al. (2017). Two additional clones of one genotype, not included in the correlation model, were iodine stained to visualize the starch pattern within a plant. Plants were harvested either at ED or at EN, washed with tap water and placed in 80% (v/v) boiling ethanol. After 2 h, when plants were transparent, they were removed and placed in Lugol's solution. After 10 min, the Lugol's solution was rinsed off to destain the non-target areas. The plants were photographed on a light-table (Hostettler et al., 2011).

Statistical Analyses

Statistical analyses were performed using the R statistical software version 3.6.0 (R Core Team, 2019). As assumptions of normality of residuals were not met, an exact Wilcoxon rank sum test was chosen to detect differences between harvest times, and a Kruskal-Wallis test for multiple pairwise comparison at $\alpha = 5\%$.

Pre-processing of Spectral Data

Spectral analysis was realized using the R package simplerspec (Baumann, 2019). The mean reflectance values of 10 measurements per sample were used. Leaf spectra were pre-processed prior to modeling. Gaps between the different detector arrays at $\lambda = 1000$ and at $\lambda = 1800$ nm were splice corrected. Spectra were smoothed with the Savitzky-Golay first derivative filter using a 3rd-order polynomial at a 21-point window (21 nm at a resampled spectrum interval of 1 nm; R package prospectr (Stevens and Ramirez-Lopez, 2020). Spectral pre-processing is crucial to reduce significant noise and baseline drift resulting from light scatter before establishing a correlation model. After smoothing the spectra with Savitzky-Golay, the spectral variables were centered and scaled prior to relating them to leaf starch using partial least squares regression (PLSR), in order to consider variables equally independent of their variation in absolute values. PLS regression is a substantial chemometric method, which can cope with multicollinearity in spectra and delivers robust calibration models with many predictors and few observations (Zhao et al., 2015). To further reduce collinearity in processed spectra, only every forth wavelength was kept for modeling, resulting in 533 spectral predictor variables.

Model Development

Leaf reflectance data from the training set was modeled by PLSR with the orthogonal scores algorithm (also NIPALS; Wold et al., 1983), using the pls R package (Mevik et al., 2020). Separate models were developed with raw or pre-processed spectra as predictors. A 5-times repeated 10-fold cross-validation scheme was used to fit the models, to determine the best number of components (ncomp), and to estimate model performance of the final model. A constant random seed was set for resampling, yielding identical hold out data across all models. Model reliability was assessed by the coefficient of determination (R^2) and the slope (b) of a linear regression with intercept, by the root mean square error (RMSE, Eq. 1), by the bias or mean error (Eq. 2), and by the ratio of performance to deviation (RPD, Eq. 3). The evaluation metrics were calculated by aggregating all holdout predictions from the repeated 10-fold cross-validation (y_i) and corresponding observed values (y_i) grouped by ncomp.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}} \quad (1)$$

$$Bias = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n} \quad (2)$$

$$RPD = \frac{\mathrm{sd}(y_i)}{\mathrm{RMSE}} \quad (3)$$

Variable influence on projection scores (VIP) is a measure of variable importance tailored to PLS regression (Wold et al., 1993; Chong and Jun, 2005). VIP scores were calculated from the PLS regression parameters taking multicollinearity into account, which is likely to occur because of the nature of spectroscopic data. VIP scores are considered as a robust measure to identify relevant predictors, here important wavelengths. A variable with VIP above 1 contributes more than the average variable to the model prediction. The VIP value *vj* was calculated for each wavelength variable *j* as

$$v_j = \frac{\sqrt{p \sum_{a=1}^{A} \left[SS_a \left(w_{aj} / \parallel w_{aj} \parallel \right)^2 \right.}}{\sqrt{\sum_{a=1}^{A} (SS_a)}} \quad (4)$$

where w_{aj} are the PLS regression weights for the ath component for each of the wavelength variables and SS_a is the sum of squares explained by the ath component (Eq. 4). The sum of squares SS_a for the ath component was calculated from the score q_a of the predicted variable y and the t_a scores of the spectral matrix X (Eq. 5):

$$SS_a = q_a^2 t_a^T t^a \quad (5)$$

Variable influence on projection scores scores were also used to filter important predictors with a threshold of VIP > 1 within the training set, and the identified predictors were used to re-calibrate the test set and assess performance. This separation in the VIP filtering by independent tests was needed to avoid overfitting and over-optimistic assessment that typically occur when identifying subsets of features on the modeling data.

In addition to the VIP based filtering, two other procedures were applied for wavelength selection. First, the 50 most relevant wavelengths to estimate starch according to Pearson's correlation coefficient (r) were taken to re-perform PLS regression. Second, four wavelengths that were assigned to starch in previous literature were taken and normalized with the reflection at the wavelength that had the smallest standard deviation across the entire wavelength range, prior to performing a multiple linear regression (MLR; Kumar et al., 2001).

Model Evaluation Using the Test Set

The best training model tuned by cross-validation and refitted on all training data, and the training models with three different wavelength selection (filtering) methods were tested on the independent test data set (60 samples). The predictive ability of these final models was again evaluated using R^2 and RMSE on the test set. Besides these test set predictions, a PLSR model was re-calibrated using only the test data. This re-calibration allowed to determine whether the test set possibly contained different or differently weighted spectral features relevant for starch prediction, so that PLSR training relationships did not generalize to this independent test experiment.

RESULTS

Wet Lab Analysis for Starch Quantification

Starch concentration of samples of the training set harvested at ED ranged from 2.0 to 120.3 mg starch per g dry weight (DW), with a median of 46.3 mg g^{-1} DW. For the samples harvested at the end of the night (EN), starch concentrations ranged from 0.1 to 47.8 mg g^{-1} DW, with a median of 9.6 mg g^{-1} DW. Starch concentrations for the test set ranged from 26.41 to 125.44 mg g^{-1} DW for ED harvested samples, with a median of 66.18 mg g^{-1} DW. Plants harvested at EN had lower starch concentrations, ranging from 3.66 to 79.51 mg g^{-1} DW, with a median of 23.28 mg g^{-1} DW. Differences between samples harvested at ED and EN were statistically significant ($p < 0.5$) for both sets (**Figure 2**). In order to test the reproducibility of the enzymatic method, three technical replicates of the 24 plants of the training set were analyzed, resulting in a standard error (SE) of 0.096 mg g^{-1} DW (data not shown). Also, there was no substantial difference in dry matter content (dry weight/fresh weight) observed between samples from different plants or sampled at different time points, indicating that water content *per se* was not responsible for the differences in starch content observed by spectral analysis (data not shown).

The iodine stained leaves displayed analogous patterns. Plants harvested at ED showed higher starch accumulation than plants harvested at EN, indicated by a dark coloration of the leaves (**Figures 3A,B**). Differences in coloration were not only visible between diurnal time points, but also across and within leaves (**Figures 3B,C**). Dark coloration indicated that old leaves accumulated more starch than young ones. Coloration varied within a plant, showing a clear pattern with young leaves accumulating less starch compared to old leaves. While starch accumulation varied within one leaf, no

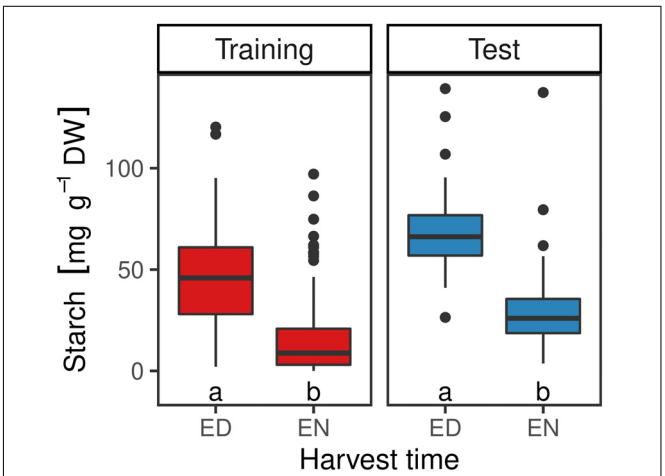


FIGURE 2 | Starch content in mg g^{-1} DW in red clover plants of the training (n = 449) and the test set (n = 60) at the end of the day (ED) and at the end of the night (EN). Medians within the two sets with no letter in common are significantly different (Wilcoxon signed-rank test; α = 5%).



FIGURE 3 | Iodine stained red clover plants (clones from the same genotype) visualizing starch distribution (dark coloration). There are clear differences in starch accumulation between harvest times **(A)** the end of the night (EN), **(B)** the end of the day (ED). Differences in starch accumulation are also visible within one plant **(B)**, and within one leaf **(C)**.

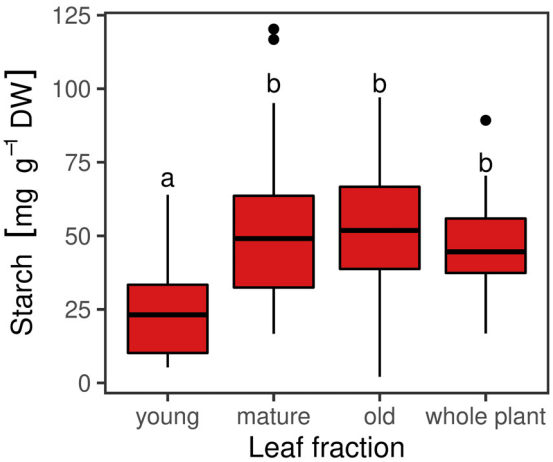


FIGURE 4 | Starch concentrations in different leaf fractions [young (n = 35), mature (n = 108), old (n = 36)] and whole plants (n = 48) of red clover genotypes harvested at the end of the day (ED). Horizontal bars represent medians. Medians with no letter in common are significantly different (Kruskal-Wallis; α = 5%).

clear pattern was distinguishable based on iodine visualization when observing starch accumulation within individual leaflets (**Figure 3C**). Observations in the iodine staining were confirmed by starch quantification in different leaf types. Young leaves had a significantly lower ($p < 0.05$) starch concentration when compared to mature leaves, old leaves or entire plants (**Figure 4**). Some differences between genotypes of the training set harvested at ED were statistically significant ($p < 0.05$), but a high variation within genotypes was observed (**Figure 5**).

Spectral Measurements and Modeling

The average reflectance spectra of the training set revealed similar patterns for both harvest time points ED and EN (**Figure 6A**). VNIR/SWIR (350–2500 nm) spectra had three main absorption regions, around the absorption bands of 700, 1400, and 1900 nm. Savitzky-Golay smoothed spectra showed very similar patterns across the entire wavelength range for samples harvested at ED and those harvested at EN (**Figure 6B**).

The best PLSR training model with pre-processed spectra resulted in an accurate starch prediction for the training set (R^2 = 0.72, RMSE = 13 mg g^{-1} DW, bias = −0.0), using seven PLS components (**Figure 7A**). Five times repeated 10-fold cross validation performed on the same data set revealed a moderate correlation coefficient (R^2_{CV}) of 0.56, a RMSE $_{CV}$ of 17 mg g^{-1}, and a residual bias of −0.2 (**Figure 7B**).

Partial least square regression models modeling using pre-processed spectra performed better than modeling using raw spectra as predictors (**Supplementary Figure S2**). Separating cross-validated predictions by ED and EN resulted in lower correlation coefficients of R^2_{CV} = 0.39 and R^2_{CV} = 0.25 for ED and EN, respectively (**Supplementary Figure S3**). The starch estimates per individual plant had a wide range of R^2_{CV} between 0 and 0.87 (**Supplementary Figure S4**). Including only the most relevant wavelengths for estimating starch content based on filtering training variable importance in the projection (VIP > 1) did not significantly improve prediction performance (**Table 1**). Prediction performance decreased compared to the full spectral model including only the best 50 with starch correlated wavelengths (PCC), and when using MLR with selected starch-assigned wavelengths (scaled), respectively (**Table 1**).

Model Evaluation With an Independent Test Set

Independent test set predictions (n = 57) using the best training PLSR model calibrated with pre-processed spectra (n = 337; ncomp = 7, all wavelengths) yielded a substantially lower R^2 of 0.36 and larger RMSE of 29 mg g^{-1} DW (**Figure 8**). The three training models calibrated with variable selection (VIP > 1, top 50 correlations, and MLR with normalized assigned starch bands) resulted in inferior accuracy when applied to the test set (**Table 1**).

Model development using different filtering methods such as variable importance in the projection (VIP), the top 50 starch correlated wavelengths (PCC), multiple linear regression (MLR) before performing partial least square regression (PLSR). Best model performance of each filtering method determined by five time's repeated 10-fold-cross validation was used to estimate leaf starch content of an independent test set.

DISCUSSION

Hyperspectral imaging on dry homogenized material is a widely used and well established technique, but applying this method on fresh plant material is not yet a standard analytical procedure (Card et al., 1988; Barton, 1991). Consequently, the correlation of hyperspectral measurements and wet lab results for starch reported in this study were clearly lower, when compared to NIRS measurements on dried plant material, where coefficients of determination (R^2) reached 0.99 for nitrogen and starch contend of cotton leaves (Hattey et al., 1994). One challenge using fresh leaf material is the water in the fresh leaves. Liquid water is a strong absorber of the infrared radiation and predominant

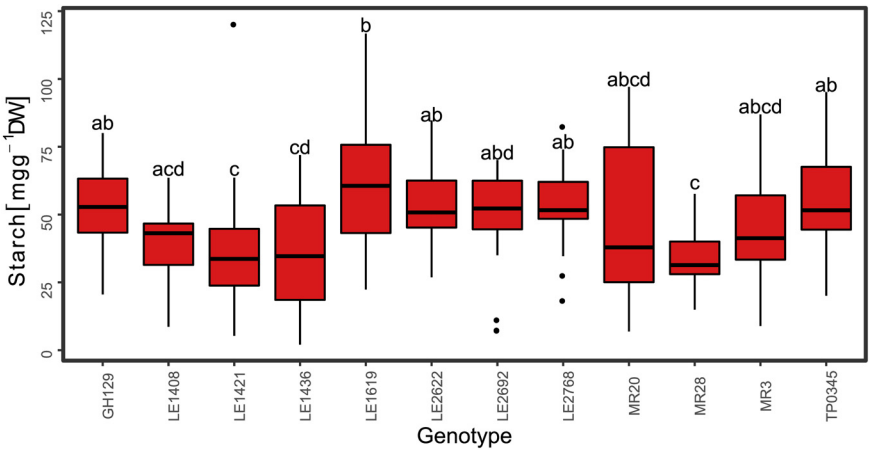


FIGURE 5 | Starch content in mg g^{-1} DW for the red clover genotypes of the training set harvested at the end of the day (ED; n = 226). Medians with no letter in common are significantly different (Kruskal-Wallis, α = 5%).

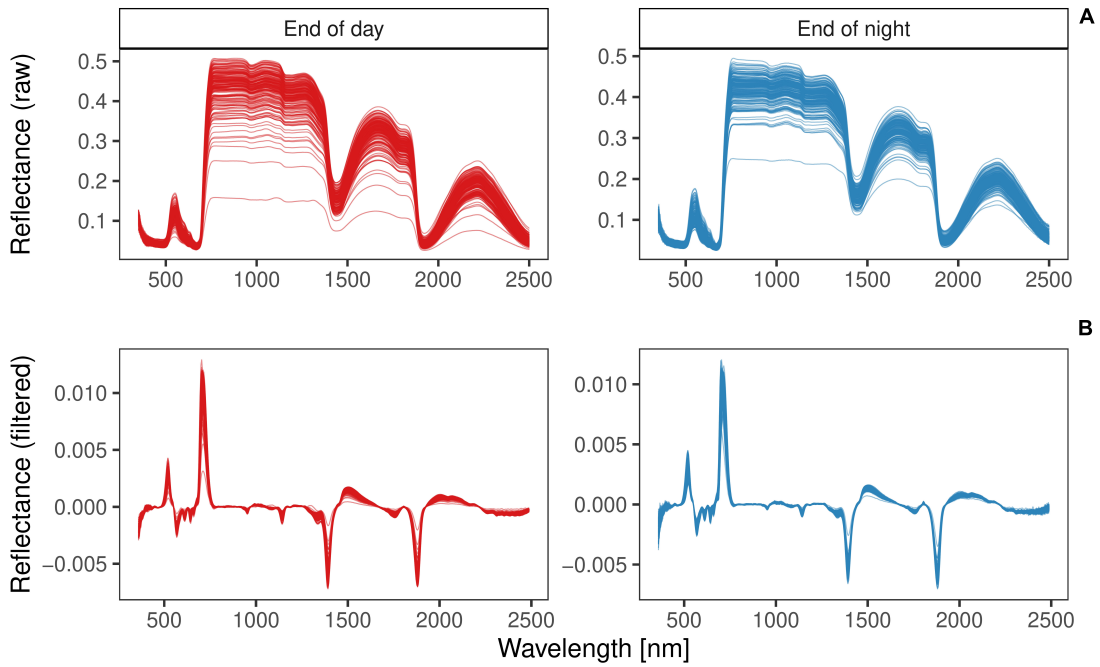


FIGURE 6 | Raw **(A)** and Savitzky-Golay pre-processed reflectance spectra **(B)** of red clover samples of the training set (n = 338) harvested at the end of day (ED; red) or end of night (EN; blue).

bands are in the regions near 1200, 1450, and 1950 nm (Fourty and Baret, 1998), where important wavelengths were present in this study (**Supplementary Figure S5**). It is therefore likely, that water absorption masked the absorption bands of starch molecules, impairing prediction of starch content to some extent (Kumar et al., 2001). Not only water absorption can obscure the starch absorption characteristics, but also the cell structure of fresh plants scattering light as it passes through multiple air and water boundaries. Furthermore, the distribution of starch in fresh leaves is not uniform with respect to the organization of cells and organelles (Kumar et al., 2001). The problems associated with the prediction of starch content in fresh leaves might be reduced, if spectral data is pre-processed (Wold et al., 1993). Indeed, pre-processing of the spectra considerably improved predictive accuracy compared to unprocessed reflectance spectra (**Supplementary Figure S2**), by removing systematic variation in spectra such as light scattering and thereby increasing the signal to noise ratio (Kuhn and Johnson, 2013).

Total starch concentration of the plant material in the training set was between 0.2 and 12% for plants harvested at ED and

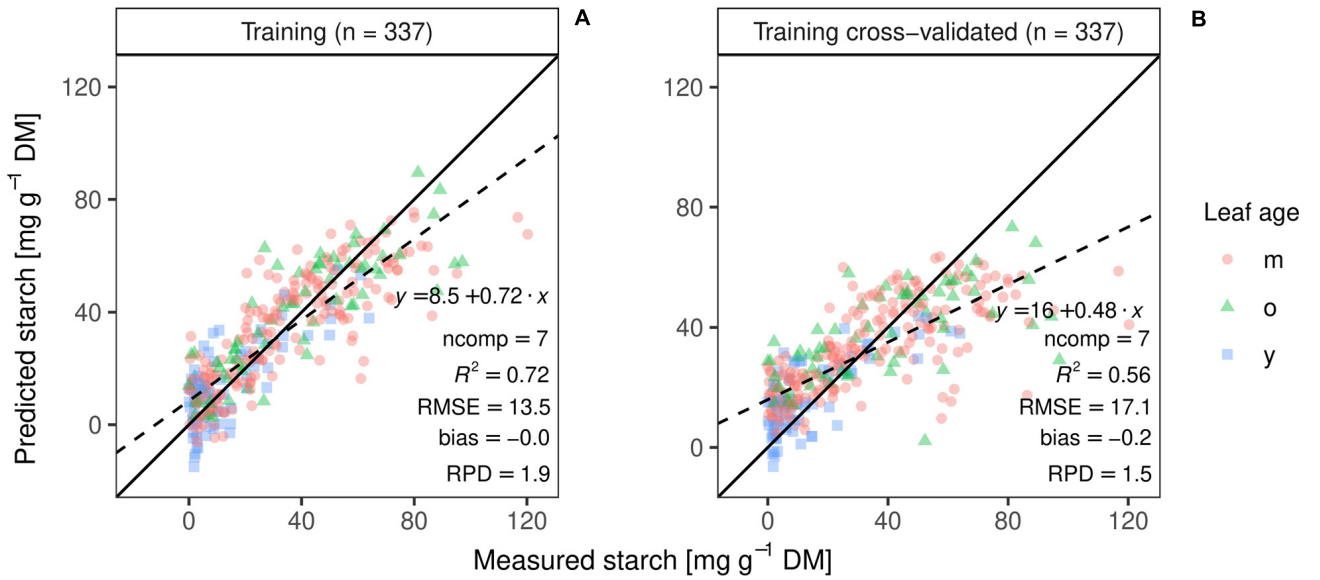


FIGURE 7 | PLS regression of the training set **(A)** and best model performance of the cross-validation on the same set (**B**; ncomp = 9; *n* = 337) Different colors and shapes indicate the age of the leaves, m for matures leaves, o for the oldest leaf, and y for the youngest fully emerged leaf. Regression line (dashed line), 1:1 line (solid black line) and summary statistics are shown.

ranged from 0.01 up to 5% for the plant material harvested at EN. The starch concentrations of the training set were slightly lower than the concentrations of the test set. The total starch concentration was substantially lower than the ones published by Ruckle et al. (2017), where leaf starch concentration ranged from 6 up to 35% for ED harvested plants. This difference occurred most likely due to different growing conditions, since light and temperature have a high impact on starch accumulation.

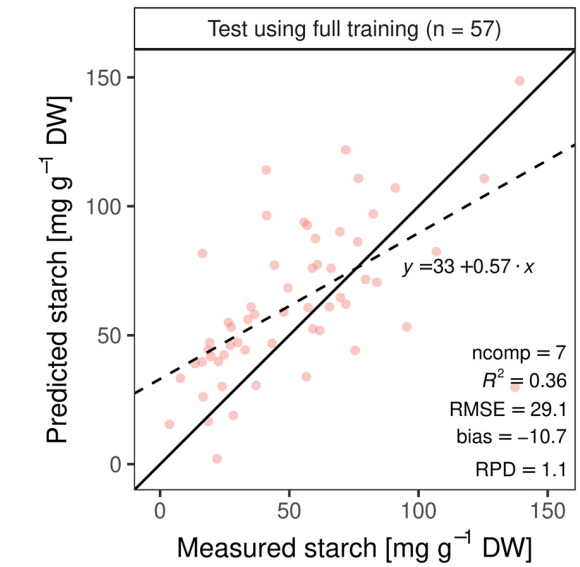


FIGURE 8 | Best performing model from the training set was used to predict leaf starch content in an independent test set (*n* = 57).

Starch content in plant leaves typically varies in a diurnal pattern (Holt and Hilst, 1969). Starch is an important form of assimilated carbohydrates in forage legumes, which is accumulated in the leaf during the day and mobilized during the night to support growth (Stitt et al., 2007). Many studies have shown that starch contents highly depend not only on the diurnal cycle, but also on weather conditions, cutting time, plant fraction, and genotypic variation (Holt and Hilst, 1969; Geiger and Servaites, 1994; Graf et al., 2010; Claessens et al., 2016). In our study, an over 3-fold difference in starch content was observed between ED harvested plants and plants harvested at EN (**Figure 2**).

Mean genotypic differences for the training set ranged from 31.6 to 59.7 mg g^{-1} DW for the ED harvested plants (**Figure 5**) and from 2.2 to 37.6 mg g^{-1} DW for the EN harvested plants (data not shown), respectively, showing high variation within genotypes. The best PLSR training model explained 56% (R^2 = 0.56) of the measured starch variation with an RMSE of 17 mg g^{-1} DW. The ratio of performance to deviation (RPD) followed the trend indicated by R^2 values (**Figures 7**, **8** and **Supplementary Figure S2, S3**). The cross-validated overall bias was almost zero for the training set, while predictions on the test set had a bias of −10.7 mg g^{-1} DW.

These results imply that the developed vis–NIR PLSR model can predict differences between harvest time points and differences between extreme genotypes (**Supplementary Figure S4**). Nevertheless, 56% of starch variation explained by our model is lower than the proportions reported by Shorten et al. (2019). They used hyperspectral imaging systems (550–1700 nm) to estimate more than ten different quality compounds in perennial ryegrass (*Lolium perenne* L.). Low and high weight sugars were estimated separately and best model prediction for

TABLE 1 | Model performance using different filtering methods.

Model strategy	Training set		Test set		Description
	R^2	RSME	R^2	RSME	
VIP > 1	0.58	16.6	0.37	38.0	Filtering based on importance of the training variable in the projection (VIP)
PCC	0.53	18.5	0.37	32.0	Filtering based on the top 50 with starch correlated wavelengths
MLR	0.26	21.8	0.18	38.0	Reflectance at 556, 702, 1300, and 1960 nm divided by reflectance at 670 nm (minimum standard deviation)

Model development using different filtering methods such as variable importance in the projection (VIP), the top 50 starch correlated wavelengths (PCC), multiple linear regression (MLR) before performing partial least square regression (PLSR). Best model performance of each filtering method determined by five time's repeated 10-fold-cross validation was used to estimate leaf starch content of an independent test set.

the high weight sugars using PLS regression resulted in an R^2 of 0.68 and a RMSE of 19.9 mg g^{-1}. Assigning two-third of the data to calibration and using the remaining data for validation resulted in a slightly lower model performance [R^2 = 0.63 and RMSE of 21.6 mg g^{-1} (Shorten et al., 2019)]. Filtering spectral variables by a variable importance in the projection threshold (VIP > 1) did not considerably improve model performance (**Table 1**). This is in contrast to comparable studies where selecting important wavelengths improved model accuracy and reduced the redundancy effects of wavelengths, which had low weight in the model (Wold et al., 1993; Chong and Jun, 2005). Our results indicate that restricting PLSR with a subset of important spectral variables is not sufficient to estimate starch with equal effectiveness compared to the full-range vis—NIR data, confirming that many spectral features are important for starch prediction. For example, the wavelengths near 550, 770, 850, 1440, 1920 nm, from 1650 nm to 1850 and 2160 nm had a relatively high model contribution for estimating starch content in the training set (VIP analysis, **Supplementary Figure S5**). The red-edge region around 700 nm, where a local maximum of the first derivative is located and which is typically indicative for chlorophyll, had relatively low model importance. However, adjacent wavelengths to the red-edge were moderately important. The highest VIP in the training set was around 550 nm. This region was shown to be the second most important region in the vis—NIR for the spectral estimation of total carbon, nitrogen, leaf mass per unit area, protein and nitrate from wet leaves of 8 crop species (Ely et al., 2019). Starch absorbance in fresh leaves was further associated with wavelengths in the regions of 556, 702, 1300, and 1960 nm (Curran et al., 1992). These absorptions partly corresponded with the VIP patterns across wavelengths for the data of the present study. In addition, performing explanatory inference for spectra-model-compound linking is hampered by spectral overlaps due to dominant water bands and signals of other compounds related to starch. In fact, plant leaves contain many biochemical compounds with vis—NIR absorption regions that overlap with starch absorptions, or whose concentration directly or indirectly correlate to starch, such as cellulose, water or lignin, all having signals from O-H vibrations in the regions around 1450 and 1940 nm. Curran et al. (2001) performed both a correlative and stepwise regression analysis between 12 abundant structural, productive and storage compounds, and vis—NIR first derivative spectra of ground and dried slash pine needles. Among the components tested, starch exhibited the lowest coefficient of determination with first derivative spectra, and selected starch wavelengths were 1208, 1418, and 2172 nm, whereas 978 and 1208 nm were linked to starch absorption features. Native plant starch consists of a variable ratio of amylose and amylopectin. Amylose content in various mixtures was accurately discriminated with vis—NIR reflectance, showing major spectral feature differences between 1700 and 1800 nm in the pure form (Fertig et al., 2004). We found two VIP peaks with moderate importance (around 1.2), that might be linked to amylose and amylopectin signals. PLSR and the variable importance analysis were thus able to explain a significant proportion of the starch variability.

A model built from a single set of training observations is often not adequate to predict an independent data set (Naes and Martens, 1985; Kuhn and Johnson, 2013). If a model is tested on the same data that was used to fit the model, performance is often overestimated (Kuhn and Johnson, 2013). Our study showed that the cross-validated PLSR model underestimated high starch contents (**Figure 7**). The independent dataset from the second experiment (test set) allowed us to further validate model performance, in addition to cross-validation during training. As expected, the test prediction resulted in a 1.7-fold increase in RMSE (**Figure 8**). Moreover, models including only a subset of wavelengths were validated on the test set, resulting in a lower predictability (**Table 1**). The VIP analysis of the two independent sets (training and test) indicated that some important wavelengths regions occurred in both sets, but with different VIP magnitudes (**Supplementary Figures S5, S6**). For example, the absorption feature near 1450 nm was less important for the test set model fitting, compared to the model developed for the training set. Further, the training model had important features between 500 and 750 nm, whereas the re-calibrated test model had important wavelengths below 500 nm. These differences in VIP magnitudes and the additional regions relevant for prediction partly explain the poorer prediction performance of the test set when applying the training model. Despite the fact that two of the three genotypes from the test set were included in the training set, the spectra and models had only limited generalization capacity for starch contents.

Recalibration using only test data led to a slight decrease in RMSE compared to test prediction, but this substantially reduced bias. Thus, a new calibration may be needed for each independent trial or the current red clover starch spectral library needs to be augmented with more measurements from different independent trials with both genotypic and phenotypic variance in starch. Various environmental growth conditions influence

starch accumulation and can thereby mask genotypic effects (Holt and Hilst, 1969). Based on the our results, we suggest follow up research that combines statistical methods to optimize knowledge transfer from such plant spectral libraries to new trials under substantial genotype x environment interaction. Thereby, the focus should be to find a trade-off between accuracy and the amount of new reference measurements needed, depending on the breeding application context. This library could be enlarged with data from new clover trials, so that it is continuously augmented with more genotypic and phenotypic variance in starch. We suggest to test methods from transfer learning research, which exploit different mechanisms to extract and transfer relevant information of collections of training data to new and partly related prediction tasks or application domains (Pan and Yang, 2010). For example, memory-based learning that constrains models based spectral similarity (Ramirez-Lopez et al., 2013), or data-driven search algorithms which filter relevant observations from spectral libraries that yield good performance on new local target samples (Lobsey et al., 2017) are candidate approaches that may be worth testing. We conducted all measurements under controlled conditions and leaves were completely removed from the plants. As a next step, it is crucial to evaluate the method under field condition.

Despite the relatively low prediction accuracy, performance of the best PLSR training model was sufficient to detect differences between red clover genotypes with very high or very low levels of starch content. Therefore, once validated in the field, the method may be valuable for in large-scale QTL studies in bi-parental populations based on strongly contrasting parental starch phenotypes. Furthermore, it has the potential to directly assist phenotypic selection in the breeding of high-energy red clover cultivars.

The success of breeding forage crops with increased energy content was previously demonstrated by breeding perennial ryegrass cultivars with high levels of water-soluble carbohydrates (WSC). These WSC cultivars can substantially increase animal performance and nitrogen use efficiency in pasture-based animal production systems (Rasmussen et al., 2009). Red clover and ryegrasses are often cultivated in mixtures, not only due to their attractive diet composition, but also due to the transfer of N between species. In addition, grass-clover mixtures require fewer pesticide and herbicide applications, and protect soils against erosion (Dhamala et al., 2016; McKenna et al., 2018). Therefore, high starch red clover cultivars in mixtures with high WSC ryegrasses appear a particularly promising option, which brings us one-step closer toward an environmental sustainable feed production meeting the high energy requirements of modern livestock production.

CONCLUSION

This study is unique in developing and testing a non-destructive method to predict leaf starch content in red clover plants. The described method is suitable to differentiate between high and low starch content in red clover genotypes. Unfortunately, model performance is not sufficient to trace small changes in starch accumulation. Therefore, the method is only partially suited to monitor starch metabolism in detail or to investigate the effect of environmental influences or management regimes throughout an entire season on the same plant. We suggest follow up studies to enlarge the current red clover starch spectral library by means of additional measurements from different independent trials, covering both genotypic and phenotypic variation in starch and to validate the method under field conditions. Currently, the level of resolution is sufficient for the method to differentiate high variance in starch and thus, can be integrated into existing breeding programs to get a rough estimate on starch levels of different red clover cultivars under controlled conditions.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/3598699#.XhOLvdko_d4.

AUTHOR CONTRIBUTIONS

RK, BS, and LF conceived the study. LF and PB performed the research, analyzed the data, and drafted the manuscript. RK, BS, and HA assisted with data analysis, interpreting the results, and drafting the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement no. 727312 (project: "EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency"). This research was also partially supported by the Coop Research Program for Sustainability in Food Value Chains of the World Food System Center at ETH Zurich (project: "High-energy red clover").

ACKNOWLEDGMENTS

We thank Dr. Michael Ruckle for initiating the research on high starch red clover and for helping with the wet lab starch analyses, Prof. Dr. Achim Walter for the allocation of the FieldSpec 4 pro device, and Dr. Dániel Carrera and Ms. Verena Knorst for the excellent technical support. This manuscript was released as a pre-print at Research Square (Frey et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.569948/full#supplementary-material

REFERENCES

Barton, F. E. (1991). New methods for the structural and compositional analysis of cell walls for quality determinations. *Anim. Feed Sci. Technol.* 32, 1–11. doi: 10.1016/0377-8401(91)90003-B

Baumann, P. (2019). *Philipp-Baumann/Simplerspec: Beta Release Simplerspec 0.1.0 for Zenodo*. Genèva: Zenodo.

Broderick, G. A. (1995). Desirable characteristics of forage legumes for improving protein utilization in ruminants. *J. Anim. Sci.* 73, 2760–2773. doi: 10.2527/1995.7392760x

Card, D. H., Peterson, D. L., Matson, P. A., and Aber, J. D. (1988). Prediction of leaf chemistry by the use of visible and near infrared reflectance spectroscopy. *Remote Sens. Environ.* 26, 123–147. doi: 10.1016/0034-4257(88)90092-2

Chong, I.-G., and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* 78, 103–112. doi: 10.1016/j.chemolab.2004.12.011

Claessens, A., Castonguay, Y., Bertrand, A., Bélanger, G., and Tremblay, G. F. (2016). "Breeding for improved nonstructural carbohydrates in alfalfa," in *Breeding in a World of Scarcity*, eds I. Roldán-Ruiz, J. Baert, and D. Reheul (Cham: Springer International Publishing), doi: 10.1007/978-3-319-28932-8

Curran, P. J., Dungan, J. L., Macler, B. A., Plummer, S. E., and Peterson, D. L. (1992). Reflectance spectroscopy of fresh whole leaves for the estimation of chemical concentration. *Remote Sens. Environ.* 39, 153–166. doi: 10.1016/0034-4257(92)90133-5

Curran, P. J., Dungan, J. L., and Peterson, D. L. (2001). Estimating the foliar biochemical concentration of leaves with reflectance spectrometry. *Remote Sens. Environ.* 76, 349–359. doi: 10.1016/S0034-4257(01)00182-1

Dhamala, N. R., Rasmussen, J., Carlsson, K., Søegaard, K., and Eriksen, J. (2016). Nitrogen fixation in red clover grown in multi-species mixtures with ryegrass, chicory, plantain and caraway. *Grassl. Sci. Eur.* 21, 576–578.

Ely, K. S., Burnett, A. C., Lieberman-Cribbin, W., Serbin, S. P., and Rogers, A. (2019). Spectroscopy can predict key leaf traits associated with source-sink balance and carbon-nitrogen status. *J. Exp. Bot.* 70, 1789–1799. doi: 10.1093/jxb/erz061

Fertig, C. C., Podczeck, F., Jee, R. D., and Smith, M. R. (2004). Feasibility study for the rapid determination of the amylose content in starch by near-infrared spectroscopy. *Eur. J. Pharm. Sci.* 21, 155–159. doi: 10.1016/j.ejps.2003.09.011

Fourty, T., and Baret, F. (1998). On spectral estimates of fresh leaf biochemistry. *Int. J. Remote Sens.* 19, 1283–1297. doi: 10.1080/014311698215441

Frey, L. A., Baumann, P., Aasen, H., Studer, B., and Kölliker, R. (2020). A non-destructive method to quantify leaf starch content in red clover. *Res. Sq.* [preprint]. doi: 10.21203/rs.2.22508/v2

Geiger, D. R., and Servaites, J. C. (1994). Diurnal regulation of photosynthetic carbon metabolism in C3 plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 45, 235–255. doi: 10.1146/annurev.pp.45.060194.001315

Goetz, A. F. H., Gao, B. C., Wessman, C. A., and Bowman, W. D. (1990). "Estimation of biochemical constituents from fresh, green leaves by spectrum matching techniques," in *Proceedings of the 10th Annual International Symposium on Geoscience and Remote Sensing*, College Park, MD.

Graf, A., Schlereth, A., Stitt, M., and Smith, A. M. (2010). Circadian control of carbohydrate availability for growth in *Arabidopsis* plants at night. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9458–9463. doi: 10.1073/pnas.0914299107

Griggs, T. C., MacAdam, J. W., Mayland, H. F., and Burns, J. C. (2007). Temporal and vertical distribution of nonstructural carbohydrate, fiber, protein, and digestibility levels in Orchardgrass Swards. *Agron. J.* 99, 755–763. doi: 10.2134/agronj2006.0036

Halling, M., Hopkins, A., Nissinen, O., Paul, C., Tuori, M., and Soelter, U. (2001). Forage legumes - productivity and composition. *Landbauforsch. Volk* 234, 5–15.

Hattey, J. A., Sabbe, W. E., Baten, G. D., and Blakeney, A. B. (1994). Nitrogen and starch analysis of cotton leaves using near infrared reflectance spectroscopy (NIRS). *Commun. Soil Sci. Plant Anal.* 25, 1855–1863. doi: 10.1080/00103629409369158

Hetta, M., Mussadiq, Z., Wallsten, J., Halling, M., Swensson, C., and Geladi, P. (2017). Prediction of nutritive values, morphology and agronomic characteristics in forage maize using two applications of NIRS spectrometry. *Acta Agric. Scand. Sect. B Soil Plant Sci.* 67, 326–333. doi: 10.1080/09064710.2017.1278782

Holt, D. A., and Hilst, A. R. (1969). Daily variation in carbohydrate content of selected forage crops. *Agron. J.* 61, 239–242. doi: 10.2134/agronj1969.00021962006100020020x

Hostettler, C., Kölling, K., Santelia, D., Streb, S., Kötting, O., and Zeeman, S. C. (2011). "Analysis of starch metabolism in chloroplasts," in *Chloroplast Research in Arabidopsis*, ed. R. P. Jarvis (Totowa, NJ: Humana Press), 387–410. doi: 10.1007/978-1-61779-237-3_21

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Kumar, L., Schmidt, K., Dury, S., and Skidmore, A. (2001). "Imaging spectrometry and vegetation science," in *Imaging Spectrometry Basic*, eds F. D. van der Meer and S. M. de Jong (Dordrecht, GE: Kluwer Academic Publishers), 111–154. doi: 10.1007/0-306-47578-2_5

Liu, W., Su, J., Li, S., Lang, X., and Huang, X. (2018). Non-structural carbohydrates regulated by season and species in the subtropical monsoon broad-leaved evergreen forest of Yunnan province, China. *Sci. Rep.* 8, 1–10. doi: 10.1038/s41598-018-19271-8

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., and Hedley, C. B. (2017). rs-local data-mines information from spectral libraries to improve local calibrations. *Eur. J. Soil Sci.* 68, 840–852. doi: 10.1111/ejss.12490

Lu, X., Sun, J., Mao, H., Wu, X., and Gao, H. (2017). Quantitative determination of rice starch based on hyperspectral imaging technology. *Int. J. Food Prop.* 20, 1037–1044. doi: 10.1080/10942912.2017.1326058

McKenna, P., Cannon, N., Conway, J., and Dooley, J. (2018). The use of red clover (*Trifolium pratense*) in soil fertility-building: a review. *Field Crops Res.* 221, 38–49. doi: 10.1016/j.fcr.2018.02.006

Mevik, B.-H., Wehrens, R., and Hovde Liland, K. (2020). *pls: Partial Least Squares and Principal Component Regression. R Package Version 2.7-3*. Available online at: https://CRAN.R-project.org/package=pls

Moraes, M. G., Chatterton, N. J., Harrison, P. A., Filgueiras, T. S., and Figueiredo-Ribeiro, R. C. L. (2013). Diversity of non-structural carbohydrates in grasses (Poaceae) from Brazil. *Grass Forage Sci.* 68, 165–177. doi: 10.1111/j.1365-2494.2012.00883.x

Mugford, S. T., Fernandez, O., Brinton, J., Flis, A., Krohn, N., Encke, B., et al. (2014). Regulatory properties of ADP glucose pyrophosphorylase are required for adjustment of leaf starch synthesis in different photoperiods. *Plant Physiol.* 166, 1733–1747. doi: 10.1104/pp.114.247759

Naes, T., and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Commun. Stat. Simul. Comput.* 14, 545–576. doi: 10.1080/03610918508812458

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pelletier, S., Tremblay, G. F., Bélanger, G., Bertrand, A., Castonguay, Y., Pageau, D., et al. (2010). Forage nonstructural carbohydrates and nutritive value as affected by time of cutting and species. *Agron. J.* 102, 1388–1398. doi: 10.2134/agronj2010.0158

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., and Scholten, T. (2013). The spectrum-based learner: a new local approach for modeling soil vis-NIR spectra of complex datasets. *Geoderma* 195–196, 268–279. doi: 10.1016/j.geoderma.2012.12.014

Rasmussen, S., Parsons, A. J., Xue, H., and Newman, J. A. (2009). High sugar grasses - harnessing the benefits of new cultivars through growth management. *Proc. N. Z. Grassl. Assoc.* 71, 167–175. doi: 10.33584/jnzg.2009.71.2746

Ruckle, M., Meier, M., Frey, L., Eicke, S., Kölliker, R., Zeeman, S., et al. (2017). Diurnal leaf starch content: an orphan trait in forage legumes. *Agronomy* 7, 1–15. doi: 10.3390/agronomy7010016

Shorten, P. R., Leath, S. R., Schmidt, J., and Ghamkhar, K. (2019). Predicting the quality of ryegrass using hyperspectral imaging. *Plant Methods* 15, 1–12. doi: 10.1186/s13007-019-0448-2

Steinfeld, H., Gerber, P., Wassenaar, T. D., Castel, V., Rosales, M. M., et al. (2006). *Livestock's Long Shadow: Environmental Issues and Options*. Rome: Food and Agriculture Organization of the United Nations.

Stettler, M., Eicke, S., Mettler, T., Messerli, G., Hörtensteiner, S., and Zeeman, S. C. (2009). Blocking the metabolism of starch breakdown products in *Arabidopsis* leaves triggers chloroplast degradation. *Mol. Plant* 2, 1233–1246. doi: 10.1093/mp/ssp093

Stevens, A., and Ramirez-Lopez, L. (2020). *An Introduction to the Prospectr Package. R Package Version 0.2.0. R Package Vignette.*

Stitt, M., Gibon, Y., Lunn, J. E., and Piques, M. (2007). Multilevel genomics analysis of carbon signalling during low carbon availability: coordinating the supply and utilisation of carbon in a fluctuating environment. *Funct. Plant Biol.* 34, 526–549. doi: 10.1071/FP06249

Taylor, N. L. (2008). A century of clover breeding developments in the United States. *Crop Sci.* 48, 1–13. doi: 10.2135/cropsci2007.08.0446

Vriet, C., Welham, T., Brachmann, A., Pike, M., Pike, J., Perry, J., et al. (2010). A suite of *Lotus japonicus* starch mutants reveals both conserved and novel features of starch metabolism. *Plant Physiol.* 154, 643–655. doi: 10.1104/pp.110.161844

Wold, S., Jonsson, J., Sjörström, M., Sandberg, M., and Rännar, S. (1993). DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta* 277, 239–253. doi: 10.1016/0003-2670(93)80437-P

Wold, S., Martens, H., and Wold, H. (1983). "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils*, eds B. Kågström and A. Ruhe (Berlin: Springer), 286–293. doi: 10.1007/BFb0062108

Yoder, B. J., and Pettigrew-Crosby, R. E. (1995). Predicting nitrogen and chlorophyll content and concentrations from reflectance spectra (400-2500 nm) at leaf and canopy scales. *Remote Sens. Environ.* 53, 199–211. doi: 10.1016/0034-4257(95)00135-N

Zhao, N., Wu, Z., Zhang, Q., Xin-yuan, S., Ma, Q., and Qiao Y.-j. (2015). Optimization of parameter selection for partial least squares model development. *Sci. Rep.* 5:11647. doi: 10.1038/srep11647

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 Frey, Baumann, Aasen, Studer and Kölliker. 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 Plant Science

ORIGINAL RESEARCH
published: 21 January 2021
doi: 10.3389/fpls.2020.578121

The History of Domestication and Selection of Lucerne: A New Perspective From the Genetic Diversity for Seed Germination in Response to Temperature and Scarification

*Wagdi Ghaleb†, Lina Qadir Ahmed†, Abraham J. Escobar-Gutiérrez and Bernadette Julier**

INRAE, URP3F, Lusignan, France

OPEN ACCESS

Edited by:
Deyue Yu,
Nanjing Agricultural University, China

Reviewed by:
Zan Wang,
Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, China
Javaid Akhter Bhat,
Nanjing Agricultural University, China

***Correspondence:**
Bernadette Julier
bernadette.julier@inrae.fr

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *30 June 2020*
Accepted: *22 December 2020*
Published: *21 January 2021*

Citation:
Ghaleb W, Ahmed LQ, Escobar-Gutiérrez AJ and Julier B (2021) The History of Domestication and Selection of Lucerne: A New Perspective From the Genetic Diversity for Seed Germination in Response to Temperature and Scarification. Front. Plant Sci. 11:578121. doi: 10.3389/fpls.2020.578121

Lucerne (*Medicago sativa*), a major perennial pasture legume, belongs to a species complex that includes several subspecies with wild and cultivated populations. Stand establishment may be compromised by poor germination. Seed scarification, deterioration and temperature have an impact on germination. The objective of this study was to analyse the genetic diversity of lucerne germination in response to three factors: (1) temperature, with seven constant temperatures ranging from 5 to 40°C, was tested on 38 accessions, (2) seed scarification was tested on the same accessions at 5 and 22°C, (3) seed deterioration was tested on two accessions and two seed lots at the seven temperatures. The germination dynamics of seed lots over time was modelled and three parameters were analysed: germinability (germination capacity), maximum germination rate (maximum% of seeds germinating per time unit), and lag time before the first seed germinates. Seed scarification enhanced germinability at both temperatures and its effect was much higher on *falcata* and wild *sativa* accessions. Incomplete loss of the hardseededness trait during domestication and selection is hypothesised, indicating that the introduction of wild material in breeding programmes should be followed by the selection for germinability without scarification. Seed lots with altered germinability had low germination at extreme temperatures, both cold and hot, suggesting that mild temperatures are required to promote germination of damaged seed lots. A large genetic diversity was revealed for germination (both capacity and rate) in response to temperature. All accessions had an optimal germination at 15 or 22°C and a poor germination at 40°C. The *sativa* varieties and landraces had a high germination from 5 to 34°C while the germination of *falcata* and the wild *sativa* accessions were weakened at 5 or 34°C, respectively. These differences are interpreted in terms of adaptation to the climate of their geographical origin regions in order to

escape frost or heat/drought risks. These new findings give insights on adaptation and domestication of lucerne in its wide geographic area. They suggest further improvement of germination is needed, especially when introducing wild material in breeding pools to remove scarification requirements and to limit differences in response to temperature.

Keywords: alfalfa, breeding, genetic variation, *Medicago sativa*, seed dormancy, adaptation

INTRODUCTION

Lucerne (*Medicago sativa* L.) is a perennial forage legume that holds a major place in cultivated grasslands worldwide. It is used for ruminant feeding, given its high yield and nutritional properties. The numerous ecosystem services it provides (Julier et al., 2017) are important in the present phase of agroecological transition (Lamichhane et al., 2018), and mainly result from its ability to fix atmospheric nitrogen. *M. sativa*, which was first cultivated in its centre of origin (Near-East to Middle-East) 9,000 years ago, is the oldest cultivated forage crop and one of the oldest crops in the world. It spread several times to North Africa and Europe through invaders (Medes, Romans and Moors) and was introduced to Asia (India and China) more than 2,000 years ago (Basigalup et al., 2014). From the 15th Century, lucerne became a popular crop in France and other European countries (De Serres, 1600) from where it accompanied explorers and migrants to America from the 16th Century onwards and to austral countries by the 19th Century (Prosperi et al., 2014). Lucerne is now grown in temperate regions worldwide and has recently became a cornerstone crop in organic agriculture. Its direct annual world value is expected to grow significant over the coming years.

Medicago sativa is a complex of several diploid or autotetraploid perennial and allogamous subspecies (ssp), wild or cultivated, that freely intercross at the same ploidy level (Small and Jomphe, 1989). Ssp *sativa* with coiled, non-spiny pods, and purple or purple yellow flowers, includes the vigorous cultivated autotetraploid varieties, some wild populations in Spain but also the diploid wild form *coerulea*. The diploid ssp *glomerata* exhibits coiled pods and yellow flowers. Ssp *falcata* with yellow flowers and slicked-shaped pods, originates from Northern Eurasia and is cold tolerant. The populations that result from hybridisation between ssp *sativa* or *coerulea* and ssp *falcata* are sometimes described as ssp x*varia* (Small and Brookes, 1990). As shown by their variegated flower colour, most European varieties exhibit introgression from ssp *falcata* that confers cold tolerance and has enabled the selection of highly productive varieties adapted to continental or subarctic climates.

As many other annual and perennial legumes, in particular of the *Papilionoideae* subfamily (Baskin and Baskin, 2014; Acosta et al., 2020), it has been well documented that physical dormancy due to hardseededness is present in lucerne to various degrees (Acharya et al., 1999). Hardseededness is characterised by a water-impermeable seed coat that ensures a maintenance of the population in the soil and prevents its germination up to several years (Baskin and Baskin, 2004; Finch-Savage and Leubner-Metzger, 2006). Despite the uncontested ecological advantages associated with the different types of seed dormancy (Baskin and Baskin, 2014), it hampers crop establishment due to low percentage of germinating seeds and slow rates of seedling emergence (Acosta et al., 2020). This type of physical dormancy is broken by different treatments to the seeds, including scarification, which makes imbibition possible by overcoming coat's impermeability to water (Baskin and Baskin, 2014). Even in cultivated varieties, the presence of hard seeds is common in commercial lots, although regulations and the market impose a minimum percentage of such seeds.

Seed deterioration, often generated by ageing, induces a reduction of germination capacity, or germinability, (Mcdonald, 1999). Evaluation of lot germinability along seed ageing is performed by the seed industry (Corbineau, 2012) and the institutions in charge of *ex situ* seed conservation (Genetic Resources Centres; Fu et al., 2015). Seed deterioration is regulated by seed storage conditions (temperature, humidity and oxygen partial pressure) but also by endogenous factors (Zhou et al., 2020). Little information is available on the impact of temperature on germination of seed lots whose germination capacity is deteriorated.

Temperature is the most important environmental factor regulating both timing and rates of seed germination. The role of temperature in seed germination is related to enzyme activity and membrane permeability, which impact respiratory metabolism, and thus the rate of germination (Bewley and Black, 1994). Plant species have different critical temperature ranges and basic thermal requirement to complete specific phenophases or the whole life cycle (Luo, 2011). Seed germination in response to temperature can be summarised by the cardinal temperatures (minimum, optimum and maximum temperatures; Bewley and Black, 1994; Shafii and Price, 2001; Durr et al., 2015). The response of different species to temperature during germination has been a topical issue for a long time (Bewley and Black, 1994; Baskin and Baskin, 2014; Durr et al., 2015; Tribouillois et al., 2016). Major differences among species depend on the climatic conditions where the species grow or originate. Cultivated species germinate faster and under a wider range of temperatures than wild species, likely as a result of human selection (Durr et al., 2015). Information on genetic variation for germination capacity is useful in breeding programmes especially if wild accessions are used to adapt varieties to new climatic conditions or practices related to ecosystem services, or for the preservation of genetic diversity in genetic resources centres. Studies at the intra-specific level are not frequent in pasture species (Moot et al., 2000; Mcgraw et al., 2003; Monks et al., 2009; Ahmed, 2015; Ahmed et al., 2019). In the model legume species *Medicago truncatula*, a significant diversity for germination was evidenced among accessions (Brunel et al., 2009) and within a mapping population (Dias et al., 2011). Knowledge of physical dormancy, seed

deterioration and genetic diversity in response to temperature, is of paramount importance in the management of lucerne breeding programmes.

In this paper, the objective was to demonstrate a genetic diversity of *Medicago sativa* for seed germination in response to constant temperatures (from 5 to 40°C). We performed experiments in controlled conditions with thirty-eight accessions of various geographical origins, from the subspecies *sativa, falcata* or *glomerata*, wild or cultivated. The effects of scarification and deterioration level of seed lots on germination at different temperatures were also tested. All the results were analysed and interpreted in terms of domestication, selection and breeding.

MATERIALS AND METHODS

Plant Material

In this study, seeds of thirty-eight accessions of wild populations, landraces and varieties of lucerne were evaluated as follows: 16 *sativa* varieties, 11 *sativa* landraces, four *sativa* wild accessions, one *falcata* variety, five *falcata* wild populations and one *glomerata* wild population (**Table 1**). In addition to the spp. membership, the status (wild, landrace or variety) that traces the breeding effort possibly applied to germination traits was used to define five groups of accessions: '*falcata* + *glomerata* wild', '*sativa* wild', '*falcata* variety', '*sativa* landraces' and '*sativa* varieties'. Seeds were obtained from INRAE Genetic Resources Centre in Lusignan, France, or from breeders. They were conserved at 5°C and 30% relative humidity until they were used. For two landraces, Flamande from France and Gabès from Tunisia, deteriorated seed lots, characterised by poor germinability, were also included. Only two wild *falcata*, the *falcata* variety and the wild *glomerata* were diploid accessions and all the others were tetraploid accessions, so the effect of ploidy was not studied here.

The material was used to test three aspects of seed germination. (1) The effect of scarification, in interaction with genetic diversity, was tested on the whole set of 38 accessions at two temperatures (5 and 22°C). Scarification consisted in scrubbing the seeds between two sheets of sandpaper (grade 150) for 30 s. (2) The effect of the deterioration level of seed lots on germination response to seven constant temperatures (5, 10, 15, 22, 28, 34 or 40°C) was analysed on two accessions by comparing, within each accession, seed lots with high or low germinability. (3) The genetic diversity for germination in response to these seven constant temperatures was evaluated on scarified seeds of the 38 accessions.

Germination Tests

For each accession and treatment, four repetitions of 100 seeds were prepared. For temperatures between 5 and 34°C, the 100 seeds of each repetition were put in 90 mm diameter Petri dishes containing two sheets of autoclave-sterilised Whatman paper (ref. 3645 Whatman, France) humidified with 5 ml of deionised and autoclave-sterilised water. For the experiment at 40°C, which was conducted in a growth chamber with 80% relative humidity, seeds were placed on two sheets of Whatman papers in transparent crystal polystyrene boxes (180 mm × 120 mm × 50 mm; GEVES trademark, Loire Plastic, France; Brunel et al., 2009) with 15 ml of sterilised water.

For each temperature, the design was a random complete block design (RCBD) with four blocks. The dishes of each block were placed in a vented plastic box. The experiment was conducted in walk-in growth chambers at constant temperature in the dark and with vapour pressure deficit <1 kPa. Temperature of the chambers were checked by six thermocouples placed at different positions within the useful volume, logged every 20 s. Temperature and relative humidity in the chambers were recorded every minute during the germination experiment.

Germination counting was carried out at variable time intervals that depended on temperature treatments: once a day at 5 and 40°C and twice a day for temperatures from 10 to 34°C. Seeds were considered as germinated when the radicle or the cotyledon leaves had protruded out of the seed and was at least 2 mm long (Bewley and Black, 1994). After each counting, if needed, deionised and autoclave-sterilised water was added to ensure non-limiting moisture. Counting was stopped when no new germination was observed during five consecutive visits. At 5°C, the experiment lasted more than 3 months. Whenever needed, counting was performed in the growth chamber in order to keep the seeds at a constant temperature and a high air humidity.

Modelling of the Dynamics of Seed Lot Germination

For both the effect of seed deterioration and the genetic diversity, the cumulative percentages of germination values over time were fitted to a modified non-rectangular hyperbole (Thornley et al., 1976), whose parameters have an eco-physiological meaning:

$$y = \left(\frac{1}{2\theta}\right) \cdot \left(\alpha \cdot (t - t_c) + y_{\max} - \sqrt{\left(\alpha . (t - t_c) + y_{\max}\right)^2 - \left(4.\theta.\alpha . (t - t_c) . y_{\max}\right)}\right).$$

Where y is the cumulated seed germination (%); θ is a unitless parameter that determines the sharpness of the knee of the curve; α is the maximum germination rate (% of germinating seeds per hour); t is the time (hour); t_c is the lag to start germination (hour) during which no seed germinates; and y_{max} is the asymptotic maximum seed germination (%). In order to limit the number of parameters to be estimated, and after several tests, θ was fixed to 0.97.

Statistical Analyses

For the effect of scarification in each of the two temperatures, the observed maximum seed germination (hereafter named as germinability) was submitted to an analysis of variance with the effects of scarification, group, accession within group, repetition, scarification × group and scarification × accession within group. The function *aov* of R was used (R Core Team, 2019).

TABLE 1 | Description of the 38 accessions of *Medicago sativa*.

Code	Subspecies	Type	Ploidy	Name	Country	Geographic coordinates		Autumn dormancy
						Longitude	Latitude	
35	*falcata*	Wild	2x	Romanica	Russia	55.74	37.61	
36	*falcata*	Wild	2x	Quasifalcata	Russia	55.74	37.61	
38	*falcata*	Wild	4x	Krasnokutskaya	Russia	50.91	47.05	
39	*falcata*	Wild	4x	Maron	France	48.63	6.04	
40	*falcata*	Wild	4x	Malzeville	France	48.70	6.18	
37	*glomerata*	Wild	2x	Glomerata	France	43.60	5.75	
20	*sativa*	Wild	4x	Monte Oscuro	Spain	41.72	−0.47	
21	*sativa*	Wild	4x	Villanueva de Jara	Spain	39.44	−1.95	
22	*sativa*	Wild	4x	Villamajor	Spain	41.68	−0.77	
23	*sativa*	Wild	4x	Pancrudo	Spain	40.76	−1.03	
34	*falcata*	Variety	2x	Anik	Canada			1.0
05	*sativa*	Landrace	4x	Flamande*	France	48.85	2.35	4.0
08	*sativa*	Landrace	4x	Poitou	France	46.58	0.34	4.0
09	*sativa*	Landrace	4x	Provence	France	43.53	5.44	6.0
24	*sativa*	Landrace	4x	Gabès*	Tunisia	33.90	10.1	9.0
25	*sativa*	Landrace	4x	Cremonese	Italy	45.13	10.02	6.0
26	*sativa*	Landrace	4x	Crau	France	43.55	4.85	
27	*sativa*	Landrace	4x	Demnate3	Morocco	31.73	−7.00	9.0
28	*sativa*	Landrace	4x	Dra15	Morocco	30.33	−5.83	9.0
29	*sativa*	Landrace	4x	Atlas	Morocco	31.11	−7.87	9.0
30	*sativa*	Landrace	4x	Ziz10	Morocco	31.93	−4.42	9.0
31	*sativa*	Landrace	4x	Baghdadi	Iran	32.13	48.18	9.0
01	*sativa*	Variety	4x	Banat VS	Serbia			5.0
02	*sativa*	Variety	4x	SW Nexus	Sweden			4.0
03	*sativa*	Variety	4x	Luzelle	France			3.0
04	*sativa*	Variety	4x	Holyna	Czech Republic			3.7
06	*sativa*	Variety	4x	Lukal	France			4.0
07	*sativa*	Variety	4x	Ludelis	France			4.0
10	*sativa*	Variety	4x	Barmed	France			7.0
11	*sativa*	Variety	4x	Harpe	France			4.0
12	*sativa*	Variety	4x	Orca	France			4.5
13	*sativa*	Variety	4x	Radius	Poland			
14	*sativa*	Variety	4x	FG-CO416C4164	United States			4.0
15	*sativa*	Variety	4x	Alforex6	United States			4.0
16	*sativa*	Variety	4x	Gongnong1	China			
17	*sativa*	Variety	4x	Magna790	United States/ARG			7.0
18	*sativa*	Variety	4x	Bauding	China			
19	*sativa*	Variety	4x	Picena GR	Italy			7.4

*The subspecies, population type (variety, landrace or wild population), name of accession and country of origin are given. For the wild populations, the geographic coordinates of the collection site are indicated. Autumn dormancy is given when known. *Two seed lots of these accessions were tested.*

For the germinability of the two seed lots with high and low germination percentages, a non-linear mixed model (Pinheiro and Bates, 2000) was used to estimate the three parameters (α, t_c and y_{max}) and to test the seed lot factor, in Flamande and Gabès separately, at each temperature. The functions *nlsList*, *nlme* and *update* of package *nlme* were sequentially used (Pinheiro et al., 2020) in R.

For the 38 accessions, the non-linear mixed model was similarly used to test the effects of group and accession. A principal component analysis (PCA) was conducted to assess the variability in the three germination parameters of the model in response to temperature (function *PCA* of R).

RESULTS

Scarification Effect on the Percentage of Seed Germination

On average for the 38 accessions, scarification had a significant positive effect on germinability at both temperatures 5 and 22°C

TABLE 2 | Mean observed germinability over 4 biological repetitions, with or without scarification at 5 and 22°C for the five groups of accessions.

Group	5°C		22°C	
	Scarified	Non-scarified	Scarified	Non-scarified
Falcata + *glomerata* wild	50.5	27.5	63.1	29.1
Sativa wild	84.1	48.9	77.1	47.3
Falcata variety	27.5	28.8	74.3	25.8
Sativa landrace	86.1	75.3	86.7	77.5
Sativa variety	86.8	80.0	88.9	83.0
Effects of factors in analysis of variance				
Scarification	572.0***		670.0***	
Group	1585.1***		843.9***	
Accession within group	78.0***		47.0***	
Scarification × group	99.4***		155.2***	
Scarification × accession within group	10.0***		6.0***	
Residual standard error	4.24		4.58	

*For the analyses of variance in each temperature, the F value and significance are given. ***Significant at $P < 0.001$.*

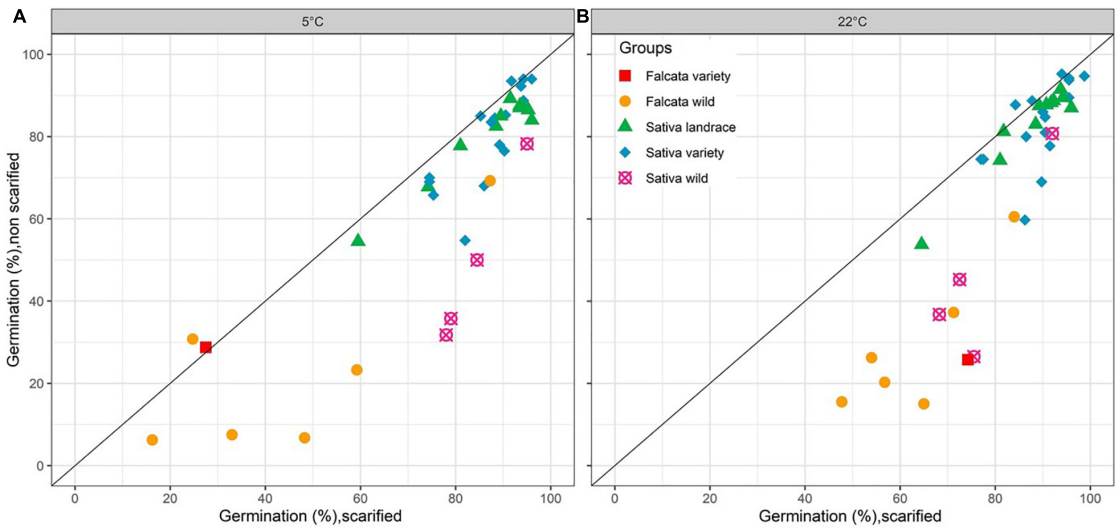


FIGURE 1 | Germinability (observed data, average of 4 biological repetitions) of 38 lucerne populations from 5 different groups with or without scarification at 5°C **(A)** and 22°C **(B)**.

(**Table 2**, **Figure 1**, and **Supplementary Table 1**). Germinability of non-scarified seeds varied from 6 to 94% at 5°C and from 15 to 95% at 22°C, depending on the accession. With scarification, germinability ranged from 16 to 96% at 5°C and from 15 to 99% at 22°C. The difference of germinability between non-scarified and scarified seeds (**Supplementary Table 1**) was significant for 12 accessions at both 5 and 22°C, for two other accessions at 5°C only and for three other accessions at 22°C only. This difference reached a maximum of 46 points of percentage at 5°C (Villanueva de Jara, *sativa* wild) and 50 points of percentage at 22°C (Anik, *falcata* variety).

An analysis of variance showed that the effects of group, accession within group, scarification, and their interactions were significant at both temperatures (**Table 2** and **Figure 1**). The scarification effect was positive for the wild accessions, either for *falcata* + *glomerata* (+34 points of percentage) or *sativa* (+30 points of percentage) at 22°C. At 5°C, the figures for the same groups were +23 and +35 points of percentage. For the *falcata* variety, the effect of scarification was large at 22°C (+48 points of percentage) but nil at 5°C, where the germination was very poor without scarification but also with scarification. For the *sativa* landraces and varieties, scarification had a small but significant effect at both temperatures (+7 and +6 points of percentage at 5°C and 22°C, respectively).

Effect of Seed Lot Deterioration on the Germination Capacity in Response to Temperature

This study was carried out on two landrace accessions, with two seed lots per accession, one with a high germinability and the other with a poor germinability.

The seed lots confirmed their expected germinability, either high or low, at all temperatures (**Table 3** and **Figure 2**). The two seed lots of the population Flamande had higher lags to start germination (t_c) and lower germination rates (α) than the two seed lots of the population Gabès. The low germinability of deteriorated seed lots were associated with low germination rates (α; **Table 3**). At 34°C, the germinability of the deteriorated seed lot, compared to the highly germinating seed lot, was much lower for Flamande than for Gabès.

Genetic Diversity for Germination in Response to Temperature

A wide diversity for the three parameters of temperature response (germinability, maximum germination rate and lag to start germination) was evidenced among the 38 lucerne accessions (**Supplementary Table 2** and **Supplementary Figure 1**, **Figure 3**). All parameters at all temperatures shown a significance among accessions (not shown). The effect of group of accessions were also significant except for the maximum germination rates at 5 and 28°C (**Table 4**).

In the PCA analysis performed with the values of germinability (y_{max}) and the estimated values of maximum germination rate (α) and lag to start germination (t_c) estimated at seven temperatures (5, 10, 15, 22, 28, 34, 40°C; **Figure 3**), the first two axes were significant (eigenvalue > 1) and contributed for 75.5% of total variance (**Supplementary Table 3** and **Figure 3A**). Axis 1 explained 65.7% of the total variability and was mainly correlated to germination rates at 10, 5 and 15°C and germinability at 10 and 34°C (**Figures 3A,B** and **Supplementary Table 2**). Axis 2, which explained 9.8% of the total variability, was mainly correlated to the germination rates at 40 and 34°C (**Supplementary Table 3**). Germinability at all temperatures was positively correlated with maximum germination rate at all temperatures (**Figure 3A**). Lag to start germination from 10 to 34°C was negatively correlated with both germinability and maximum germination rate (**Figure 3A**). As a whole, the accessions that had high germinability germinated shortly and quickly after seed imbibition.

All germination parameters were optimum at 15 or 22°C for all groups of accessions (**Table 4** and **Supplementary Figure 1**). With progressing colder temperatures, the lag to start germination increased and the germination rate decreased for all groups of accessions, while germinability decreased for the *falcata* and *glomerata* accessions only. With progressing higher temperatures, the germinability of the cultivated *sativa* accessions remained high up to 34°C, while their lag to start germination increased and their germination rate decreased. At 40°C, all germination parameters of the cultivated *sativa* accessions were altered. For the wild accessions and the *falcata* variety, germinability was still high at 28°C, but not at 34°C nor at 40°C, and lag to start germination and germination rate were strongly modified. **Figure 3B** demonstrates that the groups behaved in a different way in the PCA. The accessions of both ssp *sativa* landrace and ssp *sativa* variety groups showed a high germinability as well as a high germination rate and a small lag to start germination. This means that these cultivated accessions are able to germinate shortly after the conditions are adequate and that all the seeds germinate at once. Moreover, it is to be noted that, as shown from their position on axis 2 of the PCA, the ssp *sativa* accessions with an autumn dormancy between 3 and 7 had a lower germination rate at 34°C than the non-dormant accessions with dormancy 9 which originate from Tunisia, Morocco or Iran (**Figure 3B**). Compared to the cultivated *sativa* accessions, accessions of the wild *falcata* and *glomerata* group showed a low germinability and very low germination rate as well as a high lag to start germination. The *falcata* variety (Anik, no 34) also showed a low germinability, a low germination rate and a long lag to start germination. However, the ssp *falcata* accession Krasnokutskaya (no 38) from Russia showed a higher germinability, a higher germination rate and a lower lag to start germination than the other accessions of the ssp *falcata* wild group (**Figure 3B**). The wild *sativa* accessions that originated from Spain had a higher germinability and a higher germination rate than the wild *falcata* accessions. Interestingly, the wild accessions that originated from Spain had only slightly lower germination rate and lag to start germination than cultivated *sativa* accessions from 5 to 22°C, but all their germination parameters were strongly altered at 34°C, with a low germinability, a long lag to start germination and a slow germination rate (**Table 3**). At 5°C, the germinability of the *sativa* accessions, whether wild or cultivated, were higher than that of the *falcata* and *glomerata* accessions.

DISCUSSION

This study aims at demonstrating that lucerne seed germination is influenced by physical treatment (scarification), seed lot deterioration and genetic diversity in response to temperature. It also illustrates the interaction between these factors and gives insights about the history of domestication and selection.

Scarification Effect on the Percentage of Seed Germination

Physical seed dormancy caused by the water-impermeable seed coat that inhibits water absorption is well described in many species (Baskin and Baskin, 2004, 2014), including legumes (Taylor, 2005; Smykal et al., 2014). Scarification, by a physical or chemical alteration of the seed coat, increases seed germination. Our results on 38 accessions of lucerne indicate that, on average, non-scarified seeds of lucerne had a lower germinability than scarified seeds at both temperatures (5 and 22°C). This scarification effect on germination response depended on accessions. It was higher for the wild accessions, either *falcata* or *sativa* ssp. than for the cultivated accessions. A large effect of mechanical scarification had previously been reported on one naturalised American yellow-flowered population (Narem, 2009). This could mean that the domestication syndrome which includes germination, well described on legumes (Smykal et al., 2014), also applies to lucerne. Our results suggest that the seed dormancy mechanisms carried by wild accessions have been lost during domestication and breeding. This loss enables the seeds of cultivated accessions to germinate as soon as they are

TABLE 3 | Germinability (y_{max} in°C), lag to start germination (t_c in h) and maximum germination rate (α in% of germinating seeds per h) for the two seed lots (high and low germination) of two lucerne accessions (Flamande and Gabès) at 7 temperatures.

Temperature (°C)	Parameter	Flamande high germ.	Flamande low germ.	Gabès high germ.	Gabès low germ.
5	α	0.793	0.330 NS	0.806	0.609 NS
	t_c	81.5	90.2**	79.6	83.6 NS
	y_{max}	91.2	74.3***	85.6	72.2***
10	α	1.485	0.687*	1.506	0.988**
	t_c	28.4	29.2 NS	27.7	26.8 NS
	y_{max}	95.5	72.9***	91.1	68.8***
15	α	3.613	1.516***	4.672	2.188***
	t_c	16.0	14.5 NS	13.1	11.1*
	y_{max}	100.3	69.3***	89.6	73.3***
22	α	6.246	2.042**	7.560	4.913*
	t_c	12.3	12.3 NS	12.3	11.5*
	y_{max}	94.5	76.2***	85.9	67.6***
28	α	3.591	1.532 NS	7.506	3.035***
	t_c	14.0	14.3 NS	12.0	11.0***
	y_{max}	96.7	65.2***	84.1	61.5***
34	α	1.886	0.616 NS	7.421	3.083***
	t_c	16.6	16.7 NS	13.3	11.8***
	y_{max}	90.5	33.0***	83.1	66.2***
40	α	0.067	0.090 NS	0.120	0.052***
	t_c	51.6	20.1**	4.0	10.6 NS
	y_{max}	11.4	3.7*	17.5	5.7**

*For each accession, the significance of the difference between seed lots is given. ***, **, *: significant at $P < 0.001$, $P < 0.01$, $P < 0.05$, respectively. NS: non-significant ($P > 0.05$).*

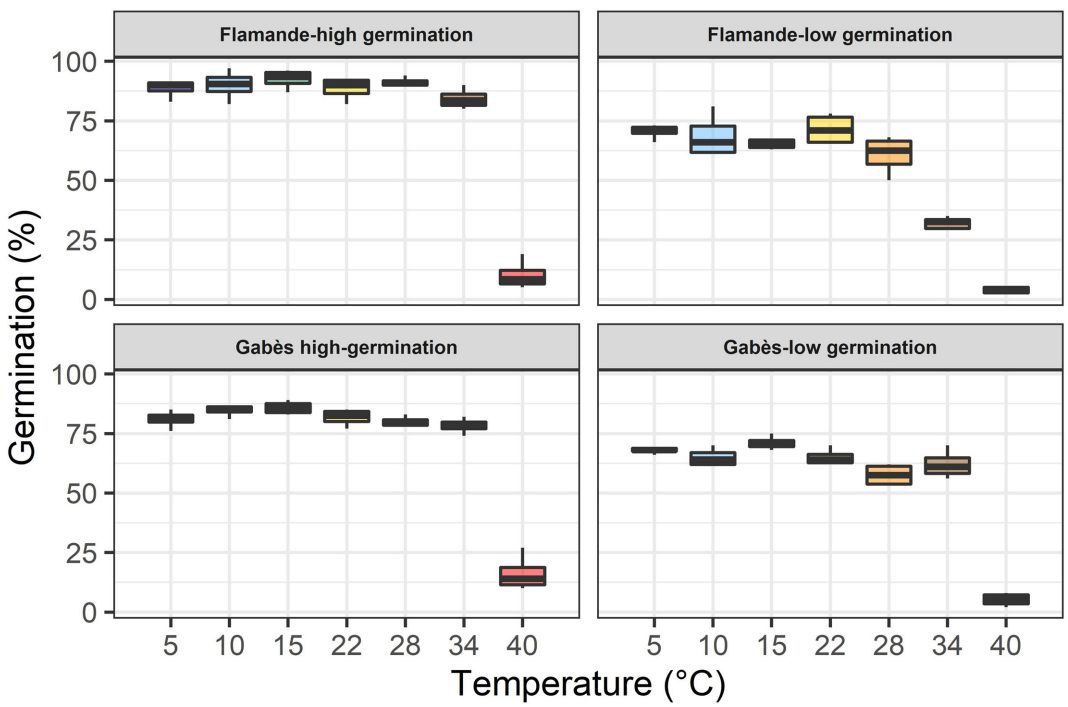


FIGURE 2 | Boxplot showing average germinability (y_{max} parameter of the model) of two seed lots for two accessions (Flamande and Gabès) in response to temperature.

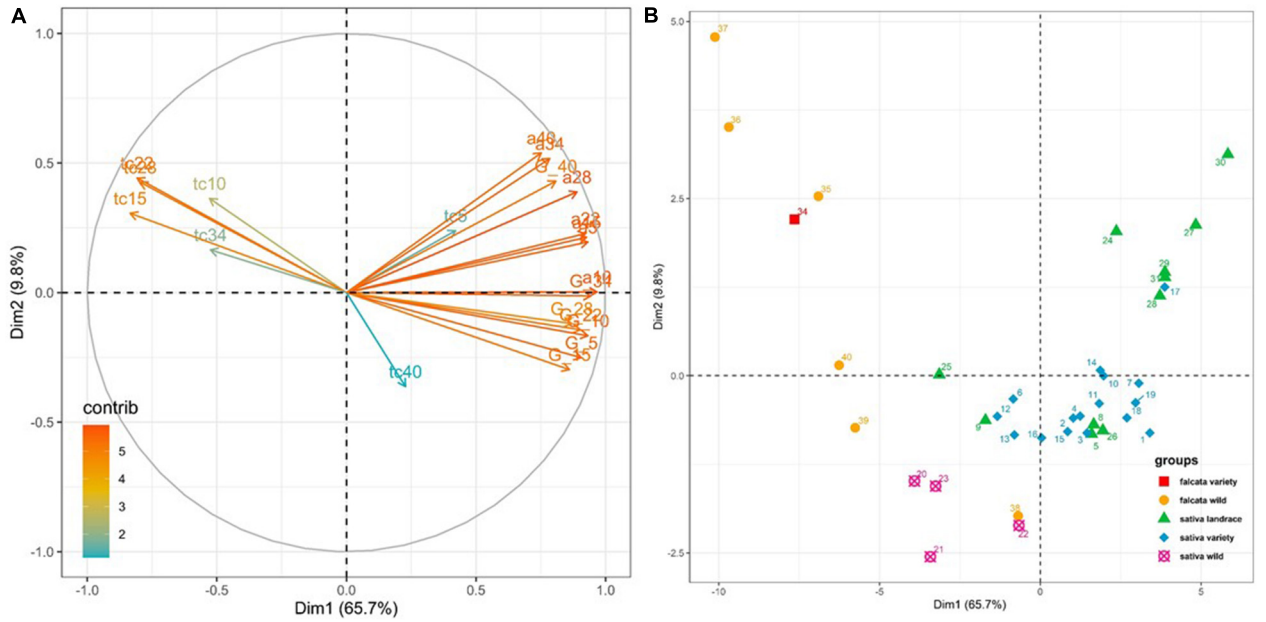


FIGURE 3 | PCA on the parameters of seed germination collected on 38 accessions at 7 temperatures. **(A)** Contribution of the variables* to the first two axes, **(B)** Repartition of the accessions on the first two axes. The codes of the varieties are given in **Table 1**. *y_{max} for germinability, t_c for the lag time to start germination, *a* for the rate of germination. y_{max}, t_c and *a* are followed by the experimental temperature.

TABLE 4 | Germinability (y_{max} in%), lag to start germination (t_c in h) and maximum germination rate (α in% of germinating seeds per h) for the 5 groups of lucerne accessions at 7 temperatures.

Temperature (°C)	Parameter	*Falcata + glomerata* wild	*Sativa* wild	*Falcata* variety	*Sativa* landrace	*Sativa* variety	*F* Value
5	α	0.074	0.109	0.007	0.801	0.701	0.9NS
	t_c	77.0	29.6	21.8	80.1	82.3	7.8***
	y_{max}	45.0	81.1	22.8	91.2	90.6	67.2***
10	α	0.346	0.832	0.062	1.547	1.440	11.2***
	t_c	33.8	29.6	72.5	28.0	28.4	7.2***
	y_{max}	65.3	83.5	74.5	95.1	95.2	43.5***
15	α	0.797	1.938	0.660	4.271	3.395	22.2***
	t_c	22.0	15.9	28.1	14.4	15.1	29.2***
	y_{max}	71.0	88.9	75.9	93.3	95.6	31.7***
22	α	0.673	1.820	0.427	6.407	5.122	4.1***
	t_c	32.8	13.6	27.8	12.3	12.4	67.3***
	y_{max}	68.6	78.9	80.7	92.6	94.7	33.4***
28	α	0.479	0.784	0.142	5.415	3.803	1.5NS
	t_c	61.9	16.4	74.7	12.6	13.0	151.1***
	y_{max}	63.3	76.2	72.1	90.3	92.8	20.9***
34	α	0.159	0.155	0.019	3.601	2.066	15.4***
	t_c	41.2	25.6	17.8	15.2	15.7	8.9***
	y_{max}	20.1	28.0	17.4	83.8	84.3	7.0***
40	α	0.024	0.013	0.007	0.131	0.091	23.5***
	t_c	49.1	36.7	26.9	31.0	41.6	39.3***
	y_{max}	3.5	5.1	3.1	18.6	15.4	42.6***

*The F value is for the test of difference among groups. ***significant at P < 0.001, NS: non-significant at P = 0.05.*

sown, even if severe stress conditions (frost, drought) could occur shortly afterwards. It is thus the responsibility of the farmer to choose an adequate sowing season to ensure good establishment and favourable growing conditions. In the wild, germination is submitted to the random occurrence of seed coat wound, thereby spreading germination over a long time period after the seeds reach the soil. As other domestication traits whose inheritance is usually simple, a small number of genes or a single gene

probably govern the susceptibility of lucerne seed germination to scarification (Smykal et al., 2014). This domestication trait comes in addition to the domestication syndrome already described in lucerne: plant vigour caused by polyploidy and erect growth habit to increase forage production and facilitate harvesting, and coiled-shape pods to limit seed shattering (Annicchiarico et al., 2015). The early farmers have probably selected plants that do not require scarification, by sampling the first emerged seedlings after sowing, and modern breeders proceed similarly. However, in our study as in others (Acharya et al., 1999; Palfi, 2007; Kandil et al., 2012), the germination of cultivated varieties was enhanced by scarification, which means that the loss of the hardseededness trait is not yet completed. Either the genetics of lucerne (autotetraploidy and allogamy) make the removal of this trait from the breeding populations difficult, or the hardseededness is under a quantitative genetic control that implies many genes (Finkelstein et al., 2008). The germination of the single *falcata* variety studied here was strongly enhanced by scarification. This variety has been bred for adaptation to Northern Canada from a wild Russian *falcata* population (Pankiw and Siemens, 1976) but germination without scarification has probably not been taken into account. In all situations, either with elite breeding pools or in breeding programs that include wild material, further selection devoted to this trait should improve germination and thus lucerne establishment.

Effect of Seed Deterioration

For both landraces, Flamande and Gabès, the low germination seed lot germinated less, with lower germination rate and longer lag to start germination than the high germination seed lot, at all temperatures. The differences between the two seed lots were higher at extreme temperatures than at average temperatures (15–22°C). This effect of temperature on the germination of deteriorated seed lots has little been studied. On pea (*Pisum sativum*) and bean (*Phaseolus vulgaris*), for example, all the seed lots of all genotypes had a high germinability, even at extreme temperatures, except for two bean seed lots at 10°C (Raveneau et al., 2011). For pea seeds produced under accelerated desiccation conditions, germination at a cold temperature (5°C) was only reduced by 10% compared to a control seed lot (Raveneau, 2012). The results obtained here suggest that the seeds with altered germinability should be sown in optimal temperature conditions, either for culture at the farm level or for multiplication in Genetic Resources Centres.

Genetic Diversity

The impact of temperature on germination is well established and cardinal temperatures have been determined on many species, such as weeds (Long et al., 2015), crops (Durr et al., 2015) and cover crops (Tribouillois et al., 2016). All accessions evaluated in our work exhibited the highest germinability and rate, as well as the shortest lag to start germination at 15°C or 22°C. The optimal constant temperature for lucerne germination is thus within this range. As discussed below, the germination was not as good at colder or warmer temperatures but the effect of extreme temperatures depended on the accessions.

The within-species diversity for germination in response to temperature has received little attention either in the number of accessions or in the ranges of temperatures that have been tested. In our study, we analysed 38 accessions from the whole *M. sativa* complex, with diploid and tetraploid accessions of wild and cultivated sub-species. This range of diversity is interesting because subspecies freely intercross at a ploidy level. Wild populations that are under-represented in breeding pools (Muller et al., 2006) could be used to integrate some traits for adaptation to new constraints, such as climate change, or new practices, such as cover crops (Labreuche, 2017; Moore et al., 2019). These accessions were studied in seven temperatures ranging from 5°C, a temperature which mimics cold autumns or springs, to a temperature of 40°C, which could occur with late spring sowings under warm climates. In our study, germinability was very low, on average 12% only, at 40°C. In contrast, germinability at 40°C ranged from 60 to 80% in a preliminary study on a subset of seven accessions among the ones studied here (Ahmed et al., 2019). We suppose that the high germinability in the latter study originates both from the temperature during counting (counting in the laboratory at 20°C) and/or from the natural light received by the seeds during counting (L. Ahmed, personal observation).

Within cultivated *sativa* accessions, germinability varied little at temperatures from 5 to 34°C. Breeding for stable germination is probably conducted, even unconsciously, when the first emerged plantlets are selected to constitute the pool of plants to be studied or crossed. In the field too, the first emerged plants gain a competitive advantage on the later emerged ones because they capture resources such as light. Indeed, there is no difference between landraces and varieties, meaning that this selection has been operating for a long time. Other studies have already established this low range of genetic diversity among varieties, but with experiments most frequently conducted under temperatures between 10 and 30°C only (Mcelgunn, 1973; Hampton et al., 1987; Brar et al., 1991; Butler et al., 2014). Anyway, the non-dormant cultivated accessions (dormancy 9, from Tunisia, Morocco and Iran) had higher germination rates at 34°C than the more dormant accessions (dormancy between 3 and 7). Non-dormant accessions are bred and used under warm climates in which hot periods may occur in the sowing time, either autumn or spring. This inability to germinate at high temperature has probably been selected as a global adaptation to local conditions.

The wild accessions, on average, had a lower germinability, a lower germination rate and a longer lag to start germination than cultivated accessions. Moreover, wild accessions adapted to cold conditions (*falcata* ssp) had a low germinability at 5°C and wild accessions from Spain (*sativa* ssp) had a low germination rate at 34°C but a high germination rate at 5°C. This counterintuitive result can easily be explained: if the falcata accessions germinated at 5°C (in the autumn), they would have a large risk of encountering cold conditions in the following weeks. It is an adaptation for populations that their seeds germinate when temperatures reach at least 10°C. Reversely, if wild accessions from Spain germinated at 34°C (in the spring), they would suffer from associated dry conditions. In the case where they germinate at 5°C, this cold

temperature is statistically not followed by colder temperatures, so establishment is preserved. These adaptations are ways to escape harsh (either cold or dry) conditions in the wild. In our study, the *falcata* variety had germination characteristics close to those of wild *falcata*. Its breeding, from a wild *falcata* accession (Pankiw and Siemens, 1976), probably did not include a selection for germination at low temperatures, as seen above for scarification. Krasnokutskaya, from Russia, received as a wild *falcata* accession by the Genetic Resources Centre of INRAE, Lusignan, France, showed germination characteristics close to those of cultivated *sativa* accessions, but its lag to start germination at 34°C was longer. This germination ability suggests that Krasnokutskaya is probably a cultivated accession selected for improved establishment. Another study (Julier et al., 1995) revealed that this accession exhibited a relative erect growth habit, long stems and large seeds, which are relevant breeding traits.

This study has been conducted with seed lots that have been stored for different lengths of time in cold chambers. With this allogamous insect-pollinated species, the seed production of many accessions in the same location and year is challenging. The storage duration may impact seed germination response to temperature but we suppose that this impact is smaller than the genetic difference among accessions. Indeed, these results show that natural selection for germination in response to temperature has shaped species diversity. The mean value of this trait, which can be considered as an adaptative trait, has changed during the evolution of the *Medicago sativa* complex in several subspecies adapted to different regions and climates. In addition, human selection has been efficient to alter the effect of temperature on germination, as also demonstrated in experimental work (Klos and Brummer, 2000). In breeding programmes, a good germination in a large range of temperatures is required. When the plant material originates from cultivated pools, no more selection is needed. When wild accessions are introduced in a breeding pool, attention must be paid to reach a high germination, especially at extreme temperatures.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://doi.org/10.15454/YIECZS.

AUTHOR CONTRIBUTIONS

AE-G and BJ conceived the study. AE-G, LA, and BJ designed the experiment. LA collected the data with contributions of BJ, AE-G, and WG. BJ, WG, and LA carried out the data analysis. All co-authors contributed to wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement no. 727312 (project EUCLEG).

ACKNOWLEDGMENTS

We thank the seed companies and institutes that provided us with seeds of their accessions: AgriObtentions, Alforex, ART, Barenbrug, CGS Sementi, CREA, DLF Seeds, Forage Genetics, Foratec Argentina, IFVCNS, Lantmännen, RAGT-2n. Cédric Perrot is thanked for the maintenance of growth chambers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.578121/full#supplementary-material

REFERENCES

Acharya, S.N., Stout, D.G., Brooke, B., and Thompson, D. (1999). Cultivar and storage effects on germination and hard seed content of alfalfa. *Can. J. Plant Sci.* 79, 201–208. doi: 10.4141/p98-043

Acosta, Y., Santiago, F., Escalante, D., Mazorra, C., Cejas, I., Martinez-Montero, M.E. et al. (2020). Cryo-exposure of Neonotonia wightii Wigth and Am seeds enhances field performance of plants. *Acta Physiologiae Plantarum* 42, 1–6.

Ahmed, L.Q. (2015). *Analyse de la variabilité inter- et intra-spécifique de cinq espèces prairiales en réponse à la température pendant la germination et la phase hétérotrophe initiale*. Poitiers: Doctorat d'Université.

Ahmed, L.Q., Durand, J.L., and Escobar-Gutierrez, A.J. (2019). Genetic diversity of alfalfa (*Medicago sativa*) in response to temperature during germination. *Seed Sci. Technol.* 47, 351–356.

Annicchiarico, P., Barrett, B., Brummer, E.C., Julier, B., and Marshall, A.H. (2015). Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* 34, 327–380. doi: 10.1080/07352689.2014.898462

Basigalup, D., Irwin, J., Mi, F., and Abdelguerfi-Laouar, M. (2014). Perspectives of alfalfa in australia, china, africa and latin america. *Legume Perspec.* 4, 9–10.

Baskin, C.C., and Baskin, J.M. (2014). *Seeds: ecology, biogeography, and evolution of dormancy and germination.* San Diego. CA, USA: Academic Press.

Baskin, J.M., and Baskin, C.C. (2004). A classification system for seed dormancy. *Seed Sci. Res.* 14, 1–16. doi: 10.1079/ssr2003150

Bewley, J.D., and Black, M. (1994). *Seeds physiology of development and germination*. London: Plenum Press.

Brar, G.S., Gomez, J.F., Mcmichael, B.L., Matches, A.G., and Taylor, H.M. (1991). Germination of 20 forage legumes as influenced by temperature. *Agron. J.* 83, 173–175. doi: 10.2134/agronj1991.00021962008300010040x

Brunel, S., Teulat-Merah, B., Wagner, M.H., Huguet, T., Prosperi, J.M., Dürr, C. et al. (2009). Using a model-based framework for analysing genetic diversity during germination and heterotrophic growth of *Medicago truncatula*. *Ann. Bot.* 103, 1103–1117. doi: 10.1093/aob/mcp040

Butler, T.J., Celen, A.E., Webb, S.L., Krstic, D., and Interrante, S.M. (2014). Temperature affects the germination of forage legume seeds. *Crop Sci.* 54, 2846–2853. doi: 10.2135/cropsci2014.01.0063

Corbineau, F. (2012). Markers of seed quality: from present to future. *Seed Sci. Res.* 22, S61–S68.

De Serres, O. (1600). *Le théâtre d'agriculture et mesnage des champs (new edition, 2001)*. France: Actes Sud.

Dias, P.M.B., Brunel-Muguet, S., Dürr, C., Huguet, T., Demilly, D., Wagner, M.H. et al. (2011). QTL analysis of seed germination and pre-emergence growth at extreme temperatures in *Medicago truncatula*. *Theor. Appl. Genet.* 122, 429–444. doi: 10.1007/s00122-010-1458-7

Durr, C., Dickie, J.B., Yang, X.Y., and Pritchard, H.W. (2015). Ranges of critical temperature and water potential values for the germination of species worldwide: contribution to a seed trait database. *Agric.Forest Meteorol.* 200, 222–232. doi: 10.1016/j.agrformet.2014.09.024

Finch-Savage, W.E., and Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. *New Phytol.* 171, 501–523.

Finkelstein, R., Reeves, W., Ariizumi, T., and Steber, C. (2008). Molecular aspects of seed dormancy. *Annu. Rev. Plant Biol.* 59, 387–415.

Fu, Y.B., Ahmed, Z., and Diederichsen, A. (2015). Towards a better monitoring of seed ageing under ex situ seed conservation. *Conserv. Physiol.* 3:cov026. doi: 10.1093/conphys/cov026

Hampton, J.G., Charlton, J.F.L., Bell, D.D., and Scott, D.J. (1987). Temperature effects on the germination of herbage legumes in New Zealand. *Proc. New Zealand Grassland Assoc.* 48, 177–183. doi: 10.33584/jnzg.1987.48.1772

Julier, B., Gastal, F., Louarn, G., Badenhausser, I., Annicchiarico, P., Crocq, G. et al. (2017). "Alfalfa (lucerne) in European cropping systems,". in *Legume Futures*. (Wallingford: CABI), 168–192. doi: 10.1079/9781780644981.0168

Julier, B., Porcheron, A., Ecalle, C., and Guy, P. (1995). Genetic variability for morphology, growth and forage yield among perennial diploid and tetraploid lucerne populations (*Medicago sativa* L). *Agronomie* 15, 295–304. doi: 10.1051/agro:19950505

Kandil, A.A., Sharief, A.E., and Odam, A.M.A. (2012). Dormancy overcoming of some alfalfa varieties. *Res. J. Seed Sci.* 5, 19–31. doi: 10.3923/rjss.2012.19.31

Klos, K.L., and Brummer, E.C. (2000). Response of six alfalfa populations to selection under laboratory conditions for germination and seedling vigor at low temperature. *Crop Sci.* 40, 959–964. doi: 10.2135/cropsci2000.404959x

Labreuche, J. (2017). *"Impacts de couverts permanents sur le blé tendre d'hiver", in: 13èmes Rencontres de la fertilisation raisonnée et de l'analyse COMIFER-GEMAS*. (Nantes: COMIFER).

Lamichhane, J.R., Debaeke, P., Steinberg, C., You, M.P., Barbetti, M.J., Aubertot, J.N. et al. (2018). Abiotic and biotic factors affecting crop seed germination and seedling emergence: a conceptual framework. *Plant Soil.* 432, 1–28. doi: 10.1007/s11104-018-3780-9

Long, R.L., Gorecki, M.J., Renton, M., Scott, J.K., Colville, L., Goggin, D.E. et al. (2015). The ecophysiology of seed persistence: a mechanistic view of the journey to germination or demise. *Biol. Rev.* 90, 31–59. doi: 10.1111/brv.12095

Luo, Q.Y. (2011). Temperature thresholds and crop production: a review. *Clim. Change* 109, 583–598. doi: 10.1007/s10584-011-0028-6

Mcdonald, M.B. (1999). Seed deterioration: physiology, repair and assessment. *Seed Sci. Technol.* 27, 177–237.

Mcelgunn, J.D. (1973). Germination response of forage legumes to constant and alternating temperatures. *Can. J. Plant Sci.* 53, 797–800. doi: 10.4141/cjps73-154

Mcgraw, R.L., Shockley, F.W., and Elam, T.K. (2003). Effects of temperature on germination of 10 native legume species. *Native Plants Spring.* 4, 5–9. doi: 10.3368/npj.4.1.5

Monks, D.P., Sadatasilan, K., and Moot, D.J. (2009). Cardinal temperatures and thermal time requirements for germination of annual and perennial temperate pasture species. *Agron. N. Zealand* 39, 95–108.

Moore, K.J., Anex, R.P., Elobeid, A.E., Fei, S.Z., Flora, C.B., Goggi, A.S. et al. (2019). Regenerating agricultural landscapes with perennial groundcover for intensive crop production. *Agronomy* 9:458. doi: 10.3390/agronomy9080458

Moot, D.J., Scott, W.R., Roy, A.M., and Nicholls, A.C. (2000). Base temperature and thermal time requirements for germination and emergence of temperate pasture species. *N. Z. J. Agric. Res.* 43, 15–25. doi: 10.1080/00288233.2000.9513404

Muller, M., Poncet, C., Prosperi, J., Santoni, S., and Ronfort, J. (2006). Domestication history in the *Medicago sativa* species complex: inferences from nuclear sequence polymorphism. *Mol. Ecol.* 15, 1589–1602. doi: 10.1111/j.1365-294x.2006.02851.x

Narem, D. (2009). Effects of stratification, warm treatment, and mechanical and acid scarification on the emergence of yellow-flowered alfalfa (*Medicago sativa* subsp. *falcata*) seeds. *J. Undergraduate Res.* 7, 99–105.

Palfi, M. (2007). Alfalfa (*Medicago sativa* L.) seed scarification by acid [HR]. *Sjemenarstvo* 24, 5–16.

Pankiw, P., and Siemens, B. (1976). Anik alfalfa. *Can. J. Plant Sci.* 56, 203–205. doi: 10.4141/cjps76-031

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., Authors, E., Heisterkamp, S. et al. (2020). "nlme: Linear and nonlinear mixed effects model". R package version 3, 1–148.

Pinheiro, J.C., and Bates, D.C. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Prosperi, J.M., Muller, M.H., Ronfort, J., Fourtier, S., Sampoux, J.P., Jenczewski, E. et al. (2014). Alfalfa domestication history, genetic diversity and genetic resources. *Legume Perspec.* 4, 13–14.

R Core Team (2019). *"A language and environment for statistical computing"*. (Vienna, Austria: R Foundation for Statistical Computing).

Raveneau, M.P. (2012). *Effet des vitesses de dessiccation de la graine et des basses températures sur la germination du pois protéagineux.* PhD. France: Angers University.

Raveneau, M.P., Coste, F., Moreau-Valancogne, P., Lejeune-Hénaut, I., and Durr, C. (2011). Pea and bean germination and seedling responses to temperature and water potential. *Seed Sci. Res.* 21, 205–213. doi: 10.1017/s0960258511000067

Shafii, B., and Price, W.J. (2001). Estimation of cardinal temperatures in germination data analysis. *J. Agric. Biol. Environ. Stat.* 6, 356–366. doi: 10.1198/108571101317096569

Small, E., and Brookes, B. (1990). A numerical taxonomic analysis of the *Medicago littoralis - M. truncatula* complex. *Can. J. Bot.* 68, 1667–1674. doi: 10.1139/b90-214

Small, E., and Jomphe, M. (1989). A synopsis of the genus *Medicago* (*Leguminosae*). *Can. J. Bot.* 67, 3260–3294. doi: 10.1139/b89-405

Smykal, P., Vernoud, V., Blair, M.W., Soukup, A., and Thompson, R.D. (2014). The role of the testa during development and in establishment of dormancy of the legume seed. *Front. Plant Sci.* 5:351. doi: 10.3389/fpls.2014.00351

Taylor, G.B. (2005). Hardseededness in mediterranean annual pasture legumes in australia: a review. *Aust. J. Agric. Res.* 56, 645–661. doi: 10.1071/ar04284

Thornley, J.H.M., Sutcliffe, J.F., and Mahlberg, P. (1976). *Mathematical models in plant physiology*. London: Academic Press.

Tribouillois, H., Dürr, C., Demilly, D., Wagner, M.H., and Justes, E. (2016). Determination of germination response to temperature and water potential for a wide range of cover crop species and related functional groups. *Plos One* 11:e0161185. doi: 10.1371/journal.pone.0161185

Zhou, W.G., Chen, F., Luo, X.F., Dai, Y.J., Yang, Y.Z., Zheng, C. et al. (2020). A matter of life and death: molecular, physiological, and environmental regulation of seed longevity. *Plant Cell Environ.* 43, 293–302. doi: 10.1111/pce.13666

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 © 2021 Ghaleb, Ahmed, Escobar-Gutiérrez and Julier. 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 Plant Science

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fpls.2021.631767

A Genome-Wide Genetic Diversity Scan Reveals Multiple Signatures of Selection in a European Soybean Collection Compared to Chinese Collections of Wild and Cultivated Soybean Accessions

*Aamir Saleem[1,2], Hilde Muylle[1], Jonas Aper[1], Tom Ruttink[1], Jiao Wang[3], Deyue Yu[3] and Isabel Roldán-Ruiz[1,2]**

[1] Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium, [2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [3] National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China

OPEN ACCESS

Edited by:
Sergio J. Ochatt,
INRA UMR 1347 Agroécologie, France

Reviewed by:
Umesh K. Reddy,
West Virginia State University, United States
Marinus J. M. Smulders,
Wageningen University and Research, Netherlands

***Correspondence:**
Isabel Roldán-Ruiz
isabel.roldan-ruiz@ilvo.vlaanderen.be

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 20 November 2020
Accepted: 01 February 2021
Published: 26 February 2021

Citation:
Saleem A, Muylle H, Aper J, Ruttink T, Wang J, Yu D and Roldán-Ruiz I (2021) A Genome-Wide Genetic Diversity Scan Reveals Multiple Signatures of Selection in a European Soybean Collection Compared to Chinese Collections of Wild and Cultivated Soybean Accessions. Front. Plant Sci. 12:631767. doi: 10.3389/fpls.2021.631767

Targeted and untargeted selections including domestication and breeding efforts can reduce genetic diversity in breeding germplasm and create selective sweeps in crop genomes. The genomic regions at which selective sweeps are detected can reveal important information about signatures of selection. We have analyzed the genetic diversity within a soybean germplasm collection relevant for breeding in Europe (the EUCLEG collection), and have identified selective sweeps through a genome-wide scan comparing that collection to Chinese soybean collections. This work involved genotyping of 480 EUCLEG soybean accessions, including 210 improved varieties, 216 breeding lines and 54 landraces using the 355K SoySNP microarray. SNP calling of 477 EUCLEG accessions together with 328 Chinese soybean accessions identified 224,993 high-quality SNP markers. Population structure analysis revealed a clear differentiation between the EUCLEG collection and the Chinese materials. Further, the EUCLEG collection was sub-structured into five subgroups that were differentiated by geographical origin. No clear association between subgroups and maturity group was detected. The genetic diversity was lower in the EUCLEG collection compared to the Chinese collections. Selective sweep analysis revealed 23 selective sweep regions distributed over 12 chromosomes. Co-localization of these selective sweep regions with previously reported QTLs and genes revealed that various signatures of selection in the EUCLEG collection may be related to domestication and improvement traits including seed protein and oil content, phenology, nitrogen fixation, yield components, diseases resistance and quality. No signatures of selection related to stem determinacy were detected. In addition, absence of signatures of selection for a substantial number of QTLs related to yield, protein content, oil content and phenological traits suggests the

presence of substantial genetic diversity in the EUCLEG collection. Taken together, the results obtained demonstrate that the available genetic diversity in the EUCLEG collection can be further exploited for research and breeding purposes. However, incorporation of exotic material can be considered to broaden its genetic base.

Keywords: selective sweeps, breeding, European soybean, genetic diversity, Chinese soybean, *Glycine max*, SNP markers, signatures of selection

INTRODUCTION

Crop improvement relies on genetic diversity of plant genetic resources. A high genetic diversity provides an opportunity for plant breeders to develop cultivars with desirable characteristics (Savchenko, 2017; Byrne et al., 2018). Today's improved cultivars of various crops, with specific characteristics depending on their use and environmental adaptation, are the result of historical domestication events and intentional as well as unintentional selections by farmers and breeders (Bradshaw, 2017; Stoskopf et al., 2019). The processes of domestication and selection lead to an increased frequency of favorable alleles, and in the most extreme situation may cause complete fixation at genomic loci underlying beneficial traits in the genepool of the crop (Smýkal et al., 2018; Weigand and Leese, 2018). At chromosomal scale, the locus that underlies a beneficial trait is surrounded by other linked loci carrying neutral mutations. The selection process targeting the advantageous allele also results in an increase of the frequency of alleles at those neutral loci that are in linkage disequilibrium with the advantageous allele. This causes a so-called "selective sweep" in the genome. Genomic regions that have undergone a selective sweep are characterized by high levels of homozygosity, an increase in low- and high-frequency alleles, a high linkage disequilibrium and a low genetic diversity (Nielsen, 2005; Hufford et al., 2012).

Selective sweeps have been analyzed in several crops to reconstruct their history of domestication and selection and to identify the genetic loci and their underlying genome sequence that were mainly affected by these processes. For example, Hufford et al. (2012) identified 3,040 genes through selective sweep analysis, revealing signatures of domestication and improvement of maize in the United States. The dispersion history and adaptive evolution of wheat throughout the agro-ecological zones of China have been inferred by population genetics analysis including selective sweep analysis (Zhou et al., 2018). Ndjiondjop et al. (2019) identified 37 candidate selective sweep regions harboring genes related to biotic and abiotic stress tolerance in African rice, and demonstrated that those regions displayed low genetic diversity as a result of strong positive selection and domestication in African rice compared to Asian rice and its wild progenitor (Oryza barthii A. Chev.). A selective sweep analysis in barley by Pankin et al. (2018) provided molecular evidence of multiple domestication origins and allowed to distinguish domestication-related traits (i.e., non-brittle rachis) from improvement-related traits (i.e., naked grain).

Modern cultivated soybean was domesticated approximately 5000 years ago from its wild progenitor Glycine soja, which is distributed throughout Eastern Asia, including most of China, South Korea, and Japan (Jeong et al., 2019). Soybean is the world's most grown high-value legume crop with beans containing high percentages of protein and oil (Pagano and Miransari, 2016). Being a restorative crop, soybean fixes atmospheric nitrogen in symbiosis with Rhizobium bacteria and delivers environmental services by minimizing the need for mineral nitrogen fertilizer. Between 2008 and 2018, global soybean production has grown from 212 to 337 million tons per year, while the total cultivated surface increased from 97 to 124 million hectares (IDH and IUCN NL, 2019). Soybean was first introduced to Europe during the second half of the 19th century. The current soybean acreage of 5.5 million hectares in Europe, representing a mere 3.4% of the world soybean production (FAOSTAT., 2019), can meet only 34% of the current European need for soybean (IDH and IUCN NL, 2019)). To meet the increasing European demand and to reduce the dependency on import, it is crucial to expand soybean cultivation and to adapt soybean genotypes to new cultivation zones in Europe. This requires a good understanding of the origin and genetic architecture of European soybean germplasm and how it relates to the germplasm from other origins.

Based on the responsiveness of soybean flowering and maturity to photoperiod and temperature, a total of 13 distinct maturity groups have been defined, of which only early maturing types (maturity groups MG000 to MGII) are suitable for cultivation in Europe (Kurasch et al., 2017). Previous studies have shown a narrow genetic base of the European soybean germplasm (Hahn and Würschum, 2014; Žulj Mihaljević et al., 2020), which can be due to the use of only a few ancestors originating from Canada, North America, Japan and China for breeding in Europe (Ristova et al., 2010; Hahn and Würschum, 2014; Miladinović et al., 2018). In addition, the original material used for breeding probably carried a low level of genetic diversity, as both pedigree and molecular marker data have indicated a narrow genetic base of North American and Canadian germplasm (Gizlice et al., 1996; Vaughn and Li, 2016; Bruce et al., 2019). In contrast, the Chinese soybean breeding pool contains a high level of genetic diversity because of a long history of cultivation over diverse eco-geographical zones with varying ranges of temperature and photoperiod (Liu et al., 2017). Selective sweep analysis has also been applied in soybean to understand the domestication and selection history. For example, Wen et al. (2015); Zhou et al. (2015) and Wang et al. (2016) report candidate selection regions harboring genes potentially involved in traits of agronomic relevance such as grain yield, seed size, flowering date, maturity date, seed protein and oil content and traits related to stress tolerance. In addition, selective sweep analysis

by Jeong et al. (2019) reported domestication-related signals in soybean using mainly germplasm from Japan and Korea. Thus, available studies have considered materials of Chinese, Japanese and Korean origin. However, similar studies have not been performed in the soybean genepool available in Europe.

Our current knowledge on the origin and genetic relationships within European soybean germplasm is still fragmented. Main reasons are the limited number of accessions (covering only a fraction of the total genetic diversity) that were included in previous studies (e.g., 28 accessions in Ristova et al. (2010), 93 in Hahn and Würschum (2014), 75 in Kurasch et al. (2017) and 97 in Žulj Mihaljević et al. (2020)), and/or the low number of genetic markers that were used for screening. Miladinović et al. (2018), genotyped 445 accessions at 85,000 SNP loci, but only used the 38 SNPs located in maturity genes for analysis, and focused on materials from one European breeding program. To fill this gap in our knowledge about the genetics of soybean germplasm relevant for breeding in Europe and to develop breeding tools for legume crops including soybean, a consortium was established within the European Union project EUCLEG[1]. In this context a unique collection of 480 soybean accessions considered relevant for European breeders, originating from 25 countries and covering a broad range of genetic diversity was assembled (named the EUCLEG collection in what follows). This offers a unique opportunity to compare the genetic diversity of the EUCLEG collection to that contained in reference materials from China, helping us to understand the main forces that have shaped the soybean genepool currently being used in breeding programs outside China.

Here, we present an analysis of the genetic diversity within the EUCLEG collection and identify selective sweeps through a genome-wide diversity scan between the EUCLEG and a Chinese soybean collection (Wang et al., 2016) ("NJAU collection"). Specific objectives of this study were: (i) to explore the structure and genetic relatedness of accessions in the EUCLEG collection; (ii) to determine the level of genetic diversity in the EUCLEG collection compared to that of the NJAU collection; (ii) to identify genomic regions that putatively underwent selective sweeps in the EUCLEG collection and their significance for future soybean improvement efforts in Europe.

MATERIALS AND METHODS

EUCLEG Collection

The EUCLEG collection consists of 480 accessions belonging to maturity groups (MG) 000, 00, 0 and I/II, and includes 210 improved varieties, 216 breeding lines and 54 landraces. A detailed description of this collection is provided in **Supplementary Table 1**. This collection was made in collaboration with seed companies and gene banks by considering the accessions with diverse geographical origins, with varying levels of selection (landraces, varieties and breeding lines), and their relevance for European soybean breeders.

[1]http://www.eucleg.eu/

For DNA extraction, one fully developed unifoliate leaf was sampled per accession from plants grown in the field. Leaf samples were lyophilized and stored under vacuum conditions until use. Genomic DNA was extracted using the DNeasy® Plant Mini Kit[2]. Pure and good quality DNA samples with an average concentration of 106 ng/μl (range 24–731 ng/μl) were used for genotyping using the 355K SoySNP Axiom microarray (Wang et al., 2016) from Affymetrix (Thermo Fisher Scientific), via Eurofins, DK.

NJAU Collection

The NJAU collection originates from the Germplasm Storage of Chinese National Center for Soybean Improvement, Nanjing Agricultural University, China, and comprises 122 wild and 272 cultivated accessions. It covers the three ecological habitats of soybean in China including the regions of Northern China, Huang-Huai and Southern China. A full description of the NJAU collection is available in Wang et al. (2016). This collection has previously been genotyped using the 355K SoySNP microarray (Wang et al., 2016). Here, we have combined the raw microarray fluorescence data of the EUCLEG samples with the raw fluorescence data of the NJAU samples to perform a "joint" SNP calling.

SNP Calling

SNP calling was performed using the software Axiom Analysis Suite (AAS) from Affymetrix®, following the instructions provided in the Axiom Analysis Suite 3.1 user guide[3]. Before SNP calling of the 874 samples of the combined data set (EUCLEG and NJAU), we first checked the performance of the 355K SoySNP microarray on the 480 EUCLEG samples separately. This step was considered necessary, as the 355K SoySNP microarray was developed using the NJAU collection and might perform sub-optimally with plant materials of a different origin. In brief, the Affymetrix® Power Tools (APT) software package, version 1.15.0 implemented in AAS performed sample quality control based on 20,000 non-polymorphic probe sets and considering the parameters Dish Quality Control (DQC; determines the intensity of contrast between signal and noise) and Sample Call Rate (QC-CR; refers to the ratio of genotype-called SNPs to attempted SNPs in a sample). Based on criteria DQC $>$ 0.82 and QC-CR $\geq$ 97, AAS filtered out four poor-quality samples. The R package SNPolisher version 1.3.6.7 implemented in AAS was used for SNP calling using 609,883 probe sets targeting 355,595 SNPs. Its Ps_Classification function classified the SNPs/probe sets into six categories based on the following SNP QC metrics: call rate (CR) $\geq$97%, Fisher's linear discriminant (FLD) $\geq$ 3.6, heterozygous strength offset (HetSO) ≥ -0.1, and homozygote ratio offset (HomRO) $\geq$0.3 for one-cluster or two-cluster SNPs or ≥ -0.9 for three-cluster SNPs. A summary of the SNP classification was obtained for the 476 good quality samples of the EUCLEG collection. We compared this summary with the SNP classification summary obtained from the NJAU collection by Wang et al. (2016).

[2]https://www.qiagen.com/be/resources/

[3]https://www.thermofisher.com/

In a second step, we genotyped the combined dataset (EUCLEG and NJAU), starting from the raw fluorescence data following the procedure described above. In the quality control step, 69 poor quality samples were excluded. SNP calling was performed on the remaining 805 good quality samples. After genotyping, low quality SNPs based on SNP QC metrics were excluded and a final genotyping dataset containing 229,557 SNPs was generated. This dataset was divided in three subsets for further processing: EUCLEG, NJAU-Wild and NJAU-Cultivated, comprising 477, 82, and 246 accessions, respectively. For the divisions NJAU-Wild and NJAU-Cultivated, we refer to Wang et al. (2016). In further analyses, we considered either the whole collection (EUCLEG and NJAU) or some of these subsets.

For some of the downstream analyses, the genomic coordinates of the SNPs were required. Because during the development of the 355K SoySNP microarray SNP coordinates were assigned using an older version of the soybean reference genome sequence (Glyma.Wm82.a1), we positioned SNPs onto the novel reference genome sequence Glyma.Wm82.a2 (with improved assembly and gene annotation quality compared to Glyma.Wm82.a1). Finally, the 224,993 SNPs corresponding to probes that could be positioned onto the 20 soybean chromosomes using a blast query were considered for further analyses.

Population Structure Analysis

The results of population structure of the NJAU collection are available in Wang et al. (2016). Here, we performed a population structure analysis of EUCLEG and NJAU combined (805 accessions in total). Two approaches were applied. In the first approach, a model-based structure analysis was performed in fastSTRUCTURE 1.0 (Raj et al., 2014) including 179,812 SNPs with minor allele frequency (MAF) of at least 5% across the sample set. The K value was varied from 2 to 10, while for other parameters default settings were used. The optimum value of K was determined using the best marginal likelihood value of fastSTRUCTURE-output from K = 2 to K = 10. The results of fastSTRUCTURE were graphically visualized using the R-package pophelper v. 2.1.0 (Francis, 2017). In the second approach, a principal component analysis (PCA) was performed in TASSEL 5 (Bradbury et al., 2007). The missing values of SNPs were imputed using the unweighted average method and PCA was performed on the genetic correlation matrix of accessions considering the first five principal components.

Population structure was also inferred for the 477 genotyped accessions of the EUCLEG collection separately, using the settings and methods described in the previous paragraph. The analysis included 139,986 SNPs with MAF of at least 5% across the sample set. The number of subgroups (K) was determined considering the delta log-likelihood criterion. For the interpretation of the results of fastSTRUCTURE, once an optimal value of K was identified, each accession was assigned to a subgroup "n" for which the ancestral coefficient reached a value $Qn \geq 0.4$; where n is the number of subgroup (1 to K). Accessions for which the two highest Q values differed by less than 0.2 were considered "Admixed". Finally, the degree of divergence between the EUCLEG and the NJAU collections was estimated by calculating a fixation index (F_{ST}) value per SNP site in VCFtools v. 0.1.15 (Danecek et al., 2011) following the methods of Weir and Cockerham (1984).

Hierarchical cluster analysis of EUCLEG and NJAU combined was performed to determine the relationship among accessions of both collections, and to see the relationship between the EUCLEG part from the Chinese origin with other accessions. For this, a dendrogram was constructed following Ward's D2 method (Murtagh and Legendre, 2014) and using Nei's standard genetic distances between accessions (Nei, 1972).

Genetic Diversity Estimates

Linkage disequilibrium (LD) analysis was performed in VCFtools v. 0.1.15 considering the filtered genotyping data including 139,986 SNPs, 162,098 SNPs and 185,194 SNPs with MAF of at least 5% in the sample sets EUCLEG, NJAU-Cultivated, and NJAU-Wild, respectively. LD was estimated for each chromosome by computing the r^2 for all pairwise comparisons of two SNPs located at a maximum of 1000 kbp inter-SNP distance. The LD decay distance per chromosome was estimated as the point at which r^2 dropped to half of its maximum value. The genome-wide LD decay was estimated by pooling the LD output across all chromosomes.

Diversity estimates were determined for EUCLEG, NJAU-Cultivated, and NJAU-Wild separately, including all 224,993 SNPs. The EUCLEG collection contained 21 accessions from Chinese origin, but to avoid any possible confounding effect of those accessions while comparing genetic diversity between EUCLEG and NJAU, these 21 accessions were not considered. The average pairwise divergence among genotypes within each collection was then determined by computing the nucleotide diversity index (π) per SNP site in VCFtools v. 0.1.15.

Selective Sweep Analysis

To detect signals of selection in the EUCLEG collection, the cross-population composite likelihood ratio test (XP-CLR) implemented in XP-CLR v. 1.0 (Chen et al., 2010) was used. XP-CLR is a site frequency spectrum (SFS)-based method that detects selective sweeps by jointly modeling the multi-locus allele frequency differentiation between two populations. Given the allele frequency of a locus in the reference population, XP-CLR maximizes the likelihood ratio of the allele frequency in an object population between a selective sweep model and a null model (Chen et al., 2010). We compared the object EUCLEG collection to the reference NJAU-Wild collection. The 21 accessions of Chinese origin in the EUCLEG collection were not considered in this analysis. XP-CLR requires as input the genetic position (expressed in recombination units) of each SNP. Because for most SNPs the genetic position was unknown, we transformed physical positions (Mbp) to recombination positions (cM) considering a homogeneous recombination rate (1 Mbp = 1 cM) throughout the soybean genome.

The whole genome was scanned in XP-CLR choosing a sliding window of 1 Mbp at steps of 5 kbp. XP-CLR options were as follows: XPCLR -xpclr genofile1.txt genofile2.txt mapfile outputfile -w1 snpWin 0.01 gridSize 5000 chrN -p0 corrLevel 0.95; where genofile1.txt and genofile2.txt correspond

to the object (EUCLEG) and reference (NJAU-Wild) collections, respectively. Because two SNP loci with high pairwise r^2 values can provide redundant information, corrLevel was set to 0.95 to weight the XP-CLR value of a window containing highly correlated SNPs. Windows with weighted XP-CLR scores in the top 1% of the empirical distribution of the genome-wide XP-CLR values were used to delineate regions of interest. To define the regions of interest (hereafter called selective sweep regions), we combined neighboring windows when the gap was less than 1 Mbp.

SoyBase[4] (Grant et al., 2010) was used to search for reported QTLs located in the selective sweep regions and to generate a list of genes positioned in respective selective sweep regions. The transcripts description was obtained from "Gmax_275_Wm82.a2.v1.annotation_info.txt"[5] and the annotation of biological functions was obtained from UniProtKB[6]. The former contains transcript definition of the best hit obtained through homology-dependent sequence analysis of soybean transcripts in the Arabidopsis genome, whereas the latter corresponds to their molecular and biological functions manually annotated and reviewed from literature and computational analysis by the UniProtKB.

RESULTS

Evaluation of the Use of the 355K SoySNP Array in the EUCLEG Collection

Analysis of the EUCLEG collection with the 355K SoySNP array revealed a total of 285,953 SNP markers (80% of the total 355,595 SNPs present on the array) belonging to the recommended categories including PolyHighResolution (PHR, total 211,593), MonoHighResolution (MHR, total 46,953) and NoMinorHom (NMH, total 27,407). These categories refer to SNPs exhibiting all three genotypic classes with a good cluster resolution (PHR), SNPs with good cluster resolution but displaying only one of the homozygous clusters (MHR) and SNPs with good cluster resolution but for which one of the two homozygous clusters is missing (NMH). These proportions correspond quite well with those previously reported for the NJAU collection by Wang et al. (2016) (**Supplementary Figure 1**), indicating that the 355K SoySNP array is not only useful for the genotyping of Chinese soybean germplasm, but also for germplasm from other origins.

Genotyping of the EUCLEG and NJAU combined collection rendered 229,557 SNPs (65% of the total 355,595) belonging to the recommended categories (PHR, MHR and NMH; a total of 194,171, 16,868 and 18,518 respectively). For the remaining 126,038 SNPs, at least one of the QC metrics were below the threshold and hence they were assigned to non-recommended categories. There were 16% (of the total 355,595) more SNPs of non-recommended categories in the combined analysis as compared to the separate analysis of the EUCLEG collection. This was because a number of SNPs (19% of the total 355,595) of recommended categories (PHR, MHR and NMH) in the EUCLEG separate analysis were assigned to the non-recommended categories in the combined analysis. Taken together, these results indicate that SNP calling on EUCLEG and NJAU combined is essential to get a more precise classification of SNP markers as compared to a separate analysis for each collection. In addition, these results indicate that the SNP dataset of EUCLEG and NJAU combined contains a high number of SNPs from the recommended categories.

For 224,993 SNPs of the total 229,557 (98%), new coordinates could be positioned onto 20 chromosomes in the Glyma.Wm82.a2 soybean reference genome sequence using a BLAST query. The probes targeting the remaining SNPs (4,564) were either missing in the novel soybean genome assembly or were assigned to the non-anchored scaffolds and not to chromosomes, and were excluded from subsequent analyses. The genome-wide distribution of the final set of 224,993 SNPs used for downstream data analyses is given in **Table 1**. The longest chromosome (18) contained the highest number of SNPs (6.3% of the total 224,993), and the shortest chromosome (11) contained the lowest number of SNPs (3.8% of the total 224,993). The average SNP density was the lowest on chromosome 1 and the highest on chromosome 13 (19 and 27 SNPs per 100 kbp, respectively). The average distance between two adjacent SNPs was 2.6 kbp (**Table 1** and **Supplementary Figure 2**).

Population Structure Analysis

We analyzed the presence of population structure using two approaches, fastSTRUCTURE and PCA. In the EUCLEG and NJAU combined analysis comprising 805 accessions, the marginal likelihood of the fastSTRUCTURE-output from K = 2 to K = 10 indicated the optimum K between 2 and 4 (**Supplementary Figure 3**). The EUCLEG sample set clustered separately from NJAU sample sets (**Figure 1A** and **Supplementary Figure 4**). The division between NJAU-Cultivated and NJAU-Wild (as defined by Wang et al., 2016) was also apparent. The F_{ST} values were 0.14, 0.34 and 0.22 for EUCLEG vs. NJAU-Cultivated, EUCLEG vs. NJAU-Wild and NJAU-Cultivated vs. NJAU-Wild comparisons, respectively. This indicates that the set of 805 soybean accessions considered in this study consists of three major groups including EUCLEG, NJAU-Cultivated and NJAU-Wild. The relatively low F_{ST} between EUCLEG and NJAU-Cultivated in relation to comparisons with NJAU-Wild supports a strong differentiation between cultivated and wild soybean accessions.

In the separate analysis of the EUCLEG collection the marginal likelihood of the fastSTRUCTURE-output from K = 2 to K = 10 indicated the optimum at K = 5 (**Supplementary Figure 5**). Although some small subgroups were formed at K > 5, we worked further with five subgroups as this clustering was in concordance with the background information of the accessions. Population structure in the EUCLEG collection is presented in **Figures 1B,C** and is summarized in **Table 2**. Additional information about the type and maturity group of accessions within each subgroup can be found in **Supplementary Table 1**. Of the total of 477 accessions, 362 were assigned to one of five subgroups (G1 to G5), and the remaining 115 accessions

[4]https://www.soybase.org/GWAS/list.php

[5]https://data.jgi.doe.gov/refine-download/phytozome2

[6]https://www.uniprot.org/

TABLE 1 | Genomic distribution of the 224,993 SNPs considered in this study and their distribution across the 20 chromosomes of Glycine max.

Chromosome	Length (bp)	Number of SNPs	SNP Density[a]	SNP Spacing[b] (bp)	Polymorphic SNPs[c]			LD decay distance (kbp)			π		
					EUCLEG	NJAU-Cultivated	NJAU-Wild	EUCLEG	NJAU-Cultivated	NJAU-Wild	EUCLEG	NJAU-Cultivated	NJAU-Wild
1	56,831,624	11,255	19(54)	3,099	0.77	0.86	0.89	175	145	75	0.20	0.26	0.31
2	48,577,505	11,994	24(57)	2,626	0.85	0.88	0.90	165	105	35	0.28	0.28	0.30
3	45,779,781	10,711	23(57)	2,548	0.79	0.85	0.91	145	80	45	0.24	0.26	0.31
4	52,389,146	10,782	20(56)	2,946	0.83	0.87	0.88	190	100	45	0.29	0.25	0.29
5	42,234,498	9,938	23(72)	2,647	0.73	0.86	0.91	160	100	50	0.20	0.25	0.34
6	51,416,486	12,176	23(60)	2,620	0.82	0.87	0.87	175	75	35	0.24	0.25	0.28
7	44,630,646	10,618	23(88)	2,666	0.80	0.85	0.90	165	90	40	0.22	0.27	0.31
8	47,837,940	12,145	25(71)	2,482	0.79	0.88	0.92	160	80	30	0.20	0.26	0.32
9	50,189,764	11,901	23(59)	2,666	0.84	0.85	0.90	170	105	40	0.27	0.29	0.31
10	51,566,898	11,717	22(67)	2,840	0.80	0.86	0.89	135	105	85	0.23	0.26	0.32
11	34,766,867	8,536	24(68)	2,512	0.74	0.86	0.91	110	60	50	0.17	0.24	0.31
12	40,091,314	9,846	24(63)	2,467	0.68	0.86	0.91	235	95	60	0.19	0.19	0.33
13	45,874,162	12,658	27(80)	2,339	0.78	0.88	0.91	155	75	20	0.24	0.26	0.30
14	49,042,192	11,730	23(63)	2,476	0.81	0.86	0.89	225	100	140	0.23	0.23	0.33
15	51,756,343	11,776	22(57)	2,767	0.85	0.88	0.91	265	120	45	0.26	0.28	0.30
16	37,887,014	9,683	25(62)	2,316	0.86	0.89	0.90	261	115	55	0.27	0.29	0.31
17	41,641,366	10,598	25(71)	2,543	0.84	0.88	0.91	140	100	40	0.24	0.30	0.30
18	58,018,742	14,389	24(59)	2,266	0.86	0.89	0.91	180	155	65	0.20	0.32	0.29
19	50,746,916	11,960	23(64)	2,446	0.83	0.87	0.91	345	170	70	0.25	0.23	0.33
20	47,904,181	10,580	22(64)	2,615	0.77	0.87	0.90	200	105	70	0.21	0.24	0.32
Average	47,459,169	11,250	23(65)	2,594	0.80	0.87	0.90	188	104	55	0.23	0.26	0.31

The main characteristics of the SNPs per chromosome are shown, together with the LD decay distance and average nucleotide diversity (π) per chromosome.
[a]*Average number of SNPs per window of 100 kbp; the maximum number of SNPs for a window of 100 kbp is shown between brackets.*
[b]*Average distance between neighboring SNPs.*
[c]*Proportion of polymorphic SNPs out of total SNPs per chromosome.*

displayed substantial admixture and could not be assigned unequivocally to a specific subgroup. Some association was found between subgroup and origin namely the regions of Southern Europe, Eastern Europe, Western Europe and Northern Europe, while no clear association between subgroup and maturity group (MG) was found. Subgroup G1 contains 93 accessions, 66 of which originate from Southern Europe. They are medium late (MG 0) and late maturing (MG I/II) varieties and breeding lines from Institute of Field and Vegetable Crops, Serbia (IFVCNS), Maize Research Institute Zemun Polje, Serbia (MRIZP) and Agenzia regionale per lo Sviluppo Rurale del Friuli Venezia Giulia, Italy (ERSA FVG), and they group closely with a set of accessions from United States and Canada. Subgroup G2 contains 59 accessions, 26 of which (of all four MGs) originate from Eastern Europe; these accessions group closely with accessions from China. A total 21 of G2 accessions originate from Germany and the Czech Republic and the remaining accessions are from Bulgaria, Poland, Russia, Ukraine and China. The largest subgroup is G3 with 127 accessions, 72 of which originate from Western Europe. In G3, 57 accessions including very early (MG 000) to early maturing (MG 00) breeding lines are from Flanders Research Institute for Agriculture, Fisheries and Food, Belgium (ILVO) and Storm Seeds, Belgium. A set of 16 early maturing varieties from Canada is also part of G3. Finally, G4 and G5 contain 44 and 39 accessions, respectively. G4 contains a unique group of edamame-types from Storm Seeds (Belgium) and from Japan. Subgroup G5 contains accessions from Eastern, Western and Northern Europe; they are mainly MG 000 accessions.

The results of the cluster analysis were similar with those of the fastSTRUCTURE analysis of EUCLEG and NJAU combined (see above). The accessions of the two collections were assigned to different clusters. NJAU-cultivated and NJAU-wild grouped into two separate clusters as in Wang et al. (2016). Interestingly, the accessions of Chinese origin included in the EUCLEG collection clustered mostly among the EUCLEG germplasm, spread over several clusters and only one of them clustered within NJAU-Cultivated (**Supplementary Figure 6**). This indicates that EUCLEG accessions originating from China resemble more closely with other EUCLEG accessions than the Chinese accessions from NJAU.

Genetic Diversity in the EUCLEG and NJAU Collections

A higher number of fixed SNP sites (MAF = 0) was observed in the EUCLEG collection (20%) compared to the NJAU collections (13% and 10% in NJAU-Cultivated and NJAU-Wild, respectively) (**Supplementary Figure 7**). Moreover, the

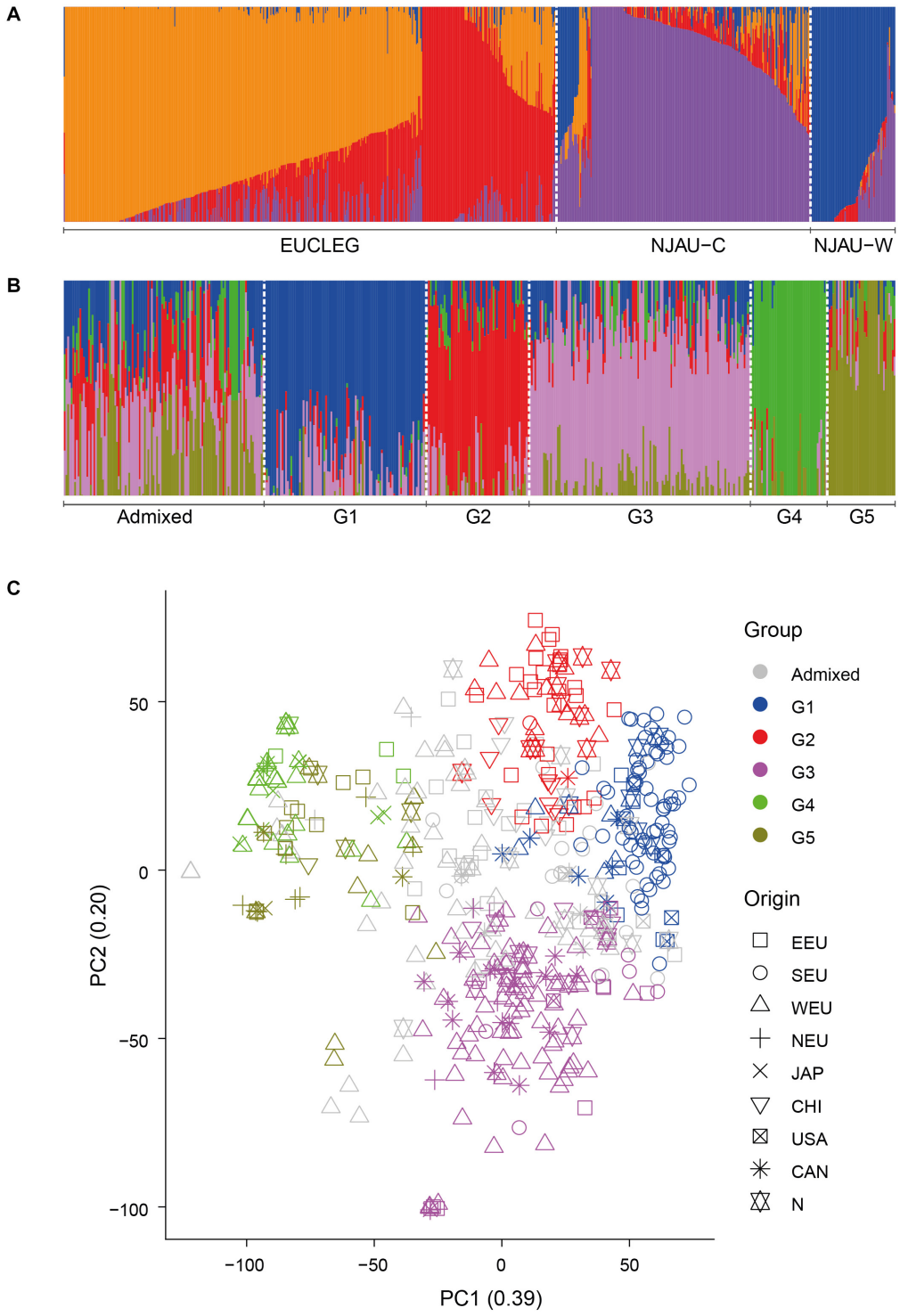


FIGURE 1 | (A) Graphical representation of the fastSTRUCTURE results for EUCLEG and NJAU combined sample sets at optimum K = 4. "NJAU-C" and "NJAU-W" represent the NJAU-Cultivated and NJAU-Wild parts of the NJAU collection. **(B)**. Graphical representation of the fastSTRUCTURE results for the EUCLEG collection at optimum K = 5. "G1" to "G5" are the five subgroups identified by fastSTRUCTURE. "Admixed" are the accessions that could not be assigned unequivocally to one of the subgroups (see main text for further details). **(C)**. Graphical representation of the first two dimensions of a principal components analysis for the EUCLEG collection. PC1 and PC2 explained 39 and 20% of total genetic variation in the EUCLEG collection. EEU: Eastern Europe, SEU: Southern Europe, WEU: Western Europe, NEU: Northern Europe, JAP: Japan, CHI: China, USA: United States of America, CAN: Canada, N: Unknown origin.

TABLE 2 | Summary of fastSTRUCTURE results for the EUCLEG collection.

Geographical Origin*	Number of accessions	G1	G2	G3	G4	G5	Admixed
Eastern EU	77	5	26	10	3	13	20
Southern EU	92	66	1	8			17
Western EU	179	5	13	72	30	10	49
Northern EU	14			3	1	8	2
Japan	9				7	2	
China	21		8	5		1	7
United States	11	5		3			3
Canada	33	6	1	16	2	2	6
Unknown	41	6	10	10	1	3	11
Total	477	93	59	127	44	39	115

The table shows the classification of 477 accessions in different subgroups ("G1 to G5"). "Admixed" corresponds to the group of accessions that could not be assigned unequivocally to any of the five subgroups identified by fastSTRUCTURE. For additional information about the accessions, see ***Supplementary Table 1****.*
**Bulgaria, Czech Republic, Hungary, Moldova, Poland, Romania, Russia, and Ukraine are grouped into Eastern Europe. Italy and Serbia are grouped into Southern Europe. Austria, Belgium, France, Germany, Netherlands, and Switzerland are grouped into Western Europe. Belarus, Estonia, Lithuania and Sweden are grouped into Northern Europe.*

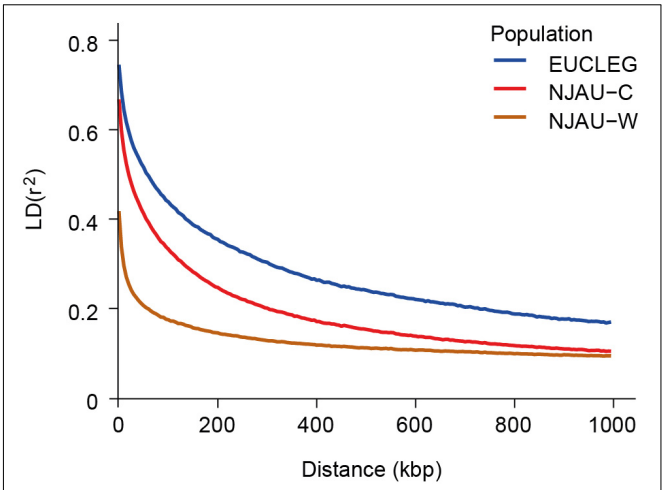


FIGURE 2 | Linkage disequilibrium decay at increasing genetic distances in the EUCLEG collection, the NJAU-Cultivated collection and in the NJAU-Wild collection. X-axis and Y-axis show the average distance between two SNPs in kbp and mean linkage disequilibrium (r^2) within bins of 5 kbp distance, respectively.

proportion of polymorphic SNPs (MAF $\geq$ 5%) was lower in EUCLEG (62%) than in NJAU-Cultivated (72%) and NJAU-Wild (82%). These results suggest an overall higher level of homozygosity in the EUCLEG collection compared to the NJAU collections.

Linkage disequilibrium (LD) dropped to half of its maximum at 175, 100, and 50 kbp in EUCLEG, NJAU-Cultivated and NJAU-Wild, respectively (**Figure 2**), suggesting lower effective population size in the EUCLEG and NJAU-Cultivated collections than in the NJAU-Wild collection. In addition, the three collections showed different patterns of LD per chromosome (**Table 1**), which indicates different histories of recombination and selection in these three collections.

Average pairwise divergence among genotypes per site (π) decreased from 0.31 in NJAU-Wild to 0.26 in NJAU-Cultivated and to 0.23 in EUCLEG (**Table 1**). This is in agreement with a loss of genetic diversity due to domestication and selection in cultivated soybean. There is also a clear tendency toward less variation in average π values per chromosome in NJAU-Wild compared to NJAU-Cultivated and to EUCLEG (coefficient of variation of average π values per chromosome is 4.6, 10.4, and 13.1%, respectively). This is consistent with a scenario of a genetic bottleneck and selection that might have more prominently affected specific chromosomes in the cultivated genepools suggesting that selective sweep analysis comparing EUCLEG and NJAU-Wild may help to identify chromosomal regions that have undergone selection and domestication in the past and that probably are involved in the determination of important agronomic traits.

Selective Sweep Analysis

XP-CLR analysis revealed 23 selective sweep regions with an average length of 1.8 Mbp (range 1.14 Mbp–3.75 Mbp) (**Table 3**), accounting for 4% of the total sequenced genome length. Selective sweep regions were present on 12 different chromosomes including 1, 2, 6, 7, 8, 9, 10, 12, 15, 18, 19, and 20 and some chromosomes had multiple selective sweep regions (**Figure 3** and **Supplementary Figure 8**). Exploration of SoyBase[7] delivered 248 of the total 2,880 previously published QTLs coinciding with the selective sweep regions (**Table 4** and **Supplementary Table 2**). A total of 3,811 genes were positioned within the selective sweep regions. The description of genes is provided in **Supplementary Table 3**. Consistent with the selected proportion of total chromosome size, selective sweep regions on chromosome 7 and 19 contained the highest (1,104) and the lowest (44) number of genes, respectively (**Supplementary Table 3**).

A nearly equal and relatively low nucleotide diversity (π) was observed in selective sweep regions 6.2, 12.2, and 15.1 in both EUCLEG (0.14, 0.06, and 0.23) and NJAU-Cultivated (0.13, 0.06, and 0.21) when compared to NJAU-Wild (0.39, 0.35, and 0.34) (**Table 3** and **Supplementary Figure 9**). This indicates that the traits regulated by these regions may have undergone similar histories of selection in EUCLEG and NJAU-Cultivated. Known QTLs for water use efficiency (WUE) and time to flowering and maturity coincide with region 6.2 and 15.1 (**Table 4**). In addition, genes conferring resistance to pathogens or controlling time to flowering are located in these regions (**Supplementary Table 3**). Strikingly, no QTL coincided with selective sweep region 12.1, although this region harbors genes for important functions such as control of time to flowering (**Supplementary Table 3**).

Interestingly, π was extremely low (0.04 – 0.13) in selective sweep regions 1.1, 1.2, 7.4, 7.5, 8.2, 18.2 and 19.1 in EUCLEG as compared to NJAU-Cultivated (0.16 – 0.33) and NJAU-Wild (0.25 – 0.35), indicating a greater strength of selection on these regions in EUCLEG (**Table 3** and **Supplementary**

[7]https://www.soybase.org/GWAS/list.php

TABLE 3 | Selective sweep regions determined by XP-CLR analysis between EUCLEG and NJAU-Wild.

Selective sweep region	Start (bp)	End (bp)	Number of SNPs	Average XP-CLR	π		
					EUCLEG	NJAU-Cultivated	NJAU-Wild
1.1	6,895,000	8,575,000	388	748	0.08	0.33	0.25
1.2	8,685,000	10,380,000	359	562	0.04	0.22	0.30
2.1	12,600,000	14,860,000	671	705	0.19	0.29	0.27
6.1	5,960,000	7,525,000	343	640	0.20	0.26	0.33
6.2	8,355,000	10,020,000	523	600	0.14	0.13	0.39
7.1	1,540,000	3,615,000	755	1,153	0.21	0.35	0.34
7.2	3,835,000	5,930,000	657	513	0.21	0.33	0.30
7.3	35,825,000	37,295,000	596	509	0.18	0.23	0.33
7.4	38,145,000	40,000,000	602	1,043	0.08	0.17	0.35
7.5	40,275,000	44,025,000	1,103	886	0.08	0.16	0.35
8.1	7,895,000	10,320,000	840	585	0.21	0.30	0.28
8.2	15,190,000	16,905,000	565	558	0.13	0.33	0.31
9.1	2,915,000	4,750,000	669	482	0.20	0.31	0.28
10.1	41,245,000	42,715,000	457	628	0.16	0.26	0.34
10.2	44,055,000	47,045,000	973	426	0.11	0.21	0.32
12.1	5,565,000	6,760,000	424	451	0.14	0.26	0.33
12.2	11,845,000	13,205,000	376	485	0.06	0.06	0.35
12.3	38,140,000	39,650,000	527	489	0.21	0.26	0.33
15.1	560,000	1,720,000	352	471	0.23	0.21	0.34
18.1	4,190,000	5,555,000	457	516	0.20	0.28	0.32
18.2	44,170,000	46,055,000	519	830	0.05	0.26	0.31
19.1	6,260,000	7,400,000	335	425	0.07	0.16	0.31
20.1	33,710,000	35,135,000	515	478	0.25	0.27	0.30
Average			565	616	0.15	0.25	0.32

"Selective sweep region" refers to the name (left side of decimal indicates the chromosome on which the region is located and the right side is an ordinal number), "Start" and "End" delineate the chromosome coordinates of the selective sweep region. "Number of SNPs" is the total number of SNPs contained in the candidate selective sweep region. "Average XP-CLR value" is the average of XP-CLR values for all the windows contained in the selective sweep region. π is the average nucleotide diversity.

Figure 9). Such a low diversity in EUCLEG can be caused by the effects of domestication and/or selection. These regions contain known QTLs for seed fatty acids, seed oil, yield components, resistance to biotic stresses including Sclerotinia stem rot (SCL), time to flowering, and WUE (**Table 4**). Genes controlling flowering and maturity, resistance against pathogens, uptake of minerals, and abiotic stress response are also located in these selective sweep regions (**Supplementary Table 3**). Some of the candidate genes for seed isoflavone content reported by Meng et al. (2016) are located in region 8.2 (*Glyma.08G190300*, *Glyma.08G190500*).

Other selective sweep regions (2.1, 6.1, 7.1, 7.2, 7.3, 8.1, 9.1, 10.1, 10.2, 12.1, 12.3, 18.1, 20.1) have low to medium π (0.11 − 0.25) in the EUCLEG collection compared to that in NJAU-Cultivated (0.21 − 0.35) and NJAU-Wild (0.27 − 0.34) (**Table 3** and **Supplementary Figure 9**). Of the various known QTLs coinciding with these regions, some are related to improvement traits including seed composition (protein, oil and isoflavone content), seed yield (yield components), pathogen resistance, and time to flowering and maturity (**Table 4**). Interestingly, some QTLs associated with tofu quality (tofu hardiness and tofu value) reported by Kurasch et al. (2018) coincide with region 8.1. Different genes known to be involved in nodulation (nodulin *MtN3*, nodulin *MtN21*; Gamas et al. (1996)), regulating Zn, Mn, Ca, and Fe uptake, involved in flowering and maturity including *E2* (*Glyma.10G221500*) and *E4* (*Glyma.20G090000*; in close proximity of selective sweep region 20.1), and genes related to hormonal control of plant growth including auxin response factor, gibberellin-regulated protein, brassinosteroid signaling, jasmonic acid biosynthesis and strigolactone biosynthesis are located in these selective sweep regions (**Supplementary Table 3**).

We also observed known QTLs related to hilum color and seed coat color that coincide with selective sweep region 8.1 (**Table 4**). This region contains a group of chalcone synthase (CHS) genes that are part of the flavonoid and anthocyanin biosynthesis pathway required for seed coat color (Akada and Dube, 1995) (**Supplementary Table 3**).

DISCUSSION

Genotyping Non-Chinese Soybean Accessions With the 355K SoySNP Array

In soybean, different genotyping microarrays are available including the SoySNP50K iSelect Bead chip from Illumina (Song et al., 2013) and the SoySNP180K Axiom microarray from Affymetrix (Lee et al., 2015) containing probes for

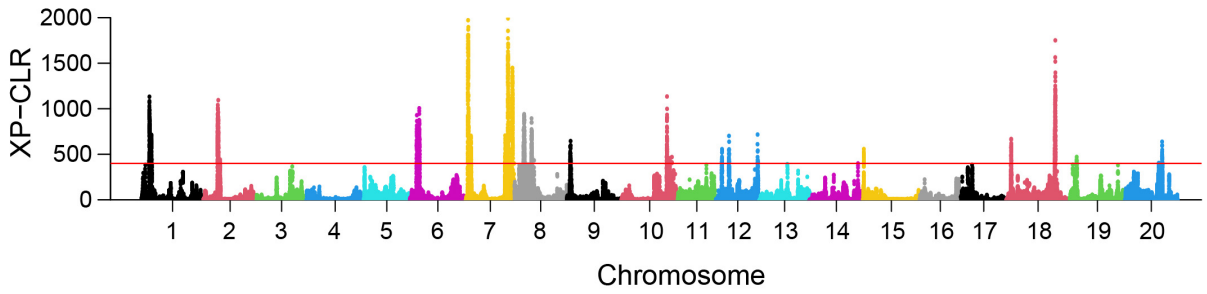


FIGURE 3 | Genome-wide distribution of XP-CLR values in the comparison EUCLEG versus NJAU-Wild. The red line indicates the threshold of the 99th percentile of XP-CLR values. Each dot represents the XP-CLR value obtained for a window of 1 Mbp size, sliding at steps of 5 kbp.

52,041 and 180,961 SNPs, respectively. Here we have used the recently developed NJAU 355K Affymetrix SoySNP array containing probes for 355,595 SNPs (Wang et al., 2016). The proportion of high quality SNPs detected in this study for the EUCLEG collection corresponded quite well with that for the NJAU collection reported by Wang et al. (2016). This indicates that the 355K SoySNP microarray, which was originally developed using plant materials of Chinese origin, is also useful for genotyping soybean from non-Chinese origin. It was therefore possible to perform SNP calling on combined EUCLEG and NJAU collections. This joint analysis offered the advantage that genotyping a larger sample set minimizes the chance of misclassification of SNPs, which reduces the type I error (Mascha and Vetter, 2018). This combined SNP dataset is therefore of great value to analyze the genetic diversity available in the EUCLEG collection and to contrast this with the genetic diversity present in the Chinese collection.

Genetic Structure of the EUCLEG and the NJAU Collections

Soybean is native to China, Japan and Korea, and has been introduced to Europe and other parts of the world (Singh and Hymowitz, 1999; Liu et al., 2017). Evidence from previous studies has shown that the earlier soybean breeding programs of different parts of the world have used Chinese soybean accessions as ancestors (Gizlice et al., 1994; Wysmierski and Vello, 2013), although an earlier study showed a clear distinction between soybean collections from United States and China (Liu et al., 2017). Similarly, population structure analysis in our study revealed a clear genetic differentiation between the EUCLEG and both the NJAU-Wild, and NJAU-Cultivated collections. This reflects breeding efforts in different parts of the world over many decades that have concentrated on improving the local adaptation of soybean to different environmental conditions. Accessions of Chinese origin included in the EUCLEG collection clustered closely with other accessions from EUCLEG instead of clustering with NJAU accessions. This is probably because these accessions have been used for breeding purposes outside China and in this way show closer relationships with their descendants.

The level of genetic diversity (π) in this study was the lowest for the EUCLEG collection, followed by NJAU-Cultivated, and with NJAU-Wild containing the highest level of genetic diversity. This agrees with the model for soybean breeding history presented by Hyten et al. (2006), in which domestication and further selection has reduced the genetic diversity in Asian germplasm. This was followed by genetic bottlenecks during introduction of soybean to other regions of the world and further selection. As a consequence, the substantially lower level of diversity in the EUCLEG collection in comparison to the NJAU collections reflects the combined effect of all three processes (domestication, introduction bottlenecks and selection).

Nevertheless, the EUCLEG collection is strongly structured, with a distribution of genetic diversity over five subgroups. In agreement with previous reports by Žulj Mihaljević et al. (2020), our analysis confirms that soybean accessions from Southern Europe are closely related to those from United States and Canada. Soybean accessions from Eastern and Western Europe contain a range of diversity as they were distributed over all five subgroups (G1 to G5). These results indicate frequent exchange of genetic resources across countries of Eastern and Western Europe, as well as the incorporation of diversity from different geographical origins including Japan, China, United States and Canada into European breeding activities (Tavaud-Pirra et al., 2009; Hahn and Würschum, 2014).

Selective Sweeps in the EUCLEG Collection

We have applied the XP-CLR methodology to determine selective sweeps because compared to other approaches, XP-CLR is robust to determine selective sweeps even in structured populations and has a higher power to detect signals of selection. Moreover it can be used with un-phased genotyping data (Vatsiou et al., 2016). We have identified 23 selective sweep regions spread over 12 chromosomes, that together account for 4% of the total sequenced genome length. This is in accordance with Zhou et al. (2015) who similarly found 5% of the total sequenced genome length affected by selective sweeps when comparing cultivated soybean accessions from different origins to wild Chinese soybean accessions.

TABLE 4 | QTLs in the selective sweep regions.

Selective sweep region	Start (bp)	End (bp)	QTL[a]
1.1	6,895,000	8,575,000	FAT 9-2
1.2	8,685,000	10,380,000	PROT 7-1
2.1	12,600,000	14,860,000	AMIN 10-3, SCN 4-2, K 1-1, K 1-2, FAT 9-4, Pod 1-2, Pod 1-3, Pod 4-1
6.1	5,960,000	7,525,000	DTF 2-3, AMIN 4-1, AMIN 11-1, AMIN 12-1, AMIN 14-2, AMIN 16-1, DTF 7-3, Oil 3-5, SCN 5-12, DTF 4-22
6.2	8,355,000	10,020,000	SDS 1-5, WUE 1-29
7.1	1,540,000	3,615,000	DTF 3-5, SW 4-7, DTF 6-6, FAT 2-5, FAT 3-34, FAT 2-6, FAT 3-35, Zn 1-17, Zn 1-18, Zn 1-19, Zn 1-20, Zn 1-21, DTF 4-23, Oil 8-8, PROT 7-6
7.2	3,835,000	5,930,000	Oil 8-10, K 1-10, K 1-11, K 1-12, K 1-13, K 1-14, K 1-15, K 1-16, K 1-17, DTF 6-7, Pod 1-9, Pod 1-10, Pod 1-11, FAT 1-4, FAT 4-8, WUE 1-46
7.3	35,825,000	37,295,000	SCN 3-1, SCN 1-5, SDS 1-8, P 1-16, SDS 1-9, P 1-17, SDS 1-10, P 1-18, SDS 1-11, P 1-19, SDS 1-12, SDS 1-13, SDS 1-14, SCN 3-2, P 1-20, SDS 1-15, P 1-21, SDS 1-16, P 1-22, SCN 4-4, SDS 1-17, P 1-23, SDS 1-18, P 1-24, SDS 1-19, P 1-25, SDS 1-20, SDS 1-21, SDS 1-22, WUE 1-49, SDS 1-23, SDS 1-24, SDS 1-25, SDS 1-26, SDS 1-27, SDS 1-28, SDS 1-29, SDS 1-30, NF 1-69, NF 1-70, NF 1-71, SDS 1-31, SDS 1-32, SDS 1-33, SDS 1-34, SDS 1-35, SDS 1-36, SDS 1-37, SDS 1-38, WUE 1-50
7.4	38,145,000	40,000,000	Fe 1-8, DTF 2-9, DTF 7-9, P 1-26
7.5	40,275,000	44,025,000	SCN 5-17, SW 4-8, Mn 1-7, Oil 3-7, DTF 6-8, SCN 4-5
8.1	7,895,000	10,320,000	SC 4-1, AMIN 4-4, AMIN 18-1, AMIN 18-2, AMIN 14-3, AMIN 16-2, AMIN 20-1, AMIN 10-10, AMIN 10-11, Pod 4-11, HC 2-2, AMIN 10-12, SC 3-4, HC 2-3, SC 4-2, SC 3-5, HC 1-1, SC 1-6, HC 2-4, SCN 4-6, FAT 5-2, PROT 7-7, SC 1-7, Oil 8-13, AMIN 22-5, AMIN 26-1, AMIN 22-6, AMIN 18-3, AMIN 14-4, AMIN 20-2, AMIN 23-1, AMIN 24-1, AMIN 25-1, FAT 9-7, Salt 1-7, Mg 1-6, FAT 6-4, TH*, TV*
8.2	15,190,000	16,905,000	PUE 2-6, SIFC 1-25, SCN 3-10
9.1	2,915,000	4,750,000	DTF 8-5, Pod 1-16, Pod 1-17, Pod 1-18, AMIN 27-2, WUE 1-59, CAN 1-4, SCL 3-17, SCL 3-18, SCL 3-19, DTF 2-14, DTF 7-14, SCL 3-20, NF 1-72, SCL 3-21
10.1	41,245,000	42,715,000	K 1-28, S 1-9, B 1-13, SMV 2-6
10.2	44,055,000	47,045,000	SCN 1-9, SCN 4-7, DTM 5-3, NF 1-84, NF 1-85, CAN 1-5, DTF 5-25, DTF 5-26, DFTM 1-3, Pod 4-1, Seeds 4-5, PH 5-1, DFTM 1-4, NF 1-86, NF 1-87, NF 1-88, NF 1-89, NF 1-90, DFTM 1-5, FAT 6-7, PubDen 1-8, CAN 1-6, NF 1-91, DTF 5-27, DTF 5-28, PH 5-2, DTM 5-4, FAT 3-36, Nodes 1-2, DTF 5-29, DTF 5-30, DTM 5-5, DTF 8-6, DTM 10-7, DTF 8-2, SIFC 1-29, SCN 5-22
12.1	5,565,000	6,760,000	SW 14-3, Pod 1-24, SW 14-4, SW 3-6, SW 3-7, PubF 1-2, SW 14-5, SW 3-8, DTF 4-46, SCN 5-27
12.2	11,845,000	13,205,000	
12.3	38,140,000	39,650,000	DTM 8-11, WUE 1-6, Pod 1-25, Salt 1-9, WUE 3-24
15.1	560,000	1,720,000	DTF 4-57, DFTM 2-15, WUE 3-27, DFTM 4-15
18.1	4,190,000	5,555,000	DTF 4-69, FAT 9-9, Oil 3-10, Oil 8-23, AMIN 22-16, AMIN 10-22
18.2	44,170,000	46,055,000	P 1-34, P 1-35, P 1-36, PROT 5-2, PubF 1-3
19.1	6,260,000	7,400,000	WUE 1-94, WUE 1-95
20.1	33,710,000	35,135,000	BRA 2-1, DTF 5-64, DTF 5-65, SCN 4-16, LeafShape 1-12, LeafWidth 1-10, WUE 1-99

"Selective sweep region" refers to the name (left side of dot indicates the chromosome on which the region is located and the right side is an ordinal number), "Start" and "End" delineate the chromosome coordinates of the selective sweep region. "QTL" is the quantitative trait locus (retrieved from SoyBase) coinciding with the respective selective sweep region.

*[a]QTL information was retrieved from SoyBase (Grant et al., 2010; www.soybase.org/). The original names of QTLs from SoyBase were adapted. For a more detailed description of these QTLs, see "**Supplementary Table 2**". AMIN: Seed amino acid content, B: Shoot Boron, BRA: Yield component branches on main stem, CAN: Canopy cover, DFTM: Days from flowering to maturity, DTF: Days to flowering, DTM: Days to maturity, FAT: Seed fatty acid content, Fe: Shoot Fe, HC: Hilum color, K: Shoot Potassium, LeafShape: Leaflet shape, LeafWidth: Leaflet width, Mg: Shoot Mg, Mn: Shoot Mn, NF: Nitrogen fixation Ureides content, Nodes: Yield comp Number of nodes per plant, Oil: Seed oil content, P: Shoot P, PH: Plant height, Pod: Pods per plant, PROT: Seed protein, PubDen: Pubescence desnity, PubF: Pubescence form, PUE: P use efficiency, S: Shoot Sulfur, Salt: Salt tolerance, SC: Seed coat color, SCL: Sclerotinia resistance, SCN: Soybean cyst nematode, SDS: Sudden death syndrome, Seeds: Seeds per plant, SIFC: Seed isoflavone content, SMV: Soybean mosaic virus, SW: Seed weight per plant, WUE: Water use efficiency, Zn: Shoot Zinc.*

** QTLs information from Kurasch et al. (2018). TH, Tofu hardiness; TV, Tofu value.*

We have found multiple QTLs for flowering and maturity coinciding with the selective sweep regions. Both are important phenological traits relevant for adaptation of soybean to different cultivation areas. These traits are regulated by the so-called E loci (*E1* to *E10*) (Samanfar et al., 2017). Except for *E6* and *E9*, dominant alleles at other E-loci are photoperiod sensitive and confer late maturity, and photoperiod sensitivity decreases as the number of recessive alleles increases (Bonato and Vello, 1999; Destro et al., 2001; Kong et al., 2014). In this study we found two selective sweep regions in the neighborhood of loci *E2*

and *E4* (region 10.2 and proximal to region 20.1, respectively), suggesting strong signals of selection at these loci in the EUCLEG collection. These results are consistent with previous reports. According to Kurasch et al. (2017) and Miladinović et al. (2018), the European soybean accessions included in their studies contained different haplotypes of four E loci (*E1* to *E4*) with the recessive *e1* and *e2* alleles being more frequently found in Central Europe and the dominant *E3* and *E4* alleles being more frequent in Southern European accessions. The genetic diversity of region 10.2 ($\pi = 0.11$) and 20.1 ($\pi = 0.25$) indicate that the *E2* locus is more fixed than the *E4* locus in the EUCLEG collection. Furthermore, the absence of any previously reported QTL coinciding with the selective sweep region 12.2 containing genes related to time to flowering and hormonal signaling (**Table 3**) provides the opportunity to explore this region for new QTLs. In addition, further selection efforts to increase earliness in European germplasm can either focus on the selection of recessive alleles at the *E4* locus or the exploitation of available genetic diversity present in other loci related to photoperiod sensitivity.

Surprisingly, no significant signatures of selection were detected for stem determinacy, which is an important adaptive trait affecting grain yield in soybean (Kato et al., 2019). Determinate growth habit is a domestication related trait (reviewed in Sedivy et al., 2017), as a high level of determinacy contributes to synchronous seed maturation, thus avoiding undesired variability of moisture content in the harvested material. However, determinate varieties perform less well at high latitudes (Kato et al., 2019) including a large part of Northern and Western Europe (Schori et al., 2003). We have previously shown that stem determinacy is quite variable in the EUCLEG collection (Borra-Serrano et al., 2020), which may explain why the selective sweep analysis presented here failed to obtain relevant signatures of selection for this trait.

In Europe, soybean is considered a protein crop and, together with yield, seed protein content is one of the main breeding goals (Berschneider, 2016). We have identified several signatures of selection that coincide with QTLs for these traits. In addition, selective sweep regions also contain some QTLs related to nutrient use efficiency (e.g., P, Fe, K, and Ca) indicating that they might have been selected for nutrient use efficiency. However, from the total 230 QTLs for seed protein that are described in SoyBase, only 31 were located in the selective sweep regions detected in this study. This, together with the observation that the protein content is higher in wild accessions than in cultivated accessions (Chen and Nelson, 2004) indicates that during the domestication and improvement processes, either some favorable haplotypes for high protein content might have been lost or that the diversity present at those sites might not have been exploited yet in the EUCLEG collection. Therefore, there is still room to improve seed protein content by exploiting the genetic diversity available in the EUCLEG collection for this trait.

Previously known QTLs related to high seed oil content also coincided with the selective sweep regions in the EUCLEG collection compared to NJAU-Wild. This is in line with the observation that wild soybean seeds have low oil content (Chen and Nelson, 2004). Although European soybean is not specifically bred for high oil content, soybean breeding programs in Europe have incorporated breeding materials from United States, where seed oil content is an important trait (Xavier et al., 2018). A low to medium genetic diversity of the respective selective sweep regions in the EUCLEG collection as compared to NJAU-Cultivated and NJAU-Wild suggests that these QTLs are not completely fixed in the EUCLEG collection.

Our analysis has also revealed some signatures of selection related to seed fatty acids including the polyunsaturated fatty acids (linolenic acid and linoleic acid) and monounsaturated fatty acids (oleic acid). Improving the quality of seed oil by minimizing the level of polyunsaturated fatty acids is an important consideration for improved stability of soybean oil (Clemente and Cahoon, 2009). Moreover, signatures of selection were also found for traits related to tofu quality and isoflavone content, a metabolite that helps in the prevention of chronic diseases such as cancer and cardiovascular diseases (Messina and Messina, 2010). This is relevant information for European breeding programs with a particular interest in compositional traits related to food production.

Soybean, a member of the Fabaceae family, has the ability to fix atmospheric nitrogen through symbiosis with rhizobium bacteria. It has been reported that high yielding soybean varieties have a better ability to fix nitrogen (Collino et al., 2015). While direct selection of nitrogen fixation may not have been one of the objectives of current breeding programs, we have found that some previously known QTLs related to nitrogen fixation coincide with the selective sweep regions in the EUCLEG collection. This suggests that this trait might have been improved indirectly through selection for high yield. However, from the total of 145 QTLs for nitrogen fixation reported in SoyBase 11 coincide with selective sweep regions in our analysis. This suggests that still a broad genetic diversity might be present in the EUCLEG collection which can be further used to improve nitrogen fixation.

Resistance to diseases, especially to Sclerotinia stem rot (SCL) caused by *Sclerotinia sclerotiorum*, is considered important in European soybean breeding programs because SCL is widely spread throughout Europe (Rüdelsheim and Smets, 2012). Pannecoucque et al. (2018) reported the presence of genetic variation in the level of susceptibility for SCL in 14 early maturing varieties from Europe. Sources of SCL resistance have been identified on 11 soybean chromosomes and a total of 99 QTLs have been reported in SoyBase (reviewed in Neupane et al., 2019). In our analysis, only 9 QTLs coincided with a selective sweep region. Moreover, soybean mosaic virus (SMV) can be a serious issue in Western and Northern Europe (Aper et al., 2016). Of the 18 QTLs conferring resistance to different strains of SMV reported in SoyBase, only a single QTL coincided with a selective sweep region in the EUCLEG collection. Lack of selection signals for a large number of previously reported QTLs linked to resistance to SCL and SMV suggests a high genetic diversity at the corresponding genomic loci in the EUCLEG collection. A more detailed analysis of the genetic patterns at these QTLs in the EUCLEG collection might be relevant to plan future breeding efforts to improve SCL and SMV resistance in European germplasm.

To the best of our knowledge this is one of the first studies that explores the genetic diversity of a large soybean collection relevant for breeding in Europe, in comparison to Chinese germplasm. Although we have found several selective sweeps that could be linked to useful traits in soybean through XP-CLR analysis, some methodological aspects require attention. First, genotyping using whole genome sequencing rather than microarray data could provide more variants and thus a more detailed description of the genomic regions that have experienced selective sweeps (Ronen et al., 2013). For example, this would enable a gene enrichment analysis which ultimately could provide information about candidate genes. Second, some of the selective sweeps determined by XP-CLR might be false positives caused by demographic processes such as bottlenecks or population expansions (Weigand and Leese, 2018). While the first aspect can be tackled if sufficient resources are available, d it is hard to entirely overcome the second limitation unless other methods are developed that allow to differentiate among the different scenarios that can lead to a positive signal of selection.

CONCLUSION

The present study focused on the exploration of a representative sample of soybean accessions relevant for breeding in Europe, the EUCLEG collection. This is one of the first studies in which the patterns of genetic diversity in a large soybean germplasm set relevant for breeding in Europe has been compared to the genetic diversity contained in Chinese cultivated and wild soybean germplasm. Our study has demonstrated a relatively lower genetic diversity in the EUCLEG collection compared to Chinese collections of cultivated and wild accessions, which indicates a narrow genetic base of the EUCLEG collection. However, a more detailed analysis of the patterns of genetic diversity in the EUCLEG collection has revealed substantial sub-structuration in five subgroups associated with geographical origins, and without a clear association with maturity classes. A selective sweep analysis has revealed the presence of multiple signatures of selection in the EUCLEG collection, compared to Chinese wild germplasm. In particular, genomic regions previously reported to influence grain protein, yield and disease resistance have been identified, whose exploration in future work might facilitate further selection efforts. No signals of selection have been detected for loci involved in stem determinacy, probably because no directed selection has been performed for this trait among the germplasm represented in the EUCLEG collection. On the other hand, clear signatures of selection were detected for at least two loci involved in photoperiod sensitivity and time to flowering, which are main traits considered by breeders in order to adapt soybean for cultivation in Europe (only maturity classes 000 to I/II). Taken together, our results have identified relevant genomic regions that can be further exploited to improve soybean for the European agricultural sector, either through further improvement of genetic resources that are available in Europe, or through incorporation of exotic soybean material in European breeding programs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AS, HM, JA, and IR-R conceived the study. AS and HM performed the research and analyzed the data. AS and IR-R drafted the manuscript. AS, HM, JA, TR, JW, DY, and IR-R interpreted the results and elaborated the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This project has received funding from the European Union's Horizon 2020 Program for Research & Innovation under grant agreement no. 727312 (project: "EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency") and from the Ministry of Science and Technology of China (2017YFE0111000).

ACKNOWLEDGMENTS

We would like to thank Nancy Mergan for technical assistance in DNA extraction. Thanks also to Sabine Van Glabeke for assistance with SNP calling and to Dr. Rex Nelson (Computational Biologist at SoyBase) for sharing with us the description of the GWAS QTL database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.631767/full#supplementary-material

Supplementary Figure 1 | Distribution of SNP categories. PHR, PolyHighResolution; MHR, MonoHighResolution; NMH, NoMinorHomozygote; OTH, Other; OTV, Off-TargetVariants; CRT; CallRateBelowThreshold. EUCLEG, NJAU and Combined represents EUCLEG separate analysis, NJAU separate analysis according to Wang et al. (2016), and EUCLEG and NJAU combined analysis, respectively. PHR, MHR, and NMH are considered recommended categories; OTH, OTV and CRT are considered non-recommended categories. The SNP categories refer to the classification made according to the Axiom™ Analysis Suite 3.1 user guide (www.thermofisher.com/). Y-axis shows the frequency of SNPs (total 335,595).

Supplementary Figure 2 | Frequency distribution of 224,973 SNPs from 355K SoySNP with a certain spacing distance.

Supplementary Figure 3 | Results of the fastSTRUCTURE analysis of EUCLEG and NJAU combined. K (X-axis) represents the number of clusters considered. The Y-axis represents the log likelihood of the model for different number of clusters.

Supplementary Figure 4 | Principal Components analysis (PCA) of the EUCLEG and NJAU combined collection. Blue, red, and orange colors represent the EUCLEG, NJAU-Cultivated and NJAU-Wild collections respectively. PC1 and PC2 explained 44 and 22% of total genotypic variation in the EUCLEG and NJAU combined collection, respectively.

Supplementary Figure 5 | Results of the fastSTRUCTURE analysis of the EUCLEG collection. K (X-axis) represents the number of clusters considered. The Y-axis represents the log likelihood of the model for different number of clusters.

Supplementary Figure 6 | Dendrogram representing the genetic relationships within EUCLEG and NJAU combined. "G1" to "G5" are the five subgroups identified by fastSTRUCTURE. "Admixed" are the accessions that could not be assigned unequivocally to one of the subgroups (see main text for further details). "NJAU-C" and "NJAU-W" represent the NJAU-Cultivated and NJAU-Wild parts of NJAU collection. "China" corresponds to the accessions of Chinese origin that were included in the EUCLEG collection (a total of 21 accessions, indicated by a black bar in front of the tree branch ends).

Supplementary Figure 7 | Polymorphism inEUCLEG and NJAU. NJAU-C, and NJAU-W represents the cultivated and wild subpopulations of NJAU respectively. Y-axis shows the frequency of recommended SNPs (total 224,993).

Supplementary Figure 8 | Results of the XP-CLR analysis comparing EUCLEG and NJAU-Wild. X-axis represents the physical position and Y-axis represents the XP-CLR value. The red line represented in each plot corresponds to the 99th percentile of the genome-wide XP-CLR values.

Supplementary Figure 9 | Comparison of the genetic diversity (π) between EUCLEG (blue), NJAU-Cultivated (red) and NJAU-Wild (orange). The lines represent the average π value of the 1 Mbp windows considered in the XP-CLR analysis. X-axis and Y-axis are represent the genomic position in million base pairs (Mbp) and π value for each chromosome, respectively. The shaded areas represent the selective sweep regions revealed by XP-CLR analysis. The start and end positions of 1 Mbp window with a step size of 5 kbp were defined externally from bedtools (v2.29.2) because XP-CLR does not provide that information in the output. XP-CLR considers recombination frequency instead of actual physical positions to define a window. This is the reason why some shaded areas are shifted a little at both sides of the actual window from XP-CLR (i.e., chromosome 19 and 20) in the representation.

REFERENCES

Akada, S., and Dube, S. K. (1995). Organization of soybean chalcone synthase gene clusters and characterization of a new member of the family. *Plant Mol. Biol.* 29, 189–199. doi: 10.1007/BF00043645

Aper, J., De Clercq, H., and Baert, J. (2016). Agronomic characteristics of early-maturing soybean and implications for breeding in belgium. *Plant Genet. Res.* 14, 142–148. doi: 10.1017/S1479262115000180

Berschneider, J. (2016). Chances and Limitations of European Soybean Production: Market Potential Analysis. Universität Hohenheim, Stuttgart, Germany

Bonato, E. R., and Vello, N. A. (1999). E6, a dominant gene conditioning early flowering and maturity in soybeans. *Genet. Mol. Biol.* 22, 229–232. doi: 10.1590/S1415-47571999000200016

Borra-Serrano, I., De Swaef, T., Quataert, P., Aper, J., Saleem, A., Saeys, W., et al. (2020). Closing the phenotyping gap: high resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sensing* 12:1644. doi: 10.3390/rs12101644

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

Bradshaw, J. E. (2017). Plant breeding: past, present and future. *Euphytica* 213:60. doi: 10.1007/s10681-016-1815-y

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., and Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in canada. *Theor. Appl. Genet.* 132, 3089–3100. doi: 10.1007/s00122-019-03408-y

Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., and Smith, S. (2018). Sustaining the future of plant breeding: the critical role of the USDA-ARS national plant germplasm system. *Crop Sci.* 58, 451–468. doi: 10.2135/cropsci2017.05.0303

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Chen, Y., and Nelson, R. L. (2004).). Genetic variation and relationships among cultivated, wild, and semiwild soybean. *Crop Sci.* 44, 316–325. doi: 10.2135/cropsci2004.3160

Clemente, T. E., and Cahoon, E. B. (2009). Soybean oil: genetic approaches for modification of functionality and total content. *Plant Physiol.* 151, 1030–1040. doi: 10.1104/pp.109.146282

Collino, D. J., Salvagiotti, F., Perticari, A., Piccinetti, C., Ovando, G., Urquiaga, S., et al. (2015). Biological nitrogen fixation in soybean in argentina: relationships with crop, soil, and meteorological factors. *Plant Soil* 392, 239–252. doi: 10.1007/s11104-015-2459-8

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

Destro, D., Carpentieri-Pípolo, V., Afonso De Souza Kiihl, R., and Alves De Almeida, L. (2001). Photoperiodism and genetic control of the long juvenile period in soybean: a review. *Crop Breeding Appl. Biotechnol.* 1, 72–92.

FAOSTAT. (2019). *Food and Agriculture Organization of the United Nations (FAO). FAOSTAT Database*. Available Online at: http://faostat.fao.org/site/291/default.aspx (accessed January 11, 2021).

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Res.* 17, 27–32. doi: 10.1111/1755-0998.12509

Gamas, P., De Carvalho Niebel, F., Lescure, N., and Cullimore, J. V. (1996). Use of a subtractive hybridization approach to identify new Medicago truncatula genes induced during root nodule development. *Mol. Plant-Microbe Int.* 9, 233–242. doi: 10.1094/MPMI-9-0233

Gizlice, Z., Carter Jr., T. E., and Burton, J. W. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* 34, 1143–1151. doi: 10.2135/cropsci1994.0011183X003400050001x

Gizlice, Z., Carter, T. E. Jr., Gerig, T. M., and Burton, J. W. (1996). Genetic diversity patterns in north american public soybean cultivars based on coefficient of parentage. *Crop Sci.* 36, 753–765. doi: 10.2135/cropsci1996.0011183X003600030038x

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Hahn, V., and Würschum, T. (2014). Molecular genetic characterization of central european soybean breeding germplasm. *Plant Breeding* 133, 748–755. doi: 10.1111/pbr.12212

Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J. M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44, 808–811. doi: 10.1038/ng.2309

Hyten, D. L., Song, Q., Zhu, Y., Choi, I. Y., Nelson, R. L., Costa, J. M., et al. (2006). Impact of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U S A* 103, 16666–16671. doi: 10.1073/pnas.0604379103

IDH and IUCN NL. (2019). *European Soy Monitor*. Available Online at: https://www.idhsustainabletrade.com/uploaded/2019/04/European-Soy-Monitor.pdf (accessed November 23, 2019).

Jeong, S. C., Moon, J. K., Park, S. K., Kim, M. S., Lee, K., Lee, S. R., et al. (2019). Genetic diversity patterns and domestication origin of soybean. *Theor. Appl. Genet.* 132, 1179–1193. doi: 10.1007/s00122-018-3271-7

Kato, S., Sayama, T., Taguchi-Shiobara, F., Kikuchi, A., Ishimoto, M., and Cober, E. (2019). Effect of change from a determinate to a semi-determinate growth habit on the yield and lodging resistance of soybeans in the northeast region of Japan. *Breeding Sci.* 69, 151–159. doi: 10.1270/jsbbs.18112

Kong, F., Nan, H., Cao, D., Li, Y., Wu, F., Wang, J., et al. (2014). A new dominant gene E9 conditions early flowering and maturity in soybean. *Crop Sci.* 54, 2529–2535. doi: 10.2135/cropsci2014.03.0228

Kurasch, A. K., Hahn, V., Leiser, W. L., Vollmann, J., Schori, A., Bétrix, C.-A., et al. (2017). Identification of mega-environments in europe and effect of allelic variation at maturity E loci on adaptation of European soybean. *Plant Cell Environ.* 40, 765–778. doi: 10.1111/pce.12896

Kurasch, A. K., Leiser, W. L., Bachteler, K., Miersch, M., Hahn, V., and Würschum, T. (2018). Evaluation of the genetic architecture of tofu traits in soybean towards

genomics-assisted breeding. *Plant Breeding* 137, 873–882. doi: 10.1111/pbr.12651

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

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.* 8:2010. doi: 10.3389/fpls.2017.02014

Mascha, E. J., and Vetter, T. R. (2018). Significance, errors, power, and sample size: the blocking and tackling of statistics. *Anesthesia Analgesia* 126, 691–698. doi: 10.1213/ANE.0000000000002741

Meng, S., He, J., Zhao, T., Xing, G., Li, Y., Yang, S., et al. (2016). Detecting the QTL-allele system of seed isoflavone content in chinese soybean landrace population for optimal cross design and gene system exploration. *Theor. Appl. Genet.* 129, 1557–1576. doi: 10.1007/s00122-016-2724-0

Messina, M., and Messina, V. (2010). The role of soy in vegetarian diets. *Nutrients* 2, 855–888. doi: 10.3390/nu2080855

Miladinović, J., ćeran, M., Đorđević, V., Balešević-Tubić, S., Petrović, K., Đukić, V., et al. (2018). Allelic variation and distribution of the major maturity genes in different soybean collections. *Front. Plant Sci.* 9:1286. doi: 10.3389/fpls.2018.01286

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classification* 31, 274–295. doi: 10.1007/s00357-014-9161-z

Ndjiondjop, M. N., Alachiotis, N., Pavlidis, P., Goungoulou, A., Kpeki, S. B., Zhao, D., et al. (2019). Comparisons of molecular diversity indices, selective sweeps and population structure of african rice with its wild progenitor and asian rice. *Theor. Appl. Genet.* 132, 1145–1158. doi: 10.1007/s00122-018-3268-2

Nei, M. (1972). Genetic distance between populations. *Am. Nat.* 106, 283–292.

Neupane, S., Purintun, J. M., Mathew, F. M., Varenhorst, A. J., and Nepal, M. P. (2019). Molecular basis of soybean resistance to soybean aphids and soybean Cyst nematodes. *Plants (Basel)* 8:374.

Nielsen, R. (2005). *Molecular signatures of natural selection. In Annual Review of Genetics*. California: Annual Reviews. 197–218.

Pagano, M. C., and Miransari, M. (2016). *1 - The importance of soybean production worldwide*. Cambridge: Academic Pres. 1–26.

Pankin, A., Altmüller, J., Becker, C., and von Korff, M. (2018). Targeted resequencing reveals genomic signatures of barley domestication. *New Phytologist* 218, 1247–1259.

Pannecoucque, J., Goormachtigh, S., Heungens, K., Vleugels, T., Ceusters, J., Van Waes, C., et al. (2018). Screening for soybean varieties suited to Belgian growing conditions based on maturity, yield components and resistance to Sclerotinia sclerotiorum and Rhizoctonia solani anastomosis group 2-2IIIB. *J. Agric. Sci.* 156, 342–349. doi: 10.1017/S0021859618000333

Raj, A., Stephens, M., and Pritchard, J. K. (2014). FastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ristova, D., Šarčevič, H., Šimon, S., Mihajlov, L., and Pejič, I. (2010). Genetic diversity in southeast european soybean germplasm revealed by SSR markers. *Agric. Consp. Scie.* 75, 21–26.

Ronen, R., Udpa, N., Halperin, E., and Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics* 195, 181–193. doi: 10.1534/genetics.113.152587

Rüdelsheim, P. L. J., and Smets, G. (2012). Baseline information on agricultural practices in the EU Soybean (Glycine max (L.) Merr.). *Perseus BVBA*, 1–43. Available online at: https://www.europabio.org/sites/default/files/120526_report_eu_farming_practices_soybean.pdf

Savchenko, I. V. (2017). Breeding new varieties and hybrids of agricultural plants. *Herald Russian Acad. Sci.* 87, 104–110. doi: 10.1134/S1019331617020150

Samanfar, B., Molnar, S. J., Charette, M., Schoenrock, A., Dehne, F., Golshani, A., et al. (2017). Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. *Theor. Appl. Genet.* 130, 377–390. doi: 10.1007/s00122-016-2819-7

Schori, A., Charles, R., and Peter, D. (2003). Soja: sélection, agronomie et production en Suisse - agrarforschung suisse. *Agrarforschung* 10, 1–8.

Sedivy, E. J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.* 214, 539–553. doi: 10.1111/nph.14418

Singh, R. J., and Hymowitz, T. (1999). Soybean genetic resources and crop improvement. *Genome* 42, 605–616. doi: 10.1139/g99-039

Smýkal, P., Nelson, M. N., Berger, J. D., and Von Wettberg, E. J. B. (2018). The impact of genetic changes during crop domestication. *Agronomy* 8:119. doi: 10.3390/agronomy8070119

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

Stoskopf, N. C., Tomes, D. T., and Christie, B. R. (2019). *Plant Breeding: Theory and Practice*. Abingdon, UK: Rotledge.

Tavaud-Pirra, M., Sartre, P., Nelson, R., Santoni, S., Texier, N., and Roumet, P. (2009). Genetic diversity in a soybean collection. *Crop Sci.* 49, 895–902. doi: 10.2135/cropsci2008.05.0266

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.* 25, 89–103. doi: 10.1111/mec.13360

Vaughn, J. N., and Li, Z. (2016). Genomic signatures of north american soybean improvement inform diversity enrichment strategies and clarify the impact of hybridization. *G3 (Bethesda)* 6, 2693–2705. doi: 10.1534/g3.116.029215

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep.* 6:20728. doi: 10.1038/srep20728

Weigand, H., and Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zool. J. Linnean Society* 184, 528–583. doi: 10.1093/zoolinnean/zly007

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358. doi: 10.2307/2408641

Wen, Z., Boyse, J. F., Song, Q., Cregan, P. B., and Wang, D. (2015). Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genom.* 16:671. doi: 10.1186/s12864-015-1872-y

Wysmierski, P. T., and Vello, N. A. (2013). The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. *Genet. Mol. Biol.* 36, 547–555. doi: 10.1590/S1415-47572013005000041

Xavier, A., Thapa, R., Muir, W. M., and Rainey, K. M. (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet. Res.* 16, 513–523. doi: 10.1017/S1479262118000102

Zhou, Y., Chen, Z., Cheng, M., Chen, J., Zhu, T., Wang, R., et al. (2018). Uncovering the dispersion history, adaptive evolution and selection of wheat in China. *Plant Biotechnol. J.* 16, 280–291.

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

Žulj Mihaljević, M., Šarèević, H., Lovrić, A., Andrijaniæ, Z., Sudarić, A., Jukić, G., et al. (2020). Genetic diversity of European commercial soybean [Glycine max (L.) Merr.] germplasm revealed by SSR markers. *Genet. Res. Crop Evol.* 67, 1587–1600. doi: 10.1007/s10722-020-00934-3

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.

The reviewer, MS, declared a past co-authorship with one of the authors, TR, to the handling editor.

Copyright © 2021 Saleem, Muylle, Aper, Ruttink, Wang, Yu and Roldán-Ruiz. 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 Plant Science

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fpls.2022.818766

Response of a Diverse European Soybean Collection to "Short Duration" and "Long Duration" Drought Stress

*Aamir Saleem[1,2], Jonas Aper[1†], Hilde Muylle[1], Irene Borra-Serrano[1], Paul Quataert[1], Peter Lootens[1], Tom De Swaef[1] and Isabel Roldán-Ruiz[1,2]**

[1]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium, [2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

Drought causes significant damage to a high value crop of soybean. Europe has an increasing demand for soybean and its own production is insufficient. Selection and breeding of cultivars adapted to European growth conditions is therefore urgently needed. These new cultivars must have a shorter growing cycle (specifically for adaptation to North-West Europe), high yield potential under European growing conditions, and sufficient drought resistance. We have evaluated the performance of a diverse collection of 359 soybean accessions under drought stress using rain-out shelters for 2 years. The contrasting weather conditions between years and correspondingly the varying plant responses demonstrated that the consequences of drought for an individual accession can vary strongly depending on the characteristics (e.g., duration and intensity) of the drought period. Short duration drought stress, for a period of four to 7 weeks, caused an average reduction of 11% in maximum canopy height (CH), a reduction of 17% in seed number per plant (SN) and a reduction of 16% in seed weight per plant (SW). Long duration drought stress caused an average reduction of 29% in CH, a reduction of 38% in SN and a reduction of 43% in SW. Drought accelerated plant development and caused an earlier cessation of flowering and pod formation. This seemed to help some accessions to better protect the seed yield, under short duration drought stress. Drought resistance for yield-related traits was associated with the maintenance of growth under long duration drought stress. The collection displayed a broad range of variation for canopy wilting and leaf senescence but a very narrow range of variation for crop water stress index (CWSI; derived from canopy temperature data). To the best of our knowledge this is the first study reporting a detailed investigation of the response to drought within a diverse soybean collection relevant for breeding in Europe.

Keywords: phenotypic diversity, breeding, *Glycine max*, growth rate, phenology, yield, UAV, growth curve

OPEN ACCESS

Edited by:
Åshild Ergon, Norwegian University of Life Sciences, Norway

Reviewed by:
Florence Volaire, UMR5175 Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), France
Raju Bheemanahalli, Mississippi State University, United States

****Correspondence:***
Isabel Roldán-Ruiz
isabel.roldan-ruiz@ilvo.vlaanderen.be

†Present address:
Jonas Aper, Protealis, Technologiepark-Zwijnaarde, Ghent, Belgium

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *19 November 2021*
Accepted: *24 January 2022*
Published: *17 February 2022*

Citation:
Saleem A, Aper J, Muylle H, Borra-Serrano I, Quataert P, Lootens P, De Swaef T and Roldán-Ruiz I (2022) Response of a Diverse European Soybean Collection to "Short Duration" and "Long Duration" Drought Stress. Front. Plant Sci. 13:818766. doi: 10.3389/fpls.2022.818766

INTRODUCTION

Drought can cause significant damage to crops and may compromise global food security (FAO, 2017). Important yield losses related to drought have been reported for different crops including wheat (27.5%), rice (25.4%), maize (14.0%), and soybean (21.8%; Zhang et al., 2018; Wang et al., 2020a). Main consequences of drought stress in crops are the reduction in leaf area and in stem elongation, which negatively affect the productivity (Mangena, 2018). To withstand drought conditions, plants can adopt a variety of strategies that involve morphological, physiological and biochemical responses, with considerable diversity observed among or even within crop species (Farooq et al., 2012). Important plant responses to drought are changes in stomatal regulation (Pirasteh-Anosheh et al., 2016), changes in the root system (Ye et al., 2018), hormonal changes (Kaur et al., 2016), activation of antioxidant defense systems (Sun et al., 2020) or osmotic adjustments (Turner, 2018). The nature and the magnitude of the crop responses to drought also depends on the duration (and the intensity) of the stress. For example, it has been shown that mild to moderate stress affects mainly stomatal functioning, while severe stress limits the photosynthesis mainly because of damage of the photosynthetic apparatus (Pirasteh-Anosheh et al., 2016; Wang et al., 2018). Timing of the stress situation is also important to consider. For example, if a period of drought occurs early in the season, the plants can compensate for negative effects when more favorable conditions return (Dong et al., 2019; Cui et al., 2020). In contrast, a period of stress at critical developmental phases can cause irreversible damage, leading to high crop productivity losses (Wei et al., 2018; Dong et al., 2019). Given the complexity and diversity of the adaptive mechanisms to drought, the development of drought resistant crops requires the consideration of multiple traits and responses to drought stress as well as their interactions.

Soybean [*Glycine max* (L.) Merr.] is the fourth leading crop worldwide, cultivated on over 120.5 million hectares (FAOSTAT, 2021). Europe is the second largest consumer of soybean after China. With continuously increasing demand, the area of soybean cultivation in Europe almost doubled from 2011 to 2018, up to 5.5 million hectares and 11.6 million tons produced in 2019 (FAOSTAT, 2021). Despite this fast increase in acreage and production volume, local soybean cultivation accounts for only 34% of the total 34.4 million tons consumed in Europe (IDH and IUCN NL, 2019; FAOSTAT, 2021). The selection of adapted varieties through breeding, is an essential step to end European dependence on imported soybean (Van Schoote, 2012; IDH and IUCN NL, 2019; Würschum et al., 2019; Boulch et al., 2021). The industry recognizes this need, with several European breeding programs already underway including those in Austria (Saatbau Linz), Belgium (Storm Seeds and Protealis), France (RAGT), Germany (IG Pflanzenzucht and University of Hohenheim), and Serbia (IFVCNS).

European agriculture is mainly rain-fed, with a share of irrigated area of only 6% (Rossi, 2019). Under the current scenario of changing climatic conditions, the frequency of dry spells associated with low rainfall and high temperature is expected to increase (UNDRR, 2021). Moreover, the period between seedling emergence and reproductive development in soybean can coincide with the occurrence of dry spells in many European regions, as observed in recent years (Bastos et al., 2020; Fu et al., 2020; Peters et al., 2020). This poses serious concerns for European soybean production. Considering the urgent need to expand the soybean cultivation area in Europe under changing climatic conditions, soybean varieties must be developed that are adapted to cultivation in Europe as well as robust to drought stress.

The first and most important step to improve drought resistance of soybean bred for European cultivation is to define effective selection criteria. It has been shown that soybean is only moderately sensitive to drought stress at the seedling stage because water demand is low at this phase in development (Wei et al., 2018; Dong et al., 2019). The developmental stages following flowering are more critical, where soybean plants require sufficient water to achieve maximum yield potential (Kranz and Specht, 2012). Investigating the effect of drought at reproductive stages is therefore considered to be more relevant in this crop (Do Rosário Rosa et al., 2020; Du et al., 2020; Yan et al., 2020). Drought stress at flowering and pod formation reduces the growth rate, leading to shorter plants (Wei et al., 2018); this reduction in plant height has been associated with a decrease in photosynthetic performance (Mak et al., 2014; Zhao et al., 2020). Therefore, the ratio of plant height under drought compared to well-watered conditions is frequently used to estimate the effect of drought (Wang et al., 2020b; Dhungana et al., 2021). Drought at reproductive stages can also cause earlier senescence and a reduction of the leaf area (Wei et al., 2018), which can have a negative impact on yield. Plants that display low leaf senescence (LSEN) under drought maintain a relatively high leaf water content and retain their photosynthetic activity (Rivero et al., 2007), which can be a yield-protecting mechanism. Another frequently investigated trait in soybean is canopy wilting, with slow wilting genotypes tending to be more resistant to drought because of a higher water use efficiency (Ries et al., 2012). Canopy wilting has been also related to canopy temperature, which provides an indirect measure of transpiration rate and stomatal conductance (Bai and Purcell, 2018). Soybean genotypes that utilize water more efficiently under drought conditions maintain a low canopy temperature and are considered more resistant than those generating higher canopy temperatures (Kaler, 2017; Kumar et al., 2017). Drought can also affect the phenology. For example, stress at the reproductive stage can shorten flowering and pod filling duration and increase the rate of senescence (Desclaux and Roumet, 1996). Evaluation of these varied and complex phenological responses is therefore important to understand the impact of drought stress on yield (Jumrani and Bhatia, 2018).

Manual measurements and visual observations are still broadly applied, although interest in the use of unmanned aerial vehicles (UAV) equipped with different imaging sensors for soybean phenotyping is on the rise. UAVs have been employed for quantification of wilting (Zhou et al., 2020), estimation of maturity stage (Yu et al., 2016; Zhou et al., 2019; Trevisan et al., 2020), quantification of plant density (Ranđelović et al., 2020)

or leaf area index (Yuan et al., 2017), and prediction of yield (Yu et al., 2016; Zhang et al., 2019b; Herrero-Huerta et al., 2020; Maimaitijiang et al., 2020; Zhou et al., 2021). Similarly, in previous work (Borra-Serrano et al., 2020) we developed a UAV-based approach to estimate canopy cover and canopy height and to derive parameters related to growth and development in soybean. UAVs that capture images in a short time lapse allow for screening of a large number of plots under the same environmental conditions (e.g., temperature, wind, and light), which is an important advantage when estimating the plant responses to stress situations.

In light of the need to develop soybean cultivars adapted to cultivation in Europe and the anticipated increased risk of drought, the general aim of this study was to investigate how a broad set of soybean accessions relevant for breeding in Europe respond to drought conditions. This can inform soybean breeders about the most relevant characteristics to use in future breeding efforts. Specific objectives of this study were: (1) to describe the phenotypic diversity present in a diverse soybean collection and its potential for breeding efforts in Europe; (2) to evaluate the performance of this collection for traits related to growth and phenology under drought stress conditions, in relation to their performance under well-watered conditions; (3) to assess the "broad-sense" heritability of traits related to drought resistance in this diverse collection; and (4) to identify plant traits that can be considered by breeders to select for drought resistance in germplasm of relevance for Europe.

MATERIALS AND METHODS

Plant Material

A set of 359 soybean accessions originating from 25 countries in Europe, China and the United States were used in this study. This subset of the "EUCLEG collection" described in Saleem et al. (2021) includes accessions of relevance for soybean breeding in Europe (**Supplementary Table S1**). The accessions were divided into four "growth groups" (GP), based on maturity information that was either publicly available at the start of this study, delivered by the provider of the seeds or generated in previous experiments. GP1 ($n = 90$), GP2 ($n = 91$), GP3 ($n = 88$), and GP4 ($n = 90$) comprised accessions expected to belong to maturity group MGI/II, MG0, MG00, and MG000, respectively. However, as the information available was rather limited and was not completely reliable in some cases, this division into groups does not correspond perfectly with the classification in maturity groups.

Field Trials

In 2018, the accessions were sown in two adjacent fields at the same location in Melle, Belgium (51.00° N, 3.80° E) on a sandy loam soil (**Figure 1**). One of these fields was used as "control" treatment under well-watered conditions, while the other was subjected to a drought treatment using rain-out shelters as described below. The design of each field was an augmented row-column design in which three check genotypes (one from each GP1, GP2, and GP4) were replicated nine times, nine check genotypes (two from each GP1, GP2, GP4, and three from GP3) were replicated six times, and 13 check genotypes (three from each GP1, GP3, GP4, and four from GP2) were replicated three times. The remaining 334 genotypes were not replicated. This resulted a total of 454 plots in the control treatment and 454 plots in the drought treatment. The check genotypes were well-characterized varieties (**Supplementary Table S1**). Two similar setups were established on adjacent fields at the same location in 2019 with the same set of soybean accessions (**Figure 1**). Adjacent fields were used in different years to avoid "legacy effects" of the previous soybean trial. The randomization scheme was different for each year.

To apply the recommended row-to-row distance and sowing density for genotypes of different maturity groups (45, 55, 65, and 75 seeds m^{-2} for MGI/II, MG0, MG00, and MG000, respectively), the plot dimensions were slightly adapted for the different GPs. Each plot comprised three rows. For GP1 and GP2, with row-to-row distance of 0.4 m, plot dimensions were 1.20×0.75 m (area 0.90 m^2). For GP3 and GP4, with row-to-row distance of 0.25 m, plot dimensions were 0.75×1.20 m. The four GPs were sown on four different dates in an attempt to synchronize the developmental stage at which drought was imposed (i.e., when 50% of the plots had initiated flowering). Sowing dates in 2018 were 20 April, 2 May, 8 May, and 11 May for GP1, GP2, GP3, and GP4, respectively. In 2019 the sowing dates were 19 April, 30 April, 10 May, and 15 May for GP1, GP2, GP3, and GP4, respectively. These sowing dates were chosen using available data from previous soybean trials in which the optimum sowing time for accessions from different maturity groups had been determined for the study location (data not shown).

Trial Management and Drought Treatment

An overview of the different field activities is provided in **Supplementary Table S2**. In short, the seeds were inoculated with a commercial strain of *Bradyrhizobium japonicum* before sowing according to the manufacturer's instructions. In 2018 the inoculant product was BIODOZ® (De Sangosse, France) but as the same product was not available in 2019, seeds were inoculated with HiStick® (BASF, United States). Both products have been shown to render sufficient nodulation in Belgium (Pannecoucque et al., 2018). Fertilizers were applied before sowing in both years and weeds were controlled by application of herbicides directly after sowing and after manual removal in June. After emergence, thinning was carried out to standardize the plant density to 60% of the total seeding density in GP1 (27 plants m^{-2}) and GP2 (33 plants m^{-2}), while no thinning was required on GP3 and GP4 plots. During the drought treatment, one insecticide application was necessary to control spider mites. Irrigation was applied as required, first manually with a hose at seedling stage and with sprinklers at later stages to maintain sufficient soil moisture in the drought and control fields until flowering had started on 50% of the plots. After that, a period of drought was imposed between 22 June and 18 July in 2018, and between 3 July and 21 August in 2019 to the plots of

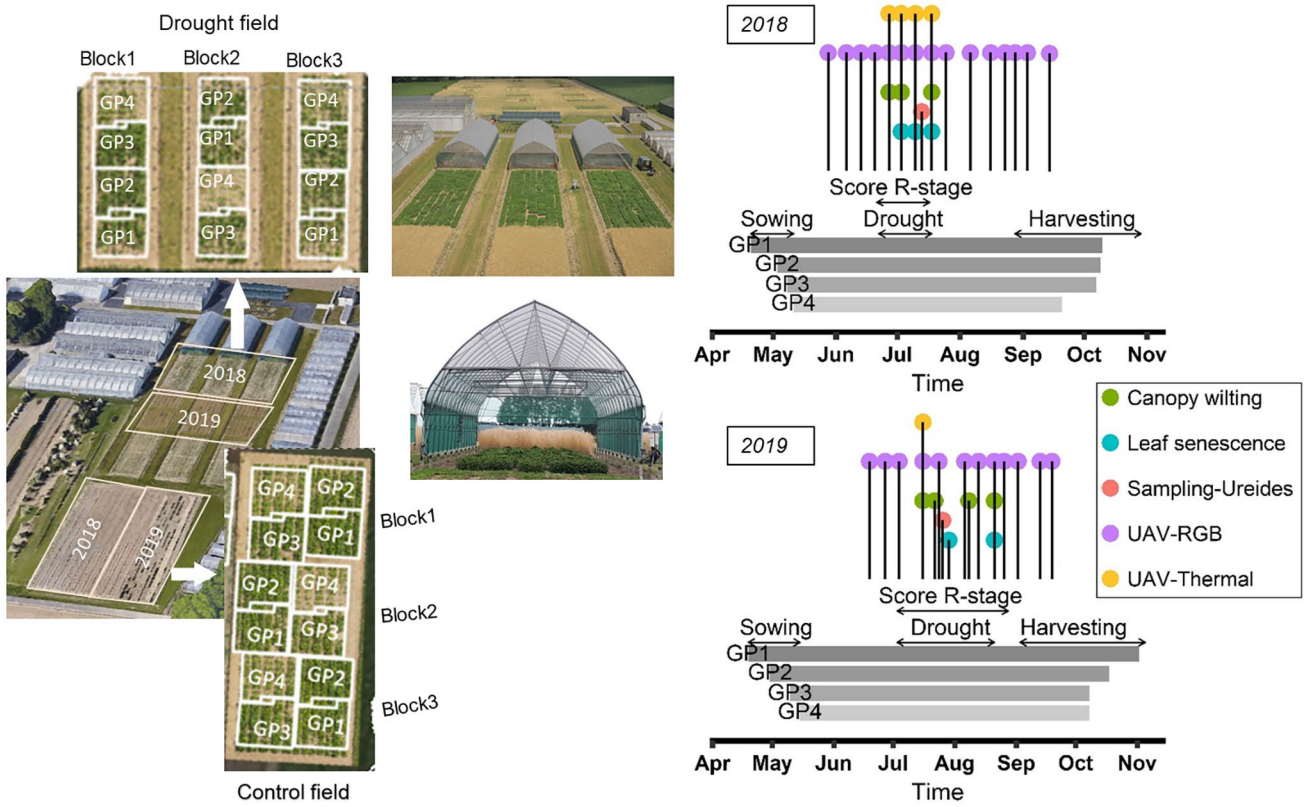


FIGURE 1 | Field plans and timing of the observations. Left: layout of control and drought fields in 2018 and 2019, and general view of the rain-out shelters. Right: schematic representation of the timing of the observations performed in 2018 and 2019. Horizontal arrows delineate the following periods: sowing, determination of R-stage, drought treatment and harvesting. Vertical lines indicate the moments when canopy wilting and leaf senescence were scored, when samples were taken for the determination of ureides concentration, and the timing of the RGB and thermal UAV flights.

the drought field. This was achieved by placing mobile rain-out shelters over the drought field (**Figure 1**). Sprinkler irrigation was continued as required in the control field. In 2018 the soil moisture was monitored in the drought field before, during and after the drought treatment using 36 time domain reflectometer sensors (TDR; type CS616, Campbell Scientific, United Kingdom; 12 per rain-out shelter, **Figure 1**) distributed throughout the field at soil depth of 30 cm and connected with a data logger (CR1000, Campbell Scientific, United Kingdom). TDR measurements were performed only during and after the drought period in 2019. The drought treatment was maintained until the canopy wilting and LSEN symptoms as observed at noon became clearly visible in most of the plots. After the drought treatment, irrigation was resumed in the drought field and was continued until the late developmental stages in control and drought fields. Meteorological conditions were recorded using a temperature and relative humidity (RH) probe (CS215, Campbell Scientific, Inc., United Kingdom), a precipitation sensor (ARG100, Campbell Scientific, Inc., United Kingdom) and a pyranometer (LP02, Hukseflux, Netherlands), all connected to an automatic data logging system (CR1000, Campbell Scientific, Inc., United Kingdom) installed at the trial site. These data were used to calculate the cumulative water deficit (CWD) for the field trial as the accumulation of the difference between daily reference evapotranspiration (ET0 in mm) and precipitation (P in mm), starting at 1 April. ET0 was calculated using the *ET.PenmanMonteith* function of the R package *Evapotranspiration* (Guo et al., 2020). The long-term CWD statistics for the region were calculated using combined weather data sets (from 1979 to 2021) from the Joint Research Centre (JRC MARS Meteorological Database) and the Royal Meteorological Institute (KMI).

Measurements

Figure 1 summarizes the time schedule of the different measurements in 2018 and 2019; **Table 1** describes the traits that were determined. Whenever possible, observations were made in the middle row of each plot.

Visual Measurements

Plant emergence was determined as the percentage of emerged plants in each plot. Plant length up to the second node (PLV) was measured at vegetative stage (between V2 and V5, where V2 and V5 are defined as the stages when a soybean plant displays two and five fully expanded trifoliate leaves, respectively;

TABLE 1 | Description of the traits determined in 2018 and 2019.

Group	Trait[a]	Description	Method of determination
Derived from manual measurements/ visual scores	E	Percentage of seedlings emerged	Expressed as a percentage of the number of seeds sown.
	PLV	Plant length up to the second node (cm)	Determined with a scale in three representative plants per plot. Average value considered. 2018: 28 May (GP1), 1 June (GP2), 5 June (GP3) and 8 June (GP4); 2019: 17 June (GP1), 20 June (GP2), 20 June (GP3) and 2 July (GP4).
	R2	Thermal time from sowing to full flowering (GDD)	From the growth curve fitted for R-stage as a function of thermal time using the sigmoid function from Yin et al. (2003).
	R5	Thermal time from sowing to beginning seed (GDD)	From the growth curve fitted for R-stage as a function of thermal time using a sigmoid function from Yin et al. (2003).
	R8	Thermal time from sowing to pod maturity (GDD)	Observed in the field when 95% of pods per plot reach mature pod color.
	R2R5	Duration of pod formation (GDD)	Determined as the difference between R2 and R5.
	R5R8	Duration of seed development (GDD)	Determined as the difference between R5 and R8.
	R2R8	Thermal time from full flowering to pod maturity (GDD)	Determined as the difference between R2 and R8.
	CW	Canopy wilting (score 1–9; low-high)	2018: 27 June, 3 July, 17 July; 2019: 17 July, 22 July, 8 August, 21 August.
	LSEN	Leaf senescence (score 1–9; low-high)	2018: 4 July, 10 July, 20 July; 2019: 29 July, 21 August.
	PPS	Number of pods on the main stem	Determined on five plants per plot. Average value considered.
	SN	Number of seeds per plant	Determined on five plants per plot. Average value considered.
	SW	Seed weight per plant (g)	Determined on five plants per plot. Average value considered.
UAV-RGB	CC75	Thermal time to canopy cover 75% (GDD)	Thermal time from sowing to canopy cover of 75%. Derived from the fitted growth curve of RGB canopy cover data.
	AGRmax	Maximum absolute growth rate (cm GDD^{-100})	Maximum rate of increase in canopy height. Derived from the fitted growth curve of RGB canopy height data.
	CH	Maximum canopy height (cm)	Maximum canopy height reached by the plot. Derived from the fitted growth curve of RGB canopy height data.
	DET	Degree of indeterminacy (GDD)	Duration of growth after initiation of flowering. Derived from the fitted growth curve of RGB canopy height data as the difference between thermal time to start of flowering and thermal to maximum canopy height.
	SNC	Rate of senescence (0–1)	Represents the rate of plant maturation at the end of cycle. Derived from the fitted growth curve of RGB canopy cover data as the difference between maximum canopy cover and the average lowest cover detected before the end of season.
UAV-Thermal	CWSI	Crop water stress index (0–1)	Index to quantify crop water stress, derived from thermal data according to De Swaef et al. (2021).

The traits have been grouped in three categories according to the mode of determination.

[a]For the determination of R2, R5, and R8 a score of developmental stage was assigned to each plot according to the scale of Fehr and Caviness (1977) during regular visits to the field (between 21 June and 17 July in 2018 and between 3 June and 27 June in 2019). R2R5, R5R8, and R2R8 were determined after correcting for spatial and residual variation in R2, R5, and R8. For the determination of PPS, SN, and SW five plants were collected on the middle row of each plot, bagged and transported to the laboratory.

Fehr and Caviness, 1977). The reproductive stages (R-stages) from R1 to R8 were visually assessed regularly (one to two times per week), using the scale of Fehr and Caviness (1977). For example, R1 corresponds to the time when 50% of the plants in a plot have started flowering. R8 is the stage when 95% of the plants in a plot have reached mature pod color. Typically, the reproductive stage in function of time follows a double sigmoid pattern (R1 to R6 and R6 to R8; Setiyono et al., 2007), and is affected by temperature. Because it was not possible to visualize all R stages for all plots at the time of scoring, R stage data from R1 to R6 was used to fit a growth curve for the R stage as a function of thermal time (growing degree days: GDD) using the sigmoid function from Yin et al. (2003). Thermal time to full flowering (R2) and thermal time to the beginning of seed formation (R5) were then derived from the fitted curve. R2, R5, and R8 data were further used to calculate three other variables after correcting for spatial and residual differences: (1) duration of pod formation (R2R5) as the difference between R2 and R5, (2) duration of seed development (R5R8) as the difference between R5 and R8, and (3) total duration of reproductive development (R2R8) as the difference between R2 and R8. Canopy wilting (CW) was scored during the drought treatment three times in 2018 and four times in 2019. LSEN was scored during and at the end of the drought periods in 2018 and 2019. At the end of the growing period, five plants in the middle row of each plot were harvested and transported to the laboratory for the determination of the number of pods on the main stem (PPS), the number of seeds per plant (SN) and seed weight per plant (SW). The average values of five plants per plot were used for analysis.

Ureides Analysis

It has been demonstrated that drought stress causes an increase in the concentration of ureides in the stems of soybean plants (King and Purcell, 2005; Ray et al., 2015; Cerezini et al., 2017). We therefore checked whether the drought treatment had an effect on the physiology of the plants by quantifying the concentration of ureides in the stem under control and drought conditions. Four plants per plot were collected from the complete set of accessions from control and drought treatments on 13 July 2018 (3 weeks after initiation of the drought treatment) and transported to the laboratory for further processing. In 2019, plant samples were collected on 26 July (3 weeks after initiation of the drought treatment) from a subset of 40 plots

representing 40 accessions (9 GP1, 7 GP2, 8 GP3, and 16 GP4) from both control and drought treatments. This subset was chosen as it represented the total genetic variation of the ureides concentration in the 2018 experiment. The samples were prepared according to the protocol of Unkovich et al. (2008). Leaves were removed and the stem fraction (including the petioles) was oven-dried for 2 days at 70°C. The dry material was ground first into coarse particles using the Peppink 200 AN Mill (Peppink, Netherlands) and then into fine particles using the Ball Mill MM 400 (Retsch, Germany). A homogeneous subsample of 100 mg was taken for the extraction and quantification of ureides according to Coleto et al. (2014).

UAV Measurements

The method developed by Borra-Serrano et al. (2020) was used to determine extra traits using UAV based approaches. RGB images (camera α6000, Sony Corporation, Japan) were taken using a drone (model Onyxstar HYDRA-12, AltiGator, Belgium) during 15 flights between 28 May and 14 September 2018 and 12 flights between 19 June and 19 September 2019 (**Figure 1**). The thermal time needed for the canopy to cover 75% of the soil (CC75), the maximum absolute growth rate (AGRmax), the maximum canopy height (CH), the degree of indeterminacy (DET) and the rate of senescence (SNC) were determined for both years from these images as described in Borra-Serrano et al. (2020). In addition, plant canopy temperature was assessed four times in 2018 (between 26 June and 28 July) and once in 2019 (16 July) using a thermal camera (Wiris 2nd gen, Workswell, Czech Republic) mounted on a drone (model Onyxstar HYDRA-12, AltiGator, Belgium in 2018; model DJI Matrice 600 Pro, DJI, China in 2019). Thermal images were preprocessed in ThermoFormat (Workswell, Czech Republic) and then stitched in Agisoft Photoscan v.1.2.6 Professional Edition (Agisoft LLC, Russia). Canopy temperature data was extracted in QGIS v3.10 (QGIS Geographic Information System; Open Source Geospatial Foundation Project)[1] with polygons defined over the middle row of each plot and only considering the pixels that correspond to vegetation as selected using a vegetation index. Air temperature, recorded at a 10 min interval by an automatic weather station (CR1000, Campbell Scientific, United Kingdom) installed at the experimental site, was used for computing mean ambient temperature for the period of 10:00 to 14:00 h. The crop water stress index (CWSI) was derived from the canopy temperature data according to De Swaef et al. (2021).

Data Analysis

The data was first filtered for plots with more than 30% emergence and then cleaned for outliers according to Tukey's rule (Tukey, 1977). As the treatments in 2018 and 2019 were different (see Results below), a separate analysis per year was carried out. The distribution of the data was inspected using Q-Q plots built in R version 3.6.3 (R Core Team, 2019) and, where necessary, a transformation was applied (indicated in **Supplementary Table S3**). The filtered and cleaned data were then analyzed using mixed linear models with the *lme4* package in R (Bates et al., 2015). The following base model was considered:

$$Y = Genotype + Block + Column + Row \quad (1)$$

Where, "*Y*" is response variable, "*Genotype*" is random effect representing the accession, and "*Block*", "*Column*," and "*Row*" are random effects representing spatial components in the experimental design.

The base model was not applied as such because it would be "overfitted" (it incorporates the "*Block*" and "*Column*" as unique components while in the trial design columns were actually nested in the blocks). For each response variable the six versions of the base model were tested and the output was evaluated using the Akaike Information Criterion (AIC; Akaike, 1974). The best model was then chosen based on the lowest AIC value (**Supplementary Table S3**). From the best model, the best linear unbiased predictor value (BLUP) was calculated for each accession as the sum of the "*Intercept*" value and the value of the random effect of "*Genotype*". Broad sense heritability (H^2) was calculated from the variance components of the best model as follows:

$$H^2 = \frac{Var(Genotype)}{Var(Genotype) + Var(Residuals)} \quad (2)$$

For the interpretation of the heritability values we followed criteria as in Khan et al. (2020): low <30%; medium >30 and < 60%; and high >60%.

Summary statistics of BLUP values including the minimum, maximum, mean and coefficient of variation (CV) for each variable were calculated for the full collection and for each GP separately. To study the stability of the performance of the accessions across 2 years under well-watered conditions, Pearson's correlations between the data of 2018 and 2019 from the control fields were calculated. Pearson's correlations were also calculated between control and drought treatments for each year separately. For the traits that were determined several times each season (CWSI, CW, and LSEN), Pearson's correlations between different time points were calculated.

To determine the response of the accessions to drought stress, a drought index (Yr) was calculated for the variables measured in both treatments (control and drought), according to Araghi and Assad (1998) as follows:

$$Yr = \frac{Control - Drought}{Control} \quad (3)$$

Where,

"Yr" = 0: equal value in control and drought.

"Yr" > 0: lower value in drought than in control.

"Yr" < 0: higher value in drought than in control.

To determine the extent of variation in the response to drought for the measured traits, the coefficient of variation (CV) of the Yr values was calculated. Pearson's correlations

[1]http://www.qgis.org/

between Yr data and control treatment data were calculated in order to check for possible dependencies between the performance of the accessions under well-watered conditions and the strength of response to drought stress. A Principal Component Analysis (PCA) was performed to uncover patterns in this dataset. We used the R software package *factoextra 1.0.7*. (Kassambara and Mundt, 2020). The results were represented using a biplot that combines score and loading plots in a single graph highlighting the most prominent patterns of variation. To simplify the analysis and to get a more accurate representation of the general trends in the data sets, we first identified highly correlated variables that could be removed without loss of information. For this purpose, separate PCAs were carried out with subsets of variables [i.e., variables describing developmental responses (R2-Yr, R5-Yr, R8-Yr, R2R5-Yr, R5R8-Yr, and R2R8-Yr), growth related responses (AGRmax-Yr, CH-Yr, and SNC-Yr), traits of drought resistance (CW-Yr, LSEN-Yr, and CWSI-Yr), and yield responses (PPS-Yr, SN-Yr, and SW-Yr)].

Pearson's correlations between Yr data of 2018 and 2019 were calculated to check the stability of how the accessions responded to drought in the 2 years investigated.

RESULTS

Weather Conditions and Characteristics of the Drought Treatments in 2018 and 2019

The weather conditions in 2018 and 2019 are summarized in **Supplementary Figure S1** and **Supplementary Table S3**. In 2018, the daily average temperature at the time of sowing was slightly higher than in 2019. During the vegetative and reproductive development of the crop, daily average temperature and daily solar radiation were also relatively higher in 2018 than in 2019 with the exception of a small period of 3 days at the end of July 2019 when the temperature rose above the critical threshold of 35°C. Also, at the time of harvesting, the average temperature was higher and the relative humidity was lower in 2018 than in 2019.

At the start of the drought treatment (when the rain-out shelters were placed over the plots) the average soil moisture content was 0.11 v/v in 2018 and 0.12 v/v in 2019 (**Supplementary Figures S1C,D**). The soil moisture content then dropped to an average of 0.06 v/v and 0.05 v/v in 2018 and 2019, respectively. Because of the fast response of the accessions to decreasing levels of soil moisture in 2018 shown as clearly visible symptoms of canopy wilting and LSEN, the rain-out shelters were removed after a period of drought treatment of 3–4 weeks and the soil moisture content was replenished to an average of 0.20 v/v.

Visual symptoms of drought stress developed more slowly in 2019. This was probably due to lower air temperature, lower solar radiation and higher air relative humidity (**Supplementary Table S4**), causing lower evapotranspiration as compared to 2018. These differences are also reflected in the cumulative water deficit index (CWD; **Supplementary Figure S2**). As the moment at which the drought treatment was applied was not exactly the same in the two seasons, it is not easy to compare the plots of 2018 and 2019 in **Supplementary Figure S2**. We see however that the CWD entered the "orange area" (representing the characteristics of "a one in 20 years" season) sooner after imposing the drought treatment in 2018 than in 2019, what might explain the quicker development of visual drought symptoms in 2018. Therefore, after a 6–7 week period of drought treatment in 2019, the rain-out shelters were removed and the soil moisture content was replenished to an average 0.14 v/v, a value which was considered sufficient at that time because most of the accessions had already progressed to advanced stages of reproductive development (many accessions were already at R6 stage) and thus required less water than the earlier reproductive stages as described by Tacker and Vories (2000).

The above results clearly show that the characteristics of the season and of the drought treatment were different in 2018 and 2019. The drought treatment was shorter in 2018, because the plants quickly developed visual symptoms of drought stress (canopy wilting or LSEN). In 2019 the drought treatment lasted longer and the plants only displayed visual symptoms of stress after several weeks of soil water depletion. In what follows, we use the terms "short duration drought stress" and "long duration drought stress" to refer to the drought treatments of 2018 and 2019, respectively.

Overall Characteristics of the Data

Seedling emergence (E) was variable in both years, with some plots displaying extremely low values. To avoid any bias in the results that might be caused by this, plots with $E < 30\%$ were not included in the analysis (for 2018, 29 and 32 plots were removed in control and drought fields, respectively, and for 2019, 30 and 68 plots were removed in control and drought fields, respectively). Considering the different characteristics of 2018 and 2019 (see above), we processed the data for each year separately. Linear mixed models were used to correct for spatial gradients (**Supplementary Table S3**) and BLUP values were calculated (**Table 2; Figure 2**).

For most traits, the level of variation observed in 2018 and 2019 in the control fields was similar. Exceptions were traits related to yield (SN and SW), which exhibited low variation in 2019 as compared to 2018. The levels of variation were also comparable between control and drought treatments for most of the traits with the exception of yield-related traits (PPS, SN, and SW) recorded in 2018 for which the variation was higher in the drought field than in the control field (**Table 2**), indicating a substantial response of the accessions to drought. In 2019 a strong overall reduction of the level of variation for yield-related traits (SN and SW) was observed, which was more pronounced in the drought field (**Table 2**).

The broad-sense heritability (H^2) was high (0.73–0.94) in both years and in both treatments for phenological traits including R2, R5, R8, and SNC (**Table 2**). For yield-related traits H^2 was medium to high (for PPS 0.38–0.71) or low to high (for SN and SW 0.16–0.87). Similarly, medium to high

TABLE 2 | Summary statistics of BLUP values from control and drought fields in 2018 and 2019.

Trait[a]	Treat[b]	2018					2019				
		nObs	Mean ± SD	CV %	H^{2c}	Yr (Mean ± SD)[d]	nObs	Mean ± SD	CV %	H^{2c}	Yr (Mean ± SD)[d]
PLV (cm)	C	334	10.77 ± 1.60	14.86	0.76		331	11.90 ± 1.31	11.01	0.64	
	D	326	10.70 ± 1.90	17.80	0.81		303	11.68 ± 1.19	10.19	0.63	
R2 (GDD)	C	336	853.62 ± 123.96	14.52	0.82	0.00 ± 0.08	336	880.47 ± 158.30	17.98	0.97	0.09 ± 0.06
	D	330	852.54 ± 129.22	15.20	0.92	(317)	308	789.72 ± 138.10	17.49	0.95	(300)
R5 (GDD)	C	321	1145.71 ± 122.89	10.73	0.74	0.03 ± 0.08	320	1208.68 ± 132.50	10.96	0.92	0.09 ± 0.06
	D	322	1101.46 ± 104	9.44	0.73	(298)	308	1098.99 ± 129.81	11.81	0.91	(287)
R8 (GDD)	C	336	1690.18 ± 118.15	6.99	0.86	0.00 ± 0.04	336	1639.16 ± 134.07	8.18	0.94	0.04 ± 0.04
	D	330	1687.91 ± 110.49	6.55	0.86	(317)	305	1563.77 ± 90.38	5.78	0.75	(297)
R2R5 (GDD)	C	311	300.09 ± 80.63	26.87		0.09 ± 0.31	315	341.18 ± 80.61	23.63		0.09 ± 0.19
	D	310	252.25 ± 62.52	24.80		(278)	306	309.31 ± 78.76	25.46		(282)
R5R8 (GDD)	C	317	534.98 ± 77.88	14.56		−0.11 ± 0.17	315	413.17 ± 63.41	15.35		−0.16 ± 0.19
	D	311	583.90 ± 65.88	11.30		(281)	298	466.42 ± 64.99	13.93		(275)
R2R8 (GDD)	C	331	834.20 ± 83.56	10.02		0.00 ± 0.08	331	754.30 ± 98.97	13.12		−0.03 ± 0.10
	D	324	831.40 ± 81.73	9.83		(305)	305	777.74 ± 104.02	13.37		(291)
PPS (–)	C	333	24.01 ± 3.08	12.83	0.39	−0.02 ± 0.22	328	25.66 ± 4.32	16.84	0.56	0.26 ± 0.13
	D	327	24.62 ± 6.47	26.30	0.71	(311)	297	18.50 ± 2.14	11.57	0.38	(283)
SN (–)	C	332	75.78 ± 23.49	31.00	0.67	0.17 ± 0.30	316	73.95 ± 7.59	10.26	0.32	0.38 ± 0.07
	D	323	64.76 ± 31.90	49.30	0.86	(307)	300	45.34 ± 3.93	8.67	0.22	(277)
SW (g)	C	329	13.07 ± 3.83	29.30	0.66	0.16 ± 0.29	321	11.33 ± 1.02	9.00	0.27	0.43 ± 0.06
	D	321	11.05 ± 4.77	43.20	0.87	(305)	294	6.44 ± 0.42	6.52	0.16	(273)
CC75 (GDD)	C	297	477.96 ± 17.32	3.62	0.32		305	545.70 ± 47.04	8.62	0.63	
	D	223	453.03 ± 63.49	14.00	0.75		274	600.89 ± 43.15	7.18	0.46	
AGRmax	C	332	12.51 ± 1.24	9.91	0.51	0.22 ± 0.08	322	12.07 ± 1.09	9.03	0.47	0.22 ± 0.09
(cm GDD^{-100})	D	307	9.70 ± 0.43	4.43	0.24	(292)	295	9.42 ± 1.07	11.11	0.38	(274)
CH (cm)	C	315	87.57 ± 7.97	9.10	0.53	0.11 ± 0.11	315	89.14 ± 10.23	11.48	0.68	0.29 ± 0.08
	D	292	78.00 ± 11.39	14.10	0.83	(270)	280	62.75 ± 3.65	6.35	0.41	(260)
DET (GDD)	C	321	473.8 ± 145.04	30.61	0.61		313	485.60 ± 157.42	32.42	0.73	
	D	293	530.65 ± 95.23	18.00	0.42		288	488.42 ± 48.19	9.87	0.28	
SNC (–)	C	309	0.40 ± 0.20	50.00	0.67	−0.26 ± 0.42	306	0.41 ± 0.29	70.73	0.9	−1.10 ± 1.33
	D	238	0.52 ± 0.20	38.50	0.67	(218)	282	0.69 ± 0.24	34.78	0.79	(223)

Only traits that were determined during or after the drought treatment are considered. "nObs" is the number of observations after removal of plots with E < 30% and outliers, "Treat" is the treatment, "SD" is standard deviation, "CV %" is % genotypic coefficient of variation and "H^2" is the broad sense heritability. PLV: Plant length up to the second node; R2: Thermal time from sowing to full flowering; R5: Thermal time from sowing to beginning seed; R8: Thermal time from sowing to pod maturity; R2R5: Duration of pod formation; R5R8: Duration of seed development; R2R8: Thermal time from full flowering to pod maturity; PPS: Number of pods on the main stem; SN: Number of seeds per plant; SW: Seed weight per plant; CC75: Thermal time to canopy cover 75%; AGRmax: Maximum absolute growth rate; CH: Maximum canopy height; DET: Degree of indeterminacy; SNC: Rate of senescence.

[a]R2R5, R5R8, R2R8 were determined from R2, R5 and R8 data after correcting for residual variation.

[b]"C" represents the control treatment and "D" represents the drought treatment.

[c]H^2 values were not calculated for R2R5, R5R8, and R2R8 as no variance components were estimated.

[d]Figure between brackets indicates the number of common genotypes observed in control and drought treatments. Yr values for PLV and CC75 were not considered as these traits correspond to moments before the initiation of drought period. Yr for DET was not calculated as it is more related to the growth habit of accessions and not directly linked to a response to drought.

H^2 values were obtained for CC75, CH, and DET (0.32–0.83), and low to high H^2 for AGRmax (0.24–0.51). For LSEN, H^2 was low at the start of the drought treatment, but it increased to values higher than 0.5 at later phases (**Table 3**). Similarly, H^2 varied strongly between different dates of observation for CW (0.16–0.51). For CWSI, H^2 was relatively high in the control (0.24–0.48) as compared to the drought field (0.01–0.13). These low H^2 values in the drought fields can be explained by low variance observed in CWSI as discussed below.

Performance Under Well-Watered Conditions

Before analyzing the response of this soybean collection to drought, we investigated the overall performance of the accessions in the control treatment during the two seasons investigated.

General Trends

On average, the plants were slightly taller in 2019 than in 2018, as reflected in the values of PLV and CH (**Table 2**), but for CH this was particularly the case for GP1 and GP4 (**Figure 2M**). The thermal time required by the different GPs to achieve the development to full flowering (R2), seed formation (R5) and pod maturity (R8) in 2018 was in accordance with expectations, with GP1 requiring the highest number of GDD and GP4 the lowest. Correspondingly, the AGRmax increased from GP1 to GP4 (**Figure 2L**). In 2019 accessions of GP4 displayed some delay in development, reaching the R2 and R5 stages later than those of GP3 (**Figures 2B,C**). Delayed emergence of some accessions and leaf damage to several plots caused by rodents in 2019 (illustrated by the higher values of CC75 and lower values of AGRmax in 2019, **Table 2**) can explain this. The values of R2 and R5 were higher in 2019

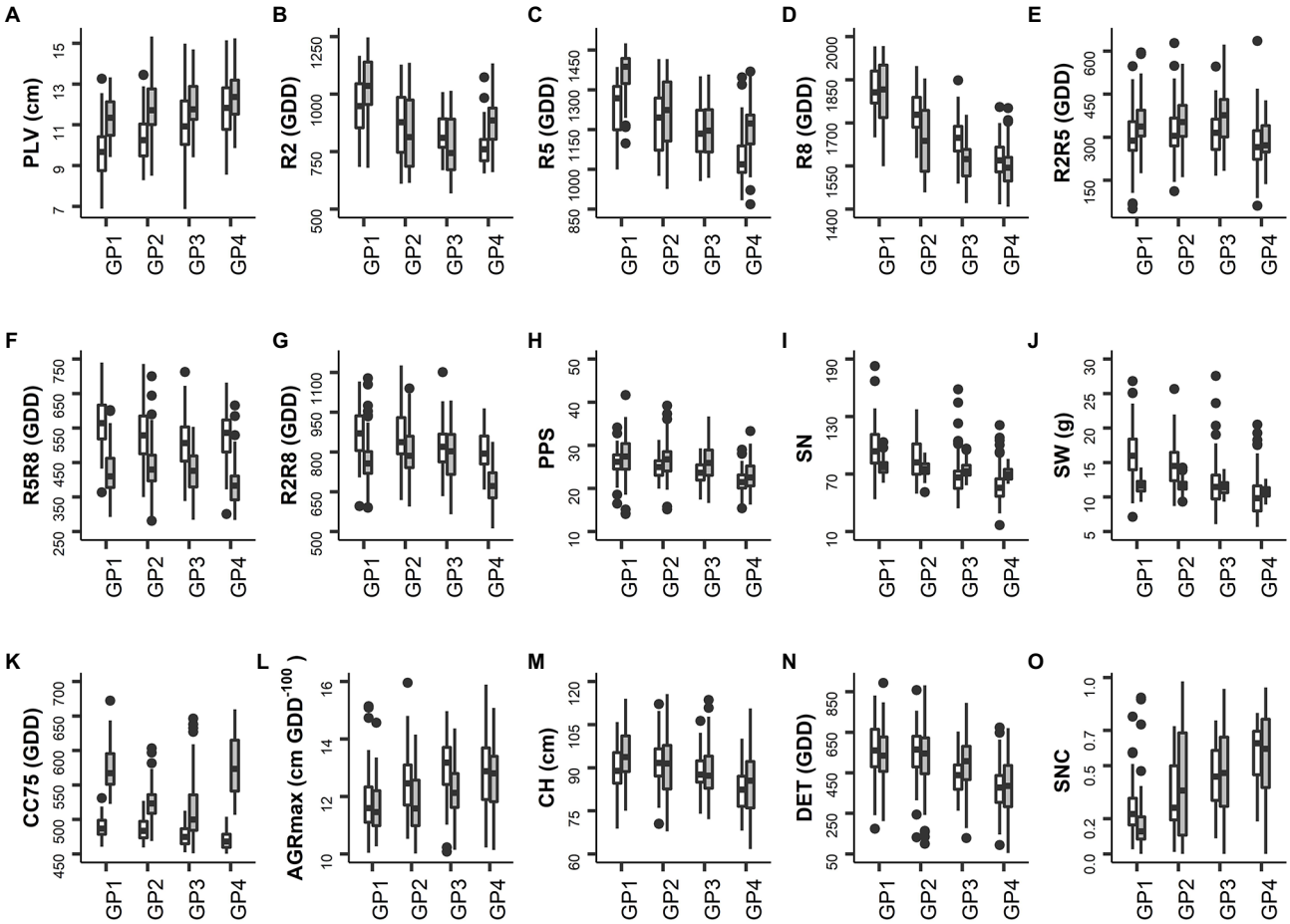


FIGURE 2 | Distribution of the BLUP values per trait in the control field. White and gray shaded whiskers represent values for 2018 and 2019, respectively. Labels on the *X*-axis represent the four groups of accessions with different sowing moments (for details see Materials and Methods). **(A)** PLV: Plant length up to the second node. **(B)** R2: Thermal time from sowing to full flowering. **(C)** R5: Thermal time from sowing to beginning seed. **(D)** R8: Thermal time from sowing to pod maturity. **(E)** R2R5: Duration of pod formation. **(F)** R5R8: Duration of seed development. **(G)** R2R8: Thermal time from full flowering to pod maturity. **(H)** PPS: Number of pods on main stem. **(I)** SN: Number of seeds per plant. **(J)** SW: Seed weight per plant. **(K)** CC75: Thermal time to canopy cover 75%. **(L)** AGRmax: Maximum absolute growth rate. **(M)** CH: Maximum canopy height. **(N)** DET: Degree of indeterminacy. **(O)** SNC: Rate of senescence.

than in 2018 (**Table 2**), indicating a slower path of vegetative development in 2019 and up to the initiation of seed formation. This trend was then reversed, with faster progress to pod maturity (lower value of R5R8 in 2019 than in 2018, **Table 2**). This can be explained by a short spell of high temperature that occurred between R5 and R8 as explained above. With some exceptions, these trends were also found when the different GPs were considered (**Figures 2B–D**).

Regarding traits directly linked to seed yield, there was a decreasing trend for values of PPS, SN, and SW from GP1 (late maturing accessions) to GP4 (early maturing accessions; **Figures 2H–J**). On average, the plants produced a similar number of pods on the main stem (PPS) in both years, but the seed number (SN) and the seed weight per plant (SW) were both lower in 2019 than in 2018 in the late maturing accessions (GP1 and GP2; **Figures 2I,J**). Also the range of variation for SN and SW was rather low in 2019 (**Figures 2I,J**). There seems to be a tendency for a higher level of indeterminacy (DET) in late maturing accessions (GP1) than in early maturing ones (GP4; **Figure 2N**). Differences between years for this trait were only marginal. As expected, the average rate of senescence (SNC) increased from GP1 to GP4 in both years (**Figure 2O**).

Correlation Between Traits and Between Years

Correlations between traits were generally low, except for traits that describe similar aspects, with the highest correlations among R2, R5, R8, and SNC and among PPS, SN, and SW (**Supplementary Figure S3**). Similar trends were found in both years under study, with only a few exceptions. Maximum canopy height (CH) was correlated positively with R2, R5, R8, PPS, SN, and SW in 2018 and 2019, confirming that late maturing accessions grew taller and produced more pods than early maturing accessions. Yield-related traits (PPS, SN, and SW), pod formation duration (R2R5) and seed development duration (R5R8) were not significantly correlated or the correlation

TABLE 3 | Summary statistics of BLUP values from control and drought fields in 2018 and 2019.

Trait	Year	Date[a]	Treatment	nObs	Mean ± SD	CV %	H^2	Yr (Mean ± SD)[b]
LSEN	2018	DAT-12	Drought	319	2.80 ± 0.20	7	0.28	
		DAT-17	Drought	321	3.32 ± 0.33	9.82	0.41	
		DAT-28	Drought	315	4.08 ± 0.52	12.87	0.56	
	2019	DAT-27	Drought	281	3.04 ± 0.08	2.62	0.13	
		DAT-50	Drought	283	4.20 ± 0.68	16.1	0.54	
CW	2018	DAT-5	Drought	329	4.18 ± 0.35	8.46	0.5	
		DAT-11	Drought	326	5.10 ± 0.43	8.37	0.36	
		DAT-25	Drought	306	3.04 ± 0.30	9.7	0.41	
	2019	DAT-15	Drought	308	2.32 ± 0.33	14.1	0.35	
		DAT-20	Drought	307	4.78 ± 0.42	8.8	0.28	
		DAT-37	Drought	302	3.56 ± 0.24	6.8	0.16	
		DAT-50	Drought	308	5.38 ± 0.79	14.7	0.51	
CWSI	2018	DAT-4	Control	319	0.17 ± 0.03	17	0.48	−2.0 ± 0.54 (280)
			Drought	304	0.50 ± 0.01	1	0.06	
		DAT-10	Control	331	0.21 ± 0.02	10.23	0.32	−2.54 ± 0.37 (307)
			Drought	323	0.73 ± 0.01	1.44	0.13	
		DAT-17	Control	327	0.36 ± 0.01	4.15	0.22	−0.43 ± 0.06 (300)
			Drought	323	0.51 ± 0.00	0	0.01	
		DAT-26	Control	330	0.13 ± 0.01	9	0.42	−3.89 ± 0.46 (295)
			Drought	310	0.64 ± 0.00	0	0.07	
	2019	DAT-13	Control	326	0.09 ± 0.01	8.69	0.24	−3.21 ± 0.36 (282)
			Drought	298	0.39 ± 0.00	1.5	0.08	

"nObs" is the number of observations after removal of plots with E < 30% and outliers, "SD" is standard deviation, "CV %" is % genotypic coefficient of variation and "H^2" is the broad sense heritability. LSEN: Leaf senescence; CW: Canopy wilting; CWSI: Crop water stress index.

[a]*"DAT" is days after treatment initiation.*

[b]*Figure between brackets indicates the number of common genotypes observed in control and drought treatments. Yr was not calculated for LSEN and CW as they were determined only in the drought field.*

coefficient was low. This indicates that in this set of accessions the length of these developmental phases (pod formation and seed development) does not linearly associate with seed yield.

In both years, the degree of indeterminacy (DET) correlated negatively with maximum absolute growth rate (AGRmax), showing that determinate accessions attain higher growth rates than indeterminate accessions. DET correlated positively with R2R5 and R2R8, confirming the expected relationship between degree of indeterminacy and duration of reproductive development.

The highest inter-year correlation values were obtained for R2, R5, and R8 ($R = 0.62$–0.8; **Supplementary Figure S4**). Correlation values for other traits including PLV, R2R8, PPS, CH, DET, and SNC were moderate ($R = 0.5$–0.6), while the inter-year correlation was low ($R < 0.4$) for R2R5, R5R8, SN, SW, CC75, and AGRmax.

Drought Stress Responses

The drought treatment was imposed each year at the start of flowering in approximately 50% of the plots in the drought and control fields (**Supplementary Figure S5**). In 2018 the R stage for accessions of different GPs was similar when the drought treatment was initiated, but in 2019 accessions of GP2 and GP3 were more advanced, especially those in the drought field (**Supplementary Figure S5**). Sequential sowing of different GPs (see Materials and Methods section) was therefore quite successful for synchronizing the developmental stage of the different accessions up to initiation of the drought treatment, but some differences were still present. As anticipated, the concentration of ureides was higher in the drought field than in the control field in all the accessions in 2018, as well as in most of the accessions sampled in 2019 (**Supplementary Figure S6**). This revealed that the imposed treatment clearly caused a stress condition.

To quantify the impact of drought, we calculated the drought index (Yr; Equation 2) for the following traits: R2, R5, R8, R2R5, R2R8, R5R8, AGRmax, CH, SNC, PPS, SN, SW, and CWSI. We also included CW and LSEN, which had only been recorded in the drought field. As CC75 was achieved by most of the plots before the initiation of the drought treatment, it was not relevant to calculate Yr for this trait. Yr for DET was also not calculated as it is more related to the growth habit of accessions and not directly linked to a response to drought. First, we investigated the effect of drought on each trait based on Yr values (**Tables 2** and **3**; **Figure 3**) and the between-year correlation (**Figure 4**). We also related the response to drought (Yr) to the performance under well-watered conditions to check whether the performance of the accessions in the control field could explain their response to the drought treatment (**Figures 4, 5**). Finally, we performed a principal component analysis (PCA) to investigate the overall reaction of the accessions (**Figure 6**).

Traits Related to Plant Growth and Development

The CV of Yr values for R2, R5, R8, R2R5, R2R8, R5R8, AGRmax, CH, and SNC ranged from 18.5% to 43.6% (**Table 2**; **Figure 3**), suggesting a broad range of responses to drought. On average, the Yr values in 2019 were higher and more positive than in 2018 for R2, R5, R8, and CH, indicating accelerated development and shorter plants as a consequence of a long duration drought treatment in 2019 (drought reduced CH by 11 and 29% on average in response to short duration and long duration drought,

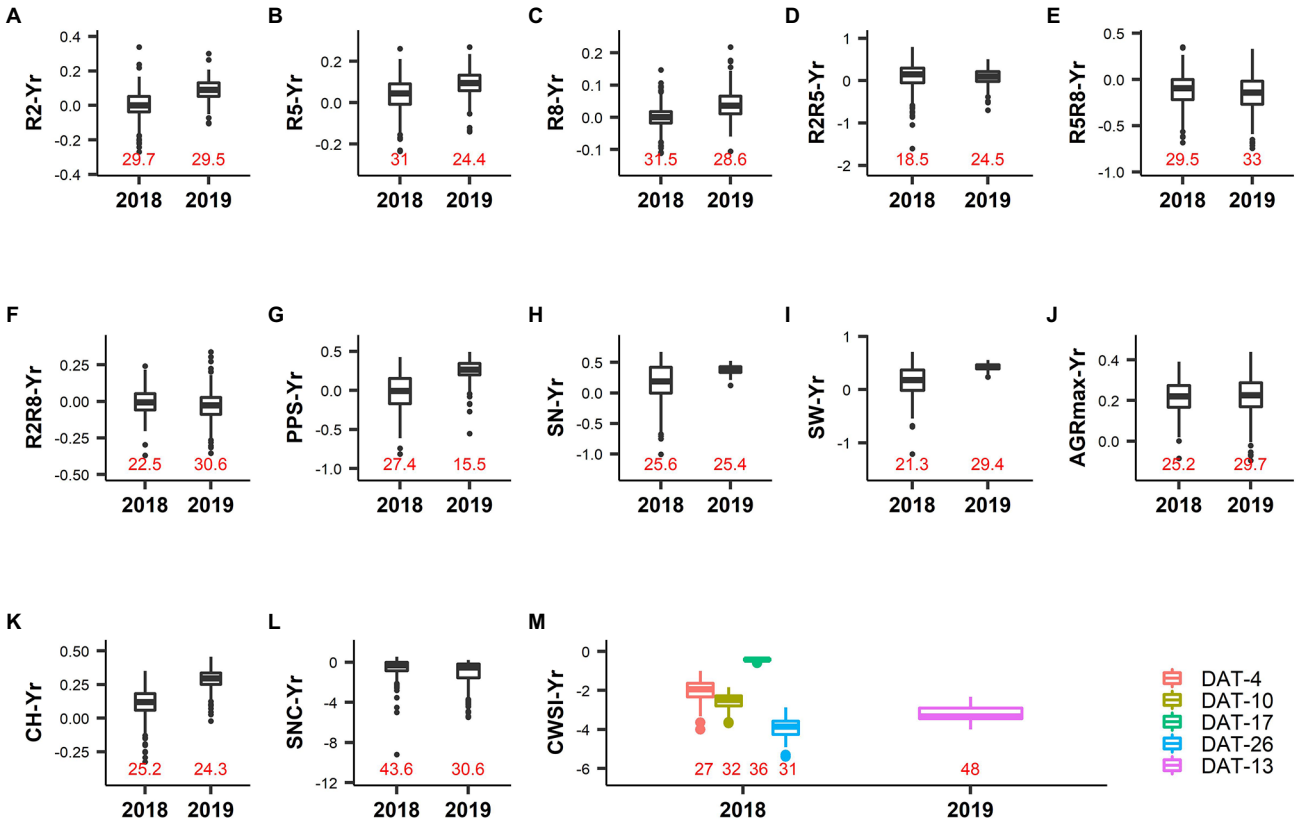


FIGURE 3 | Phenotypic variation in drought index values (Yr) of different traits. Labels on the *X*-axis represent the year of the experiment, and correspond to a short drought treatment (2018) and a long drought treatment (2019). The *Y*-axis represents the drought index value of the respective trait. Data labels inside the plot area (colored in red) indicate the value of the coefficient of variation determined from the normalized Yr data. **(A)** R2: Thermal time from sowing to full flowering. **(B)** R5: Thermal time from sowing to beginning seed. **(C)** R8: Thermal time from sowing to pod maturity. **(D)** R2R5: Duration of pod formation. **(E)** R5R8: Duration of seed development. **(F)** R2R8: Thermal time from full flowering to pod maturity. **(G)** PPS: Number of pods on the main stem. **(H)** SN: Number of seeds per plant. **(I)** SW: Seed weight per plant. **(J)** AGRmax: Maximum absolute growth rate. **(K)** CH: Maximum canopy height. **(L)** SNC: Rate of senescence. **(M)** CWSI: Crop water stress index. Legends represent the measurement day after drought treatment initiation (DAT).

respectively; **Table 2; Figure 3K**). A long drought treatment also induced much earlier senescence (negative Yr values) than a short drought treatment (26 and 110% higher SNC values in the drought fields in 2018 and 2019, respectively; **Table 2** and **Figure 3L**). In both years the vast majority of Yr values for AGRmax was positive (**Figure 3J**) confirming that the drought treatment resulted in an overall reduction of the growth rate (average reduction of AGRmax of 22% in both years; **Table 2**), irrespective of the duration of drought treatment. These results indicate the accessions responded more strongly to the long drought treatment in 2019 as compared to the short duration drought in 2018, in terms of larger reduction of canopy height as well as accelerated development and senescence.

Strong correlations were found between control and drought fields for traits related to developmental stages (R2, R5, R8, R2R5, R5R8, and R2R8) in both years investigated (**Figures 4A–F**). This agrees with the high heritability calculated for these traits (see above). In 2018 (short duration drought treatment) no strong relationship was found between the Yr values and the control treatment values for R2 and R8 (**Figures 4G,I**). The accelerated development observed in 2019 as consequence of the long duration drought treatment (data are below the 1:1 line in **Figure 4C**), was stronger on the late maturing accessions (i.e., $R = 0.67$ between Yr values and values in control for R8, **Figure 4I**). Both of the short duration and the long duration drought treatments caused a prolongation of the duration of the seed development (R5R8) by 11–16% (**Table 2**; the majority of the dots are above the 1:1 line in **Figure 4E**). This was especially the case for many accessions of the early maturing GP4 (**Supplementary Figure S7E**). Conversely, the duration of pod formation (R2R5) was reduced by 9% in response of both of the short duration and the long duration drought treatments (**Table 2**).

In both years, drought caused a reduction of AGRmax and CH (**Figures 4P,Q**). After a long drought treatment, this reduction was stronger in the accessions that displayed the highest AGRmax values under well-watered conditions (**Figure 4V**, $R = 0.43$) or that grew taller under well-watered conditions (**Figure 4W**, $R = 0.84$). While SNC displayed a high variability, the variation was similar under drought and control treatments (**Figure 4R**). Drought accelerated senescence with a relatively stronger effect after a long drought treatment

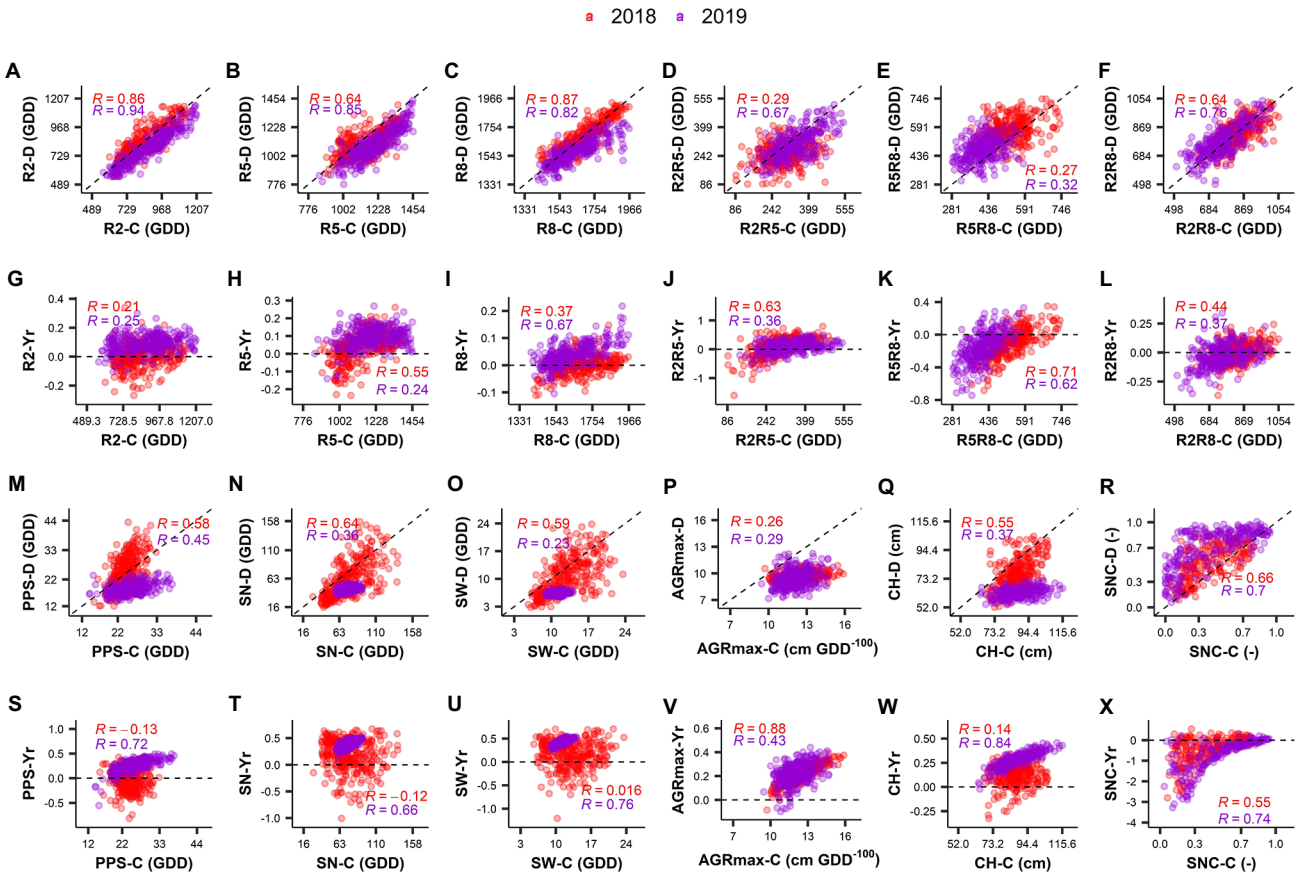


FIGURE 4 | Correlation between BLUP values obtained in the control and drought treatments **(A–F,M–R)** and between drought index and control treatment **(G–L,S–X)** for the different traits in the 2 years investigated (2018 and 2019). "-**(C)**" and "-**(D)**" in the *Y*-axis labels represent the trait in control treatment and in drought treatment, respectively. "-Yr" in the labels of the *Y*-axis represents the drought index value for the respective trait. "R" value is the Pearson's correlation coefficient. Different colors represent the year of the experiment. R2: Thermal time from sowing to full flowering; R5: Thermal time from sowing to beginning seed; R8: Thermal time from sowing to pod maturity; R2R5: Duration of pod formation; R5R8: Duration of seed development; R2R8: Thermal time from full flowering to pod maturity; PPS: Number of pods on the main stem; SN: Number of seeds per plant; SW: Seed weight per plant; Maximum absolute growth rate; CH: Maximum canopy height; SNC: Rate of senescence.

(**Figure 4X**), especially for many accessions of GP1 characterized by later senescence (**Supplementary Figure S7L**).

Physiological Traits

CWSI-Yr values were negative in all cases, indicating a reduction of the transpiration rate caused by drought in all accessions relative to their potential transpiration in the control (**Figure 3M**). In 2018 (short duration drought treatment) the largest difference between control and drought was recorded at DAT-26 (DAT: Days After Drought initiation), followed by DAT-10. This difference was small at DAT-17, indicating a lower level of stress in the drought field relative to DAT-10 and DAT-26. This was probably due to less severe environmental conditions at DAT-17 when the air relative humidity was higher (69% as compared to 41% at DAT-10, data not shown). However, the low level of variation of CWSI values prevents any further interpretation.

While LSEN increased as the drought treatment progressed, CW showed a more erratic behavior over time (**Figures 5K,L**). This was probably due to the way that CW was scored in this work. In comparison to LSEN, CW was more affected by environmental conditions (e.g., air temperature, humidity, solar radiation) when observations were made. In general, the correlation between LSEN values recorded at different dates was thus higher than the correlation between CW values recorded at different dates (**Supplementary Figures S8**, **S9**).

Traits Related to Yield

The response of seed yield-related traits to drought was stronger in 2019 (after a longer period of drought) than in 2018 (after a shorter period of drought), as indicated by a larger deviation of values from the diagonal in the comparisons of control vs. drought (**Figures 4M–O**). Interestingly, after a short drought treatment (2018), a substantial number of accessions displayed higher PPS values in the drought treatment compared to the control (as indicated by negative Yr values in **Figure 4S**). These accessions were mostly late maturing, belonging to GP1 and GP2 (**Supplementary Figure S7G**). In many cases this did not result in an increase of SN or SW (**Figures 4T,U**). Remarkably,

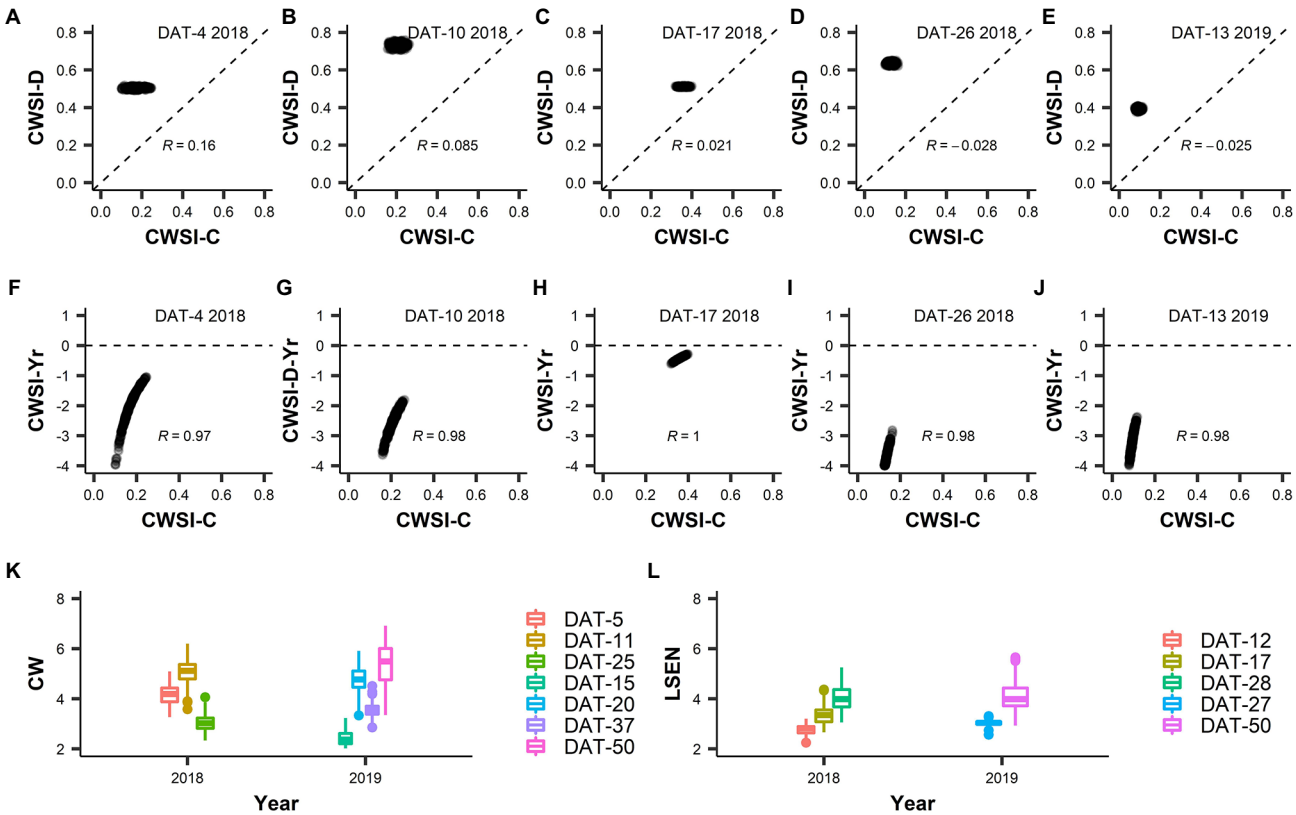


FIGURE 5 | Crop Water Stress Index (CWSI), Canopy wilting (CW) and Leaf senescence (LSEN). **(A–E)** present the correlation of CWSI values between control and drought treatments. **(F–J)** present the correlation between CWSI drought index values (Yr) and values obtained in the control treatment for CWSI. In data labels, "DAT" is measurement day after drought treatment initiation, "2018" and "2019" represent the year of experiment and "R" value is the Pearson's correlation coefficient. Legends in **(K,L)** represent the measurement days (DAT) for CW and LSEN, respectively.

the SN and SW values were very low in 2019 in both the drought and control fields, even though PPS in the control field of 2019 was comparable to that in 2018 (**Figures 4M–O**).

Multiple Trait Responses to Drought Stress

To get an overall view of the response to drought of this soybean collection, a PCA was performed using Yr data as well as CW and LSEN. For CW and LSEN, that were determined on different dates, the highest value observed for the accession at any date was considered in the analysis, and for CWSI, Yr data at the last measurement time was considered. As comparison of Yr values in 2018 (short duration drought stress) and 2019 (long duration drought stress) revealed a lack of correlation between years (**Supplementary Figure S10**), we carried out the analysis for each year separately. Preliminary data inspection for subsets of variables indicated that the following variables could be removed without loss of information: R2-Yr, R8-Yr, R2R8-Yr, and SN-Yr (**Supplementary Figures S11**, **S12**).

In general, a high level of consistency was found for the response of the accessions after a short and a long period of drought treatment, and the overall distribution of the GPs over the biplots (**Figure 6**). In both years, PC1 represented mainly the contrast between duration of pod development (R2R5-Yr) and duration of seed development (R5R8-Yr). The early maturing accessions of GP4 displayed the strongest responses for LSEN and canopy wilting (CW) after a short (2018) and long period (2019) of drought. The main difference between the biplots of short duration and long duration drought treatments is the length and direction of the arrows representing yield-related components (SW-Yr and PPS-Yr) and canopy height (CH-Yr). After a long period of drought (2019), a strong reduction in canopy height (mainly in late maturing accessions of GP1 and GP2) was associated with a strong reduction in yield-related traits. Remarkably, a strong response for CW and LSEN (mainly in early maturing accessions of GP4) was associated with less negative consequences for yield-related traits, as indicated by the opposite direction of arrows representing CW and LSEN on the one side, and SW-Yr and PSS-Yr on the other side (**Figure 6B**). This relationship is not apparent in the biplot of the short drought treatment (**Figure 6A**).

DISCUSSION

A Collection With High Genetic and Phenotypic Diversity

Here we investigate the general characteristics and the response to drought of a large subset of the EUCLEG soybean collection described in Saleem et al. (2021). This collection is of relevance

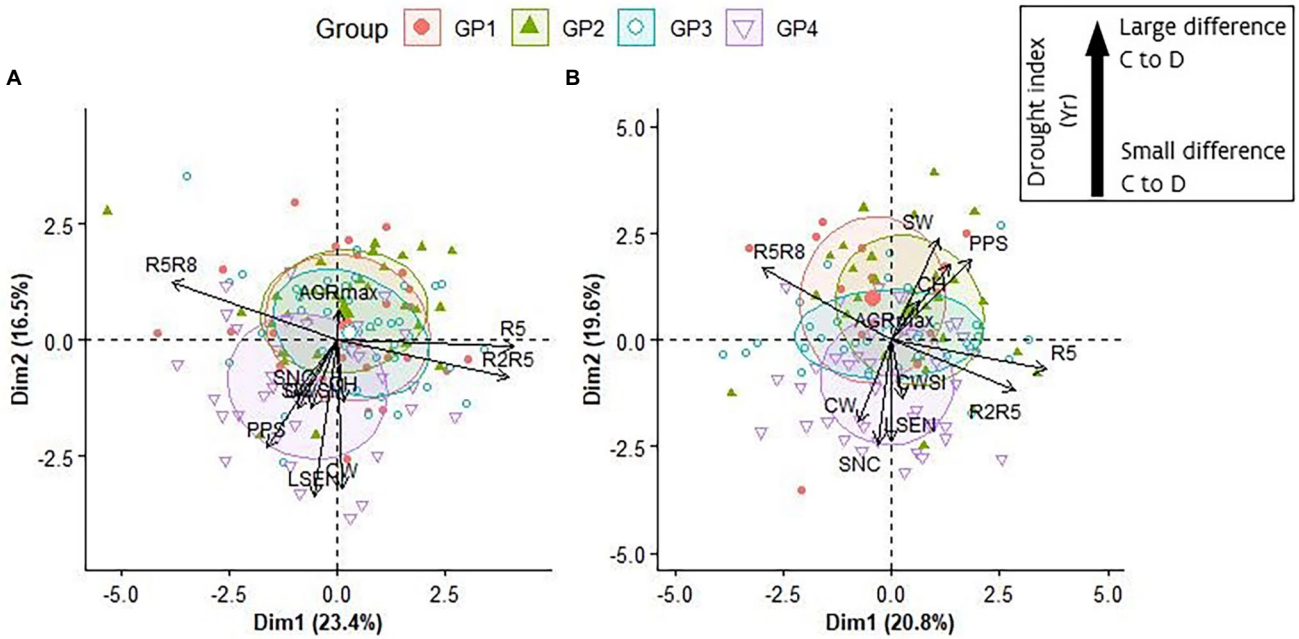


FIGURE 6 | Representation of the first two axes of a Principal Component Analyses (PCA) performed on Yr data of R5, R2R5, R5R8, PPS, SW, AGRmax, CH, SNC and CWSI and BLUP values for LSEN and CW in 2018 **(A)** and in 2019 **(B)**. For representation purposes the "Yr" indicator has been removed from the variable names (e.g., in these plots "R5R8" represents "R5R8-Yr"). The inset in the figure does not apply to the interpretation of SNC values.

for breeding efforts in Europe and contains accessions from maturity classes I/II, 0, 00 and 000. For the purposes of this study, these were classified into four groups (GP1, GP2, GP3, and GP4). Phenotypic evaluation under well-watered conditions during 2 years revealed moderate to high levels of variation for most traits investigated. This was expected given the diverse origin of the accessions and the overall high level of genetic diversity within the collection (Saleem et al., 2021). A longer period of development of GP1 and GP2 accessions was observed which was expected as these accessions were comprised of relatively late maturing MGI/II and MG0 types. Conversely, accessions from GP3 and GP4, displayed a shorter period of development and relatively faster growth. As discussed by Aper et al. (2016), in the location where this study was conducted, the strong vegetative development of early maturing accessions confers them good weed suppression capabilities, and enables the accumulation of the sufficient photosynthates for flowering and seed filling. A decreasing trend for yield traits (PPS, SN, and SW) was observed over maturity duration, with larger values in GP1 and GP2 and smaller values in GP3 and GP4. A similar trend was also described by Aper et al. (2016). Remarkably, the inter-year differences for yield traits (SN and SW) was particularly high for GP1 and GP2 accessions. These accessions originate mostly from Eastern and Southern Europe, while GP3 and GP4 accessions originate mostly from Western and Northern Europe (**Supplementary Table S1**). Given the narrow range of adaptation of soybean varieties bred for a specific region due to sensitivity to photoperiod and temperature (Song et al., 2019), the more stable yield observed in this study for GP3 and GP4 can be explained by a better adaptation of these accessions to Northwest European conditions.

We observed also a wide range of variation for phenological traits (R2, R5, and R8) and a decreasing trend in duration of developmental phases from GP1 to GP4. Surprisingly there was no clear relationship between yield-related traits and either duration of pod formation (R2R5) or seed development (R5R8). The thermal time from sowing to pod maturity (R8) was also not correlated with R2R5 or R5R8. A long duration of reproductive development has been proposed as a strategy to improve soybean yield without affecting the total length of the growing cycle (Metz et al., 1985; Cooper, 2003), and a simulation study predicted a positive relationship between yield and the thermal time to flowering and pod maturity in Northern Europe (Boulch et al., 2021). Our results indicate that the EUCLEG collection contains accessions with the required combination of a long duration of reproductive development and high values of yield traits, with possibly no direct impact on the total length of the growing period.

It has been reported that semi-determinate soybean genotypes can compensate for short adverse periods because of their capacity to produce reproductive organs for longer than determinate types (Zhang et al., 2019a). Semi-determinacy is also considered a good characteristic to introduce in early maturing soybean varieties, provided that taller plants do not have increased risk of lodging (Kato et al., 2019). Correspondingly, multiple studies have expressed the need to develop semi-determinate early maturity soybean material in Europe (Rosenzweig et al., 2003; Schori et al., 2003; Aper et al., 2016). We have found a high range of variation for indeterminacy (DET) in the EUCLEG collection, but we also found that late

maturing types (GP1) are on average more indeterminate than early maturing ones (GP4). However, DET correlated positively with the length of the period between flowering and maturity (R2R8) in GP3 and GP4 ($R = 0.5–0.6$, data not shown), suggesting that by using this collection, semi-determinacy could be combined with a long duration of reproductive development in early maturing types.

The high heritability of traits related to phenological development including R2, R5, and R8 was according to expectations (Zhang et al., 2015; Li et al., 2019). A variable performance for yield-related traits as noticed in our study in the control fields was also expected given the complex quantitative nature of yield, with a strong influence of the environment (Kuswantoro, 2019; Xavier and Rainey, 2020). In contrast, canopy height (CH) and number of pods on the main stem (PPS) were relatively more stable across 2 years under well-watered conditions ($R = 0.55$ and 0.48, respectively). A significant positive correlation between these traits highlights their significance to improve yield in soybean.

Anticipated Multiple and Diverse Responses to Drought in Soybean

Given the high level of diversity contained in the EUCLEG collection, we anticipated that multiple and diverse responses to drought would be present among the genotypes investigated. We therefore performed a thorough evaluation of multiple traits. Traits that are directly related to drought response including LSEN, canopy wilting (CW) and canopy temperature (CWSI), were combined with traits that describe the growth (AGRmax, CH), the developmental path (R2, R5, and R8) and the duration of reproductive development phases (R2R5, R5R8, and R2R8). In addition, because drought resistance should not compromise productivity, yield-related parameters (PPS, SN, and SW) were also considered in the evaluation.

We observed a low to medium level of variation for canopy wilting (CW) and LSEN (**Table 3**). In general, the variation increased as the treatment progressed, indicating a differential response of the accessions to drought. Multiple mechanisms have been proposed to result in slow canopy wilting including low stomatal conductance, deep rooting, constant transpiration under high vapor pressure deficit and low radiation use efficiency (Kunert and Vorster, 2020). While slow or delayed wilting is considered useful in soybean because it can protect yield under drought conditions (Ye et al., 2020), it was not easy to evaluate wilting symptoms in our experiment. This was probably due to the influence of the environmental conditions prevalent at the time of evaluation. Nonetheless, we found a significant correlation between CW and LSEN. As the evaluation of LSEN seems to be less prone to the particular environmental conditions at the time of the evaluation, we consider this a more robust indicator of the response of soybean to water deficit. We did not find a clear relationship between LSEN and drought response for yield-related traits after a short period of drought, but accessions of GP4 displayed higher LSEN values on average than those of other GPs. After a long period of drought, GP4 accessions displayed higher LSEN values on average and were relatively less affected for yield-related traits (determined as number of pods on the main stem and seed weight per plant) than those of other GPs. As GP4 accessions generally grow faster and mature earlier, it is possible that these characteristics help them to show a less reduction in yield when the drought condition is maintained for long. Anyhow our results suggest that under long drought stress, the stronger signs of LSEN might be associated with a high resistance, at least among early maturing accessions. As the level of variation for LSEN within GP4 is substantial, a further study especially in early maturing accessions may help to clarify the relationship between LSEN and response for yield under drought conditions.

Based on a comparison of different indices and approaches, De Swaef et al. (2021) concluded that CWSI can be a complementary criterion to detect differential responses to drought stress in perennial grasses. CWSI has also been employed to determine the level of stress and to schedule irrigation in soybean (Candogan et al., 2013; Tekelioğlu et al., 2017), and Anda et al. (2019) reported higher CWSI values in soybean under drought as compared to a control treatment. Our results were in accordance with this, as higher CWSI values were noted for the drought treatments than well-watered conditions. However, the variation for CWSI in the drought treatments was low. As soybean is a rather isohydric species, plants probably close their stomata even when they are exposed to moderate drought stress (Tardieu and Simonneau, 1998), what might explain the lack of variation observed.

In soybean, a longer period of grain filling has been shown to be advantageous for yield potential at high latitudes (Cooper, 2003). A simulation study estimated an optimum grain filling duration of 60 days for soybean in Northern France (Boulch et al., 2021). In our experiments, drought lengthened the duration of seed development (R5R8, corresponding to the seed filling duration), which was due to an earlier shift to R5 (start of seed formation) which led a shortening of R2R5 (corresponding to the period during which new flowers and pods are formed). Drought response for pod formation duration (R2R5) or grain filling duration (R5R8) was independent from drought response for yield-related traits, as illustrated by the PCA analysis. The reduction of seed yield-related traits under the short duration drought was less prominent in later maturing accessions of GP1 and GP2. The late maturing accessions of GP1 and GP2 in the EUCLEG collection also displayed a higher degree of indeterminacy. It is possible that the drought treatment caused the cessation of flower and pod formation in these accessions, but as they are also more indeterminate, they might have produced more flowers and pods after the end of the drought treatment with less yield penalty than accessions from other maturity groups. As accessions of GP4 are more determinate, it is possible that only the flowers and pods that had been formed before the drought treatment were able to produce seeds, causing a larger difference for yield-related traits between control and drought fields.

Also after a long period of drought, the response for pod formation duration (R2R5) or grain filling duration (R5R8) was independent from drought response for yield-related traits, as illustrated by the PCA analysis in 2019. The prominent responses to a long drought treatment in late maturing accessions of GP1 and GP2 were a reduction in maximum growth rate (AGRmax) and canopy height (CH) along with a reduction in yield-related traits (PPS and SW). Contrary to what was observed after a short period of drought in 2018, the indeterminate

behavior of accessions of GP1 and GP2 was insufficient to protect the yield after a longer drought treatment in 2019. This is probably a reflection of irreversible damage that is common under severe stress (Sehgal et al., 2017; Wei et al., 2018).

Replication of a Drought Experiment in the Field Proved Difficult

Rain-out shelters such as those used for this study allow a good evaluation of the response of plants to drought in the field, as they have limited impact on the air temperature or light conditions and root growth is not limited by the size of the plot as in greenhouse or growth chamber experiments (De Swaef et al., 2021). Using moisture sensors, we also succeeded in attaining similar soil conditions over the 2 years under study. When the rain-out shelters were first positioned over the plots, the average soil moisture content was similar (0.11 v/v in 2018 and 0.12 v/v in 2019), and it dropped to similar levels during the treatment (0.06 v/v and 0.05 v/v in 2018 and 2019, respectively). However, other environmental parameters that strongly affect soybean phenology and development (Wang et al., 1997; Salem et al., 2007; Alsajri et al., 2019; Kumagai and Takahashi, 2020) are difficult to manipulate under field conditions, making it impossible to completely eliminate year-to-year variability. In 2019 the lower air temperature, lower solar radiation and higher relative humidity retarded the development of stress symptoms as compared to 2018. This, in combination with our decision to maintain the drought treatment for a longer period in 2019 than in 2018, had important consequences for the performance of the plants. This led to different environmental conditions experienced by the plants in 2018 and 2019. Correspondingly, we found low to medium stability for most of the trait responses across the 2 years, illustrating that plant responses in a specific drought scenario are not only affected by the reduced soil moisture level but also by other environmental components including temperature, solar radiation and vapor pressure deficit (Tardieu, 2012).

One way to replicate the drought treatment would be to screen for drought resistance using only years with similar environmental characteristics, but this is not practical and perhaps not even possible. Another approach can be to choose areas associated with stable environment across years (Saryoko et al., 2017). An alternative approach, as discussed by Tardieu (2012), is the combination of different methods including phenotyping and modeling: phenotyping in controlled conditions to identify parameters of models, simulation of trait values in a large range of climatic scenarios by using a model with genotype-specific parameters and, finally, testing these models in a limited number of field experiments. This explicitly takes into account the year-to-year variability of drought scenarios, and can be combined with model assisted breeding.

CONCLUSION

We found a wide range of phenotypic diversity for absolute growth rate, canopy height, degree of indeterminacy, phenology and yield-related traits under well-watered conditions in a diverse collection of soybean accessions of relevance for breeding in Europe. Drought applied at the reproductive stage in two seasons brought about diverse responses in this collection. The long duration drought treatment (for 6–7 weeks) in 2019 caused a much stronger response as compared to the short duration drought (for 3–4 weeks) in 2018. Main responses were an average reduction of 11–29% in maximum canopy height, an average reduction of 22% in maximum absolute growth rate and an acceleration of the rate of senescence by 26–110%. Drought also caused a reduction of 9% in the duration of pod formation but conversely an increase of 11–16% in the duration of seed development.

The characteristics of the drought treatment in 2018 and 2019 were different, which resulted in differential responses of the accessions over the 2 years. When a short period of drought was applied (2018) the earlier cessation of flower and pod formation allowed a less pronounced reduction of yield-related traits. A longer duration of the drought stress treatment (2019) brought about a different response. Under these conditions (long drought stress treatment), the accessions that displayed a strong reduction in canopy height (cessation of growth) were also the most affected for yield-related traits. These results suggest that under the conditions associated with a short period of drought stress, drought resistance criteria can be based on yield-related traits, while resistance to long drought stress can be improved by selecting for genotypes that are able to maintain growth. Although stronger signs of LSEN and canopy wilting helped some accessions (mainly GP4) to protect their yield (determined as number of pods on the main stem and seed weight per plant) under long drought stress, further exploration of this relationship especially in early maturing accessions is necessary.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AS, JA, and IR-R conceived the study. AS and JA performed the research. PL and IB-S organized the UAV flights and pre-processed the data. AS, HM, and PQ analyzed the data. AS and IR-R drafted the manuscript. AS, JA, HM, IB-S, PQ, PL, TS, and IR-R interpreted the results and elaborated the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This project has received funding from the European Union's Horizon 2020 Program for Research & Innovation under grant agreement no. 727312 (project: "EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency").

ACKNOWLEDGMENTS

The authors would like to thank Nancy Mergan and Carina Pardon for technical assistance in Ureides Analysis. The authors would also like to thank the field assistants of ILVO-Plant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.818766/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Alsajri, F. A., Singh, B., Wijewardana, C., Irby, J. T., Gao, W., and Reddy, K. R. (2019). Evaluating soybean cultivars for low- and high-temperature tolerance during the seedling growth stage. *Agronomy* 9:13. doi: 10.3390/agronomy9010013

Anda, A., Simon, B., Soós, G., da Silva, J. A. T., and Kucserka, T. (2019). Crop-water relation and production of two soybean varieties under different water supplies. *Theor. Appl. Climatol.* 137, 1515–1528. doi: 10.1007/s00704-018-2660-9

Aper, J., De Clercq, H., and Baert, J. (2016). Agronomic characteristics of early maturing soybean and implications for breeding in Belgium. *Plant Genet. Res.* 14, 142–148. doi: 10.1017/S1479262115000180

Araghi, G. S., and Assad, M. T. (1998). Evaluation of four screening techniques for drought resistance and their relationship to yield reduction ratio in wheat. *Euphytica* 103, 293–299. doi: 10.1023/A:1018307111569

Bai, H., and Purcell, L. C. (2018). Aerial canopy temperature differences between fast- and slow-wilting soya bean genotypes. *J. Agron. Crop Sci.* 204, 243–251. doi: 10.1111/jac.12259

Bastos, A., Fu, Z., Ciais, P., Friedlingstein, P., Sitch, S., Pongratz, J., et al. (2020). Impacts of extreme summers on European ecosystems: a comparative analysis of 2003, 2010 and 2018. *Philos. Trans. R. Soc. B* 375:20190507. doi: 10.1098/rstb.2019.0507

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme 4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/JSS.V067.I01

Borra-Serrano, I., De Swaef, T., Quataert, P., Aper, J., Saleem, A., Saeys, W., et al. (2020). Closing the phenotyping gap: high resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sens.* 12:1644. doi: 10.3390/rs12101644

Boulch, G., Elmerich, C., Djemel, A., and Lange, B. (2021). Evaluation of soybean (*Glycine max* L.) adaptation to northern European regions under different agro-climatic scenarios. *In Silico Plants* 3, 1–13. doi: 10.1093/INSILICOPLANTS/DIAB008

Candogan, B. N., Sincik, M., Buyukcangaz, H., Demirtas, C., Goksoy, A. T., and Yazgan, S. (2013). Yield, quality and crop water stress index relationships for deficit-irrigated soybean [*Glycine max* (L.) Merr.] in sub-humid climatic conditions. *Agric. Water Manag.* 118, 113–121. doi: 10.1016/j.agwat.2012.11.021

Cerezini, P., Dos Santos Lima Fagotti, D., Pípolo, A. E., Hungria, M., and Nogueira, M. A. (2017). Water restriction and physiological traits in soybean genotypes contrasting for nitrogen fixation drought tolerance. *Sci. Agric.* 74, 110–117. doi:10.1590/1678-992x-2016-0462

Coleto, I., Pineda, M., Rodiño, A., De Ron, A., and Alamillo, J. (2014). Comparison of inhibition of N2 fixation and ureide accumulation under water deficit in four common bean genotypes of contrasting drought tolerance. *Ann. Bot.* 113, 1071–1082. doi: 10.1093/aob/mcu029

Cooper, R. L. (2003). A delayed flowering barrier to higher soybean yields. *Field. Crop. Res.* 82, 27–35. doi: 10.1016/S0378-4290(03)00003-0

Cui, Y., Ning, S., Jin, J., Jiang, S., Zhou, Y., and Wu, C. (2020). Quantitative lasting effects of drought stress at a growth stage on soybean evapotranspiration and aboveground BIOMASS. *Water* 13:18. doi: 10.3390/W13010018

Desclaux, D., and Roumet, P. (1996). Impact of drought stress on the phenology of two soybean (*Glycine max* L. Merr) cultivars. *Field Crop. Res.* 46, 61–70. doi: 10.1016/0378-4290(95)00086-0

De Swaef, T., Maes, W. H., Aper, J., Baert, J., Cougnon, M., Reheul, D., et al. (2021). Applying RGB- and thermal-based vegetation indices from UAVs for high-throughput field Phenotyping of drought tolerance in forage grasses. *Remote Sens.* 13:147. doi: 10.3390/rs13010147

Dhungana, S. K., Park, J.-H., Oh, J.-H., Kang, B.-K., Seo, J.-H., Sung, J.-S., et al. (2021). Quantitative trait locus mapping for drought tolerance in soybean recombinant inbred line population. *Plan. Theory* 10:1816. doi: 10.3390/PLANTS10091816

Dong, S., Jiang, Y., Dong, Y., Wang, L., Wang, W., Ma, Z., et al. (2019). A study on soybean responses to drought stress and rehydration. *Saudi J. Biol. Sci.* 26, 2006–2017. doi: 10.1016/J.SJBS.2019.08.005

Do Rosário Rosa, V., Da Silva, A. A., Brito, D. S., Júnior, J. D. P., Silva, C. O., Dal-Bianco, M., et al. (2020). Drought stress during the reproductive stage of two soybean lines. *Pesqui. Agropecuária Bras.* 55, 1–11. doi: 10.1590/S1678-3921.PAB2020.V55.01736

Du, Y., Zhao, Q., Chen, L., Yao, X., Zhang, H., Wu, J., et al. (2020). Effect of drought stress during soybean R2–R6 growth stages on sucrose metabolism in leaf and seed. *Int. J. Mol. Sci.* 21:618. doi: 10.3390/IJMS21020618

FAO (2017). The Future of Food and Agriculture Trends and Challenges. Available at: http://www.fao.org/3/i6583e/i6583e.pdf (Accessed October 15, 2017).

FAOSTAT (2021). Food and Agriculture Organization of the United Nations (FAO). FAOSTAT Database. Available at: http://faostat.fao.org/site/291/default.aspx (Accessed January 11, 2021).

Farooq, M., Hussain, M., Wahid, A., and Siddique, K. H. M. (2012). "Drought stress in plants: an overview," in *Plant Responses to Drought Stress: From Morphological to Molecular Features*. ed. R. Aroca (Berlin, Heidelberg: Springer), 1–33.

Fehr, W., and Caviness, C. (1977). *Stages of Soybean Development. Special Report*. Tainan: AVRDC.

Fu, Z., Ciais, P., Bastos, A., Stoy, P. C., Yang, H., Green, J. K., et al. (2020). Sensitivity of gross primary productivity to climatic drivers during the summer drought of 2018 in Europe. *Philos. Trans. R. Soc. B* 375:20190747. doi: 10.1098/rstb.2019.0747

Guo, D., Westran, S., and Peterson, T. (2020). Evapotranspiration: modelling actual, potential and reference crop evapotranspiration. R Package Version 1.15. Available at: https://CRAN.R-project.org/package=Evapotranspiration (Accessed December 29, 2020).

Herrero-Huerta, M., Rodriguez-Gonzalvez, P., and Rainey, K. M. (2020). Yield prediction by machine learning from UAS-based mulit-sensor data fusion in soybean. *Plant Methods* 16, 1–16. doi: 10.1186/s13007-020-00620-6

IDH and IUCN NL (2019). European Soy Monitor. Available at: https://www.idhsustainabletrade.com/uploaded/2019/04/European-Soy-Monitor.pdf (Accessed November 23, 2019).

Jumrani, K., and Bhatia, V. (2018). Impact of combined stress of high temperature and water deficit on growth and seed yield of soybean. *Physiol. Mol. Biol. Plants* 24, 37–50. doi: 10.1007/s12298-017-0480-5

Kaler, A. (2017). Genomic and physiological approaches to improve drought tolerance in soybean part of the agronomy and crop sciences commons, plant breeding and genetics commons, and the statistics and probability commons [University of Arkansas]. Available at: http://scholarworks.uark.edu/etdhttp://scholarworks.uark.edu/etd/2490 (Accessed July 28, 2021).

Kassambara, A., and Mundt, F. (2020). Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Available at: https://cran.r-project.org/package=factoextra (Accessed January 28, 2020).

Kato, S., Sayama, T., Taguchi-Shiobara, F., Kikuchi, A., Ishimoto, M., and Cober, E. (2019). Effect of change from a determinate to a semi-determinate growth habit on the yield and lodging resistance of soybeans in the northeast region of Japan. *Breed. Sci.* 69, 151–159. doi: 10.1270/jsbbs.18112

Kaur, R., Bhardwaj, R., Sharma, R., Kapoor, D., Kohli, S., Kumar, V., et al. (2016). "Hormonal regulation of drought stress in plants," in *Water Stress and Crop Plants: A Sustainable Approach. Vol. 2.* ed. P. Ahmad (Hoboken, NJ: John Wiley & Sons Ltd), 582–599.

Khan, M. M. H., Rafii, M. Y., Ramlee, S. I., Jusoh, M., and Mamun, A. (2020). Genetic variability, heritability, and clustering pattern exploration of Bambara

groundnut (*Vigna subterranea* L. Verdc) accessions for the perfection of yield and yield-related traits. *Biomed Res. Int.* 2020:2195797. doi: 10.1155/2020/2195797

King, C., and Purcell, L. (2005). Inhibition of N2 fixation in soybean is associated with elevated ureides and amino acids. *Plant Physiol.* 137, 1389–1396. doi: 10.1104/pp.104.056317

Kranz, W. L., and Specht, J. E. (2012). Irrigating soybean. NebGuide G1367. Available at: http://ianrpubs.unl.edu/epublic/live/g1367/build/g1367.pdf (Accessed March 5, 2021).

Kumagai, E., and Takahashi, T. (2020). Soybean (*Glycine max* (L.) Merr.) yield reduction due to late sowing as a function of radiation interception and use in a cool region of northern Japan. *Agronomy* 10:66. doi: 10.3390/agronomy10010066

Kumar, M., Govindasamy, V., Rane, J., Singh, A. K., Choudhary, R. L., Raina, S. K., et al. (2017). Canopy temperature depression (CTD) and canopy greenness associated with variation in seed yield of soybean genotypes grown in semi-arid environment. *S. Afr. J. Bot.* 113, 230–238. doi: 10.1016/j.sajb.2017.08.016

Kunert, K., and Vorster, B. J. (2020). In search for drought-tolerant soybean: is the slow-wilting phenotype more than just a curiosity? *J. Exp. Bot.* 71, 457–460. doi: 10.1093/JXB/ERZ235

Kuswantoro, H. (2019). Performance, similarity and genetic parameters of agronomical characters of soybean [*Glycine max* (L) Merrill.] germplasms. *Agric. Nat. Resour.* 53, 228–236. doi: 10.34044/j.anres.2019.53.3.03

Li, M., Liu, Y., Tao, Y., Xu, C., Li, X., Zhang, X., et al. (2019). Identification of genetic loci and candidate genes related to soybean flowering through genome wide association study. *BMC Genomics* 20, 987–913. doi: 10.1186/S12864-019-6324-7

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., and Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* 237:111599. doi: 10.1016/j.rse.2019.111599

Mak, M., Babla, M., Xu, S. C., O'Carrigan, A., Liu, X. H., Gong, Y. M., et al. (2014). Leaf mesophyll K+, H+ and Ca2+ fluxes are involved in drought-induced decrease in photosynthesis and stomatal closure in soybean. *Environ. Exp. Bot.* 98, 1–12. doi: 10.1016/j.envexpbot.2013.10.003

Mangena, P. (2018). "Water stress: morphological and anatomical changes in soybean (*Glycine max* L.) plants," in *Plant, Abiotic Stress and Responses to Climate Change*. ed. V. Andjelkovic (London: IntechOpen), 9–31.

Metz, G. L., Green, D. E., and Shibles, R. M. (1985). Reproductive duration and date of maturity in populations of three wide soybean Crosses1. *Crop Sci.* 25, 171–176. doi: 10.2135/cropsci1985.0011183X002500010041x

Pannecoucque, J., Goormachtigh, S., Ceusters, J., Debode, J., Van Waes, C., and Van Waes, J. (2018). Temperature as a key factor for successful inoculation of soybean with Bradyrhizobium spp. under cool growing conditions in Belgium. *J. Agric. Sci.* 156, 493–503. doi: 10.1017/S0021859618000515

Peters, W., Bastos, A., Ciais, P., and Vermeulen, A. (2020). A historical, geographical and ecological perspective on the 2018 European summer drought. *Philos. Trans. R. Soc. B* 375:20190505. doi: 10.1098/rstb.2019.0505

Pirasteh-Anosheh, H., Saed-Moucheshi, A., Pakniyat, H., and Pessarakli, M. (2016). "Stomatal responses to drought stress," in *Water Stress and Crop Plants: A Sustainable Approach*. *Vol. 1*. ed. P. Ahmad (Hoboken, NJ: John Wiley and Sons Ltd), 24–40.

Ranđelović, P., Đorđević, V., Milić, S., Balešević-Tubić, S., Petrović, K., Miladinović, J., et al. (2020). Prediction of soybean plant density using a machine learning model and vegetation indices extracted from RGB images taken with a UAV. *Agronomy* 10:1108. doi: 10.3390/agronomy10081108

Ray, J. D., Dhanapal, A. P., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Genome-wide association study of ureide concentration in diverse maturity group IV soybean [*Glycine max* (L.) Merr.] accessions. *G3* 5, 2391–2403. doi: 10.1534/g3.115.021774

R Core Team (2019). *RStudio: Integrated Development for R*. Boston, MA: RStudio PBC.

Ries, L. L., Purcell, L. C., Carter, T. E., Edwards, J. T., and King, C. A. (2012). Physiological traits contributing to differential canopy wilting in soybean under drought. *Crop Sci.* 52, 272–281. doi: 10.2135/cropsci2011.05.0278

Rivero, R., Kojima, M., Gepstein, A., Sakakibara, H., Mittler, R., Gepstein, S., et al. (2007). Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19631–19636. doi: 10.1073/PNAS.0709453104

Rosenzweig, V. E., Goloenko, D. V., and Davydenko, O. G. (2003). Breeding strategies for early soybeans in Belarus. *Plant Breed.* 122, 456–458. doi: 10.1046/j.1439-0523.2003.00874.x

Rossi, R. (2019). Irrigation in EU Agriculture. European Parliamentary Research Service. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/644216/EPRS_BRI(2019)644216_EN.pdf (Accessed June 7, 2021).

Saleem, A., Muylle, H., Aper, J., Ruttink, T., Wang, J., Yu, D., et al. (2021). A genome-wide genetic diversity scan reveals multiple signatures of selection in a European soybean collection compared to Chinese collections of wild and cultivated soybean accessions. *Front. Plant Sci.* 12:631767. doi: 10.3389/fpls.2021.631767

Salem, M. A., Kakani, V. G., Koti, S., and Reddy, K. R. (2007). Pollen-based screening of soybean genotypes for high temperatures. *Crop Sci.* 47, 219–231. doi: 10.2135/cropsci2006.07.0443

Saryoko, A., Homma, K., Lubis, I., and Shiraiwa, T. (2017). Plant development and yield components under a tropical environment in soybean cultivars with temperate and tropical origins. *Plant Prod. Sci.* 20, 375–383. doi: 10.1080/1343943X.2017.1356203

Schori, A., Charles, R., and Peter, D. (2003). Soja: se'lection, agronomie et production en Suisse. *Revue Suisse d'Agriculture* 35, 69–76.

Sehgal, A., Sita, K., Kumar, J., Kumar, S., Singh, S., Siddique, K. H. M., et al. (2017). Effects of drought, heat and their interaction on the growth, yield and photosynthetic function of lentil (*Lens culinaris* Medikus) genotypes varying in heat and drought sensitivity. *Front. Plant Sci.* 8:1776. doi: 10.3389/fpls.2017.01776

Setiyono, T. D., Weiss, A., Specht, J., Bastidas, A. M., Cassman, K. G., and Dobermann, A. (2007). Understanding and modeling the effect of temperature and daylength on soybean phenology under high-yield conditions. *Field Crop. Res.* 100, 257–271. doi: 10.1016/j.fcr.2006.07.011

Song, L., Jin, J., and He, J. (2019). Effects of severe water stress on maize growth processes in the field. *Sustainability* 11:618. doi: 10.3390/su11185086

Sun, Y., Wang, C., Chen, H. Y. H., and Ruan, H. (2020). Response of plants to water stress: a meta-analysis. *Front. Plant Sci.* 11:978. doi: 10.3389/fpls.2020.00978

Tacker, P., and Vories, E. (2000). "Irrigation," in *Arkansas Soybean Production Handbook (MP197)* (Fayetteville, AR: University of Arkansas Division of Agriculture Extension Service), 42–49.

Tardieu, F. (2012). Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. *J. Exp. Bot.* 63, 25–31. doi: 10.1093/JXB/ERR269

Tardieu, F., and Simonneau, T. (1998). Variability among species of stomatal control under fluctuating soil water status and evaporative demand: modelling isohydric and anisohydric behaviours. *J. Exp. Bot.* 49, 419–432. doi: 10.1093/jxb/49.Special_Issue.419

Tekelioğlu, B., Büyüktaş, D., Baştuğ, R., Karaca, C., Aydinşakir, K., and Dinç, N. (2017). Use of crop water stress index for irrigation scheduling of soybean in Mediterranean conditions. *J. Exp. Agric. Int.* 18, 1–8. doi: 10.9734/JEAI/2017/37058

Trevisan, R., Pérez, O., Schmitz, N., Diers, B., and Martin, N. (2020). High-throughput phenotyping of soybean maturity using time series UAV imagery and convolutional neural networks. *Remote Sens.* 12:3617. doi: 10.3390/rs12213617

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, PA: Addison-Wesley.

Turner, N. C. (2018). Turgor maintenance by osmotic adjustment: 40 years of progress. *J. Exp. Bot.* 69, 3223–3233. doi: 10.1093/jxb/ery181

UNDRR (2021). United Nations Office for Disaster Risk Reduction. GAR Special Report on Drought 2021, Geneva. Available at: https://www.undrr.org/publication/gar-special-report-drought-2021 (Accessed September 8, 2021).

Unkovich, M., Herridge, D., Peoples, M., Cadisch, G., Boddey, R., Giller, K., et al. (2008). *Measuring Plant-Associated Nitrogen Fixation in Agricultural Systems*. Canberra, ACT: ACIAR MN136.

Van Schoote, M. (2012). Breeding of food grade soybeans. master's thesis. Universiteit Gent Belgium, Available at: https://libstore.ugent.be/fulltxt/RUG01/001/894/331/RUG01-001894331_2012_0001_AC.pdf (Accessed April 1, 2020).

Wang, Z., Li, G., Sun, H., Ma, L., Guo, Y., Zhao, Z., et al. (2018). Effects of drought stress on photosynthesis and photosynthetic electron transport chain in young apple tree leaves. *Biol. Open* 7:bio035279. doi: 10.1242/bio.035279

Wang, C., Linderholm, H. W., Song, Y., Wang, F., Liu, Y., Tian, J., et al. (2020a). Impacts of drought on maize and soybean production in Northeast China during the past five decades. *Int. J. Environ. Res. Public Health* 17:2459. doi: 10.3390/IJERPH17072459

Wang, Z., Reddy, V. R., and Quebedeaux, B. (1997). Growth and photosynthetic responses of soybean to short-term cold temperature. *Environ. Exp. Bot.* 37, 13–24. doi: 10.1016/S0098-8472(96)01033-7

Wang, W., Zhou, B., He, J., Zhao, J., Liu, C., Chen, X., et al. (2020b). Comprehensive identification of drought tolerance QTL-allele and candidate gene systems in Chinese cultivated soybean population. *Int. J. Mol. Sci.* 21:4830. doi: 10.3390/IJMS21144830

Wei, Y., Jin, J., Jiang, S., Ning, S., and Liu, L. (2018). Quantitative response of soybean development and yield to drought stress during different growth stages in the Huaibei plain, China. *Agronomy* 8:97. doi: 10.3390/agronomy8070097

Würschum, T., Leiser, W. L., Jähne, F., Bachteler, K., Miersch, M., and Hahn, V. (2019). The soybean experiment '1000 gardens': a case study of citizen science for research, education, and beyond. *Theor. Appl. Genet.* 132, 617–626. doi: 10.1007/s00122-018-3134-2

Xavier, A., and Rainey, K. M. (2020). Quantitative genomic dissection of soybean yield components. *G3* 10, 665–675. doi: 10.1534/G3.119.400896

Yan, C., Song, S., Wang, W., Wang, C., Li, H., Wang, F., et al. (2020). Screening diverse soybean genotypes for drought tolerance by membership function value based on multiple traits and drought-tolerant coefficient of yield. *BMC Plant Biol.* 20, 321–315. doi: 10.1186/S12870-020-02519-9

Ye, H., Roorkiwal, M., Valliyodan, B., Zhou, L., Chen, P., Varshney, R. K., et al. (2018). Genetic diversity of root system architecture in response to drought stress in grain legumes. *J. Exp. Bot.* 69, 3267–3277. doi: 10.1093/jxb/ery082

Ye, H., Song, L., Schapaugh, W. T., Ali, M. L., Sinclair, T. R., Riar, M. K., et al. (2020). The importance of slow canopy wilting in drought tolerance in soybean. *J. Exp. Bot.* 71, 642–652. doi: 10.1093/jxb/erz150

Yin, X., Goudriaan, J., Lantinga, E., Vos, J., and Spiertz, H. (2003). A flexible sigmoid function of determinate growth. *Ann. Bot.* 91, 361–371. doi: 10.1093/AOB/MCG029

Yu, N., Li, L., Schmitz, N., Tian, L. F., Greenberg, J. A., and Diers, B. W. (2016). Development of methods to improve soybean yield estimation and predict plant maturity with an unmanned aerial vehicle based platform. *Remote Sens. Environ.* 187, 91–101. doi: 10.1016/j.rse.2016.10.005

Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., et al. (2017). Retrieving soybean leaf area index from unmanned aerial vehicle Hyperspectral remote sensing: analysis of RF, ANN, and SVM regression models. *Remote Sens.* 9:309. doi: 10.3390/rs9040309

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16, 217–211. doi: 10.1186/S12864-015-1441-4

Zhang, D., Wang, X., Li, S., Wang, C., Gosney, M. J., Mickelbart, M. V., et al. (2019a). A post-domestication mutation, Dt2, triggers systemic modification of divergent and convergent pathways modulating multiple agronomic traits in soybean. *Mol. Plant* 12, 1366–1382. doi: 10.1016/J.MOLP.2019.05.010

Zhang, J., Zhang, S., Cheng, M., Jiang, H., Zhang, X., Peng, C., et al. (2018). Effect of drought on agronomic traits of rice and wheat: a meta-analysis. *Int. J. Environ. Res. Public Health* 15:839. doi: 10.3390/ijerph15050839

Zhang, X., Zhao, J., Yang, G., Liu, J., Cao, J., Li, C., et al. (2019b). Establishment of plot-yield prediction models in soybean breeding programs using UAV-based hyperspectral remote sensing. *Remote Sens.* 11:2752. doi: 10.3390/RS11232752

Zhao, W., Liu, L., Shen, Q., Yang, J., Han, X., Tian, F., et al. (2020). Effects of water stress on photosynthesis, yield, and water use efficiency in winter wheat. *Water* 12:2127. doi: 10.3390/w12082127

Zhou, J., Yungbluth, D., Vong, C. N., Scaboo, A., and Zhou, J. (2019). Estimation of the maturity date of soybean breeding lines using UAV-based multispectral imagery. *Remote Sens.* 11:2075. doi: 10.3390/rs11182075

Zhou, J., Zhou, J., Ye, H., Ali, M. L., Chen, P., and Nguyen, H. T. (2021). Yield estimation of soybean breeding lines under drought stress using unmanned aerial vehicle-based imagery and convolutional neural network. *Biosyst. Eng.* 204, 90–103. doi: 10.1016/j.biosystemseng.2021.01.017

Zhou, J., Zhou, J., Ye, H., Ali, M. L., Nguyen, H. T., and Chen, P. (2020). Classification of soybean leaf wilting due to drought stress using UAV-based imagery. *Comput. Electron. Agric.* 175:105576. doi: 10.1016/j.compag.2020.105576

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 Saleem, Aper, Muylle, Borra-Serrano, Quataert, Lootens, De Swaef and Roldán-Ruiz. 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 | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 February 2023
DOI 10.3389/fpls.2023.1091875

Check for updates

OPEN ACCESS

EDITED BY
Leif Skot,
Aberystwyth University, United Kingdom

REVIEWED BY
Mohammed Ali Abd Elhammed Abd Allah,
Desert Research Center, Egypt
Fouad Maalouf,
International Center for Agricultural Research in the Dry Areas, Lebanon

*CORRESPONDENCE
Natalia Gutiérrez
natalia.gutierrez.leiva@juntadeandalucia.es

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 07 November 2022
ACCEPTED 16 January 2023
PUBLISHED 01 February 2023

CITATION
Gutiérrez N, Pégard M, Balko C and Torres AM (2023) Genome-wide association analysis for drought tolerance and associated traits in faba bean (*Vicia faba* L.). *Front. Plant Sci.* 14:1091875. doi: 10.3389/fpls.2023.1091875

COPYRIGHT
© 2023 Gutiérrez, Pégard, Balko and Torres. 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.

Genome-wide association analysis for drought tolerance and associated traits in faba bean (*Vicia faba* L.)

Natalia Gutiérrez[1]*, Marie Pégard[2], Christiane Balko[3] and Ana M. Torres[1]

[1]Área de Mejora y Biotecnología, Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA), Centro Alameda del Obispo, Córdoba, Spain, [2]INRAE P3F, 86600 Lusignan, France, INRA, Centre Nouvelle-Aquitaine-Poitiers, Lusignan, France, [3]Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Sanitz, Germany

Faba bean (*Vicia faba* L.) is an important high protein legume adapted to diverse climatic conditions with multiple benefits for the overall sustainability of the cropping systems. Plant-based protein demand is being expanded and faba bean is a good candidate to cover this need. However, the crop is very sensitive to abiotic stresses, especially drought, which severely affects faba bean yield and development worldwide. Therefore, identifying genes associated with drought stress tolerance is a major challenge in faba bean breeding. Although the faba bean response to drought stress has been widely studied, the molecular approaches to improve drought tolerance in this crop are still limited. Here we built on recent genomic advances such as the development of the first high-density SNP genotyping array, to conduct a genome-wide association study (GWAS) using thousands of genetic polymorphisms throughout the entire faba bean genome. A worldwide collection of 100 faba bean accessions was grown under control and drought conditions and 10 morphological, phenological and physiological traits were evaluated to identify single nucleotide polymorphism (SNP) markers associated with drought tolerance. We identified 29 SNP markers significantly correlated with these traits under drought stress conditions. The flanking sequences were blasted to the *Medicago truncatula* reference genomes in order to annotate potential candidate genes underlying the causal variants. Three of the SNPs for chlorophyll content after the stress, correspond to uncharacterized proteins indicating the presence of novel genes associated with drought tolerance in faba bean. The significance of stress-inducible signal transducers provides valuable information on the possible mechanisms underlying the faba bean response to drought stress, thus providing a foundation for future marker-assisted breeding in the crop.

KEYWORDS

drought stress, GWAS, genetic diversity, faba bean, PCA, heritability, SNPs markers, candidate genes

Introduction

Grain legumes are among the most important sources of high-protein for food and feed worldwide and represent key crops for sustainable, low-input, and diverse farming systems. In crop rotations, legumes enhance soil fertility through biological nitrogen fixation and break disease cycles, thus reducing the input of chemicals in agriculture (Nemecek et al., 2008). With one of the highest protein contents and a balanced amino acid profile, faba bean (*V. faba* L.) is the sixth global temperate legume in production (5.7 Million tonnes in 2.7 Mhas), after, chickpea, pea and lentil, with the average yield largely surpassing all of these crops (FAOSTAT, 2020). Faba bean can adapt to a variety of climatic and soil conditions, providing an advantage over other legume crops. Despite these benefits, faba beans still have a limited use in modern agriculture, mainly due to yield instability caused by biotic and abiotic (mainly drought) stresses. In the Mediterranean region, grain legumes are typically grown in rainfed agricultural systems and therefore yield is often variable or low due to the terminal droughts that characterise these areas (Amede et al., 2003; Khan et al., 2007). As a result of climate change, droughts are predicted to increase both in frequency and intensity, further hampering acceptance and wider use of faba beans by farmers in this area as well as in Northern Europe.

Its high sensitivity to drought stress from seedling to maturity (Khan et al., 2007; Khan et al., 2010) prevents faba bean from expressing its full performance potential. A meta-analysis synthesizing the results of field studies and drought experiments across the globe along 34 years revealed a reduction of 40% in faba bean yield following a 65% decrease in water availability, the highest drought-induced yield reduction among the surveyed legume crops (Daryanto et al., 2015). Thus, identifying drought-tolerant faba bean genotypes and developing efficient molecular breeding approaches is crucial to mitigate the devastating impact of drought stress.

A wide genetic variation has been reported in faba bean accessions for various traits related to drought adaptation (Muktadir et al., 2020). In other legume crops, selection for drought resistance based on highly heritable secondary traits, together with physiological attributes such as accumulation of proline or soluble sugars, has proven highly successful (Lafitte et al., 2003; Richards, 2006; Stoddard et al., 2006; Annicchiarico and Iannucci, 2008; Alderfasi and Alghamdi, 2010; Ammar et al., 2015; Balko et al., 2023, submitted). Although the faba bean response to drought stress has been widely studied (Ricciardi et al., 2001; Ammar et al., 2015; Siddiqui et al., 2015), few molecular approaches have been taken to improve drought tolerance in this crop. Khazaei et al. (2014) first reported quantitative trait loci (QTLs) for stomatal characteristics located on chromosome II and exploited the synteny with the model legume species *M. truncatula* to identify candidate genes within these QTLs. Subsequently, Abid et al. (2015) identified six putative drought stress responsive genes in faba bean by suppression subtractive hybridization. More recently, Ali et al. (2016) published the first and so far only drought genome-wide association analysis (GWAS) focusing on a set of 189 German winter faba bean lines derived from 11 parental founders, by assessing a number of physiological aspects related with dehydration tolerance established in previous studies (Balko et al., 1995; Lafitte et al., 2003; Balko, 2005; Stoddard et al., 2006; Richards, 2006; Annicchiarico and Iannucci, 2008; Alderfasi and Alghamdi, 2010; Ammar et al., 2015; Balko et al., 2023, submitted). Using 175 single-nucleotide polymorphisms (SNPs) and 1147 amplified fragment length polymorphisms (AFLPs), several QTLs were detected but the relatively low number of markers used and the very low linkage disequilibrium (LD) detected, limited the success of this association analysis.

In general, traits that contribute to drought tolerance in plants are quantitative and involve multiple genes. Therefore, it is crucial to exploit new genomic resources for the improvement of this crop. Recent advances in next-generation sequencing (NGS) and high-throughput genotyping have allowed the development of new faba bean genomic tools and resources, including mining of SNPs from transcriptome data (Kaur et al., 2014; Ocaña et al., 2015; Webb et al., 2016) and the development of the first high-density SNP genotyping array (O'Sullivan et al., 2019). These resources allow us to conduct genome-wide association studies (GWAS) using thousands of genetic polymorphisms distributed throughout the entire genome. GWAS aims at identifying genetic markers that are strongly associated with QTLs by using the LD between the genetic marker and the causal mutation. Compared with linkage mapping, GWAS provides higher allelic diversity at the corresponding loci and exploits ancestral recombination events in a population, leading to a better association between the marker and the phenotype of a target trait (Zhu et al., 2008).

In recent years, GWAS studies have been conducted in many plant species to dissect complex quantitative traits related to drought tolerance (Hoyos-Villegas et al., 2017; Dossa et al., 2019; Dramadri et al., 2021; Ravelombola et al., 2021; Choudhary et al., 2022). As stated above, a single GWAS study on drought tolerance has been conducted so far in faba bean (Ali et al., 2016), whose results were limited by the low LD and number of markers. The objectives of the present study were: (1) to evaluate the drought tolerance index in faba bean of ten morphological, phenological and physiological traits, (2) to conduct GWAS to identify SNP markers associated with the drought tolerance indices; (3) to investigate the potential relationship between significant loci associated with the drought tolerance indices.

Materials and methods

Plant material

A panel of 100 faba bean accessions from different countries in Africa (8 accessions), North and South America (2), Asia (27) and Europe (39) were used in the study. The original country of the remaining 24 ICARDA accessions is unknown. Europe with 9 countries is the most represented geographical area in the panel, followed by Asia, Africa and America (7, 4 and 2 countries, respectively). Spain is the country accounting for the highest number of accessions (23). The panel includes genetic stocks, landraces and breeding lines aiming at gathering a wide range of genetic diversity from diverse geographical origins. The drought panel was made in collaboration with four public institutes: ICARDA, IFAPA, INIA and INRA, holding the following genebanks SYR002,

ESP046, ESP004, FRA043, respectively. Prior to the genotyping analysis, all the Spanish lines had been selfed for at least four generations. The remaining accessions were purified for two generations by single seed descent (SSD) in insect-proof cages. A detailed description of the collection is provided in Supplementary Table 1.

Phenotypic data analysis

Phenotyping

The 100 faba bean panel was assayed in 2019 and 2020 at the Julius Kühn-Institut (JKI) in Groß Lüsewitz, Germany (54.0701 N 12.33874 E), in a slightly loamy sand soil with pH: 5,7 (Balko et al., 2023 submitted). Field management essentially followed normal local faba bean cropping practices. Plants were sown in a randomized block design with four replications under irrigated (control) and drought stress conditions created under rain-out shelters (two blocks in each shelter). The accessions were grown in single row plots of 1.2 m length with 14 plants each and a row-to-row distance of 0.5 m. Drip irrigation was scheduled in the range of 60 - 70% of field capacity of the soil, determined over winter after excessive rainfall (Balko et al., 2023, submitted). Water content in the soil was assessed by Time Domain Reflectometry (TDR) probes in about 40 cm depth. Drought stress was initiated when about 30% of the plots started flowering. Irrigation in the stress treatment was stopped and during occurring rainfall the shelters were moved over the respective plots.

Six morphological and phenological traits were recorded: maturity date (MAT), defined as the date when more than 90% of the pods have ripened; plant height in cm (PH); number of pods per plant (PP); number of seeds per plant (SP); hundred seed weight (HSW) in grams; and plot yield (PY) in kg. Moreover, four physiological traits were assessed in leaves: free proline content (PRO) (Bates et al., 1973); total content of soluble sugars (TSS) (Yemm and Willis, 1954); and chlorophyll content (SPAD1 and SPAD2). SPAD measurements were performed with a Chlorophyll Meter SPAD 502 plus (Konica Minolta) at the beginning of the stress treatment (SPAD1) and 4 weeks after the onset of stress (SPAD2). Leaf samples for determination of free proline and total content of soluble sugars were taken in the same time window and flash frozen in liquid nitrogen (Balko et al., 2023 submitted). The measurements of these traits were performed by selecting ten individual plants in the middle of the row for each accession.

Adjustment of phenotypic data

All phenotypic traits were independently adjusted for field micro-enviromental heterogeneity using the 'breedR' package (Muñoz and Sanchez, 2022). Phenotypes were combined and adjusted by years. In the model, the genomic estimated breeding values (GEBV) for each trait were determined with the genomic best linear unbiased prediction based model (GBLUP) (Whittaker et al., 2000; Meuwissen et al., 2001; Cantet et al., 2005). Within trials, a random effect was fitted thanks to the use of the tensor product of two bi-splines bases with a covariance structure for the random knot effects (RKE) to account for spatial variability along the row (r) and the column (c) of the field design (Cantet et al., 2005; Cappa and Cantet, 2007; Cappa et al., 2015) to capture the spatial heterogeneity at the plot level. The following model was used:

$$y = \mu + Zu + Ws + \epsilon$$

where y is the raw phenotype, μ the global mean, u the vector of random additive effects following $N(0,G\sigma^2_a)$ with σ^2_a the additive variance and G the relationship matrix, s the vector of random spatial effects containing the parameters of the B-splines tensor product following $N(0,S\sigma^2_s)$ with σ^2_s the variance of the RKE for row and column and S the covariance structure in two dimension, ϵ the vector of residual effects following $N(0,I\ \sigma^2_e)$ with σ^2_e the residual variance. The design matrix Z, and W are indicator matrices relating the plot to the random effects. The method used to obtain the relationship matrix is detailed in the following section. Bi-splines were anchored at a given number of knots for rows and columns, a higher number of knots smooths out the surfaces. 'breedR' optimized the knot numbers by an automated grid search based on the Akaike Information Criterion (AIC). The micro-environmental individual effect was subtracted from the observed phenotype to obtain a spatially adjusted individual phenotype. A genotypic mean of the spatially adjusted phenotypes was calculated for each trait and used for the GWAS. All measurements were tested for deviations from normality by a randomized Q-Q plot.

Genomic data analysis

Genotyping

Young leaves from individual plants were collected, ground in liquid nitrogen and total genomic DNA was isolated using the DNeasy Plant Mini Kit (QIAGEN Ltd, UK). DNA quality was checked by agarose gel electrophoresis and concentration was estimated using the QubitTM dsDNA BR Assay Kit (Invitrogen by ThermoFisher Scientific, UK) following the manufacturer's instructions.

For genotyping we used the Vfaba_v2 Axiom SNP array with 50K SNP (O'Sullivan et al., 2019; Khazaei et al., 2021). Seven of the 100 accessions showing poor DNA quality, as well as SNP markers with a call rate below 97% and a minor allele frequency (MAF) below 95% were excluded from the analysis. After quality control, a matrix consisting of 21,915 SNPs and 93 accessions with 0.89% missing data was kept for further GWAS analysis. The imputation of the missing data was performed by mean allelic frequency. To obtain the genetic position of the significant SNPs we used the information and protocols provided by Skovbjerg et al., 2022. Thus, 17,403 out of the 21,915 SNPs markers could be assigned to genomic positions. The extremely large size of chromosome 1 (> 3 Gbp) generated problems with various softwares and therefore the chromosome was split at the centromere to form Chr1S and Chr1L. In addition, to verify and complete the faba bean chromosomal positions, the SNPs flanking sequences were aligned to the *V. faba* reference genomee[1] (Jayakodi et al., 2022) using the Geneious v.7.1.9genomee[2].

1 https://projects.au.dk/fabagenome/genomics-data

2 https://www.geneious.com

Genomic relationship matrix

The genomic relationship matrix (GRM) was constructed based on VanRaden, 2008, where the matrix Z was calculated as (M - P). M is a matrix of minor allele counts (0, 1, 2 for the reference, heterozygote and alternative, respectively) with *m* column (one for each marker) and *n* rows (one for each accession). P is a matrix which contains the allele frequency, expressed as a difference from 0.5 and multiplied by 2, such that column *i* of P is $2(p_i\text{-}0.5)$. Subtraction of P from M gives Z, which sets mean values of the allele effects to 0. Genomic relationship matrix G was obtained for the first method proposed by VanRaden:

$$G = ZZ'/2\sum pi(1 - pi)$$

Genetic structure

To estimate the number of distinct genetic clusters (K) and admixture existing in the faba bean panel a bayesian based clustering analysis was performed using FastSTRUCTURE v 1.0 (Raj et al., 2014). FastSTRUCTURE was run on default settings with 10-fold cross validation on the 100 accessions testing for subpopulations (K) values ranging from 1 to 10. The most likely K number was chosen by plotting the marginal likelihood of each model as a function of K and determining when the graph begins to plateau. Accessions with membership probabilities ≥ 0.50 were considered to belong to the same group. The choice of K was further supported by applying a discriminant analysis of principal components (DAPC) based procedure for clustering using the 'fviz_pca' function in the 'factoextra' R-package (Kassambara and Mundt, 2020).

Correlation and broad sense heritability estimates

To understand the extent of the relationship among the traits, the correlation matrix for control and stress values was made by Pearson correlation analysis. Descriptive analysis and correlations were conducted in the R statistical software. The broad sense heritability ($\boldsymbol{h^2}$) for the traits was estimated using the following formula:

$$h2 = Vg/(Vg + Vsp + Vres)$$

where $\boldsymbol{V_g}$ is the genetic variance component, $\boldsymbol{V_{sp}}$ is the spatial variance component, $\boldsymbol{V_{res}}$ is the residual variance component. The genotypic mean value for each accession for each trait under control and stress conditions were represented by mean PCA biplot. PCA was performed in the R software package 'prcomp' and visualized with the 'fviz_pca' function.

Genome-wide association analysis

Association analyses were performed in 93 faba bean accessions with 21,915 high-quality SNPs. A Multi Locus Mixed Model method (MLMM) (Segura et al., 2012) was implemented in the R package 'mlmm.gwas' (Bonnafous et al., 2019) to evaluate the trait-SNP associations. The MLMM, is an iterative approach that improves power over single locus methods by incorporating multiple markers in the model simultaneously as covariates, to reduce the false-positive rates and to increase the detection power. In each step (maximum 10 steps), the variance components are estimated and then used to calculate p-values for the association of each SNP with the trait of interest. MLMM utilizes eBIC (extended Bayesian Information Criterion) to determine the number of steps and therefore the number of QTLs with a lambda value of 0.77. The Bonferroni threshold was used to label an association as significant. Significant markers were visualized with a Manhattan plot and important p-value distributions (expected vs. observed p-values on a -log10 scale) were shown with a quantile–quantile (Q-Q) plot.

Potential candidate gene

The sequences flanking associated SNPs were blasted against the NCBI *M. truncatula* reference genome[3] to annotate potential candidate genes underlying the causal variants. Gene locations were determined using the Genome Data Viewer (GDV)[4]. In addition, the sequences flanking associated SNPs were blasted against the faba bean reference genome (Jayakodi et al., 2022) to verify chromosomal positions and locate those candidates that did not show significant hits in *Medicago*. For some of these, the use of the corresponding faba bean contig allowed to infer the Medicago ortholog and include the corresponding gene annotation.

Results

Phenotypic variation and heritability

The ten morphological, phenological and physiological traits listed above were used to examine the possible existence of significant phenotypic variances among the 93 faba bean accessions, both in control and drought conditions. Descriptive statistics revealed large phenotypic variations for all the traits studied (Table 1). For MAT, PH, PP, SP, HSW, PY and SPAD2 the mean values in the drought stress treatment were lower than in the control condition. In contrast, PRO, TSS highly increased under drought stress while in SPAD1 the mean increase was smaller.

The frequency distributions of all 10 traits fit the normal distributions, indicating their quantitative nature (Figure 1). The coefficient of variation (CV%) for most of the traits was comparable for control and drought stress. CV ranged from 5.37 (MAT) to 57.01 (SP) under control condition and from 5.16 (MAT) to 38.71 (PY) under drought stress. Narrow-sense heritability (h^2) estimates ranged from 0.28 (TSS) to 0.75 (SPAD1) in the control and from 0.21 (PRO) to 0.78 (HSW) in drought stress. Heritabilities calculated for each trait were moderate to high for most of the traits, varying from 0.51 to 0.75 in control conditions and from 0.52 to 0.78 in drought stress. Slightly lower values were recorded for PY and PRO in both conditions. Except for TSS and in both treatments, similar estimates for

3 www.blast.ncbi.nlm.nih.gov/Blast.cgi

4 www.ncbi.nlm.nih.gov/genome/gdv/

10.3389/fpls.2023.1091875

TABLE 1 Statistical analysis of 10 morphological, phenological and physiological traits in controlled and drought stress conditions.

Traits	Description	Treatment	Mean	Min	Max	Range	SD	CV (%)	h^2
MAT	Maturity date (days)	Control	127.54	114.00	148.00	34.00	6.85	5.37	0.51
		Stress	115.08	101.00	132.00	31.00	5.93	5.16	0.52
PH	Plant height/cm	Control	64.59	22.67	127.60	104.93	15.35	23.76	0.56
		Stress	56.63	31.90	84.50	52.60	8.21	14.50	0.52
PP	Number of pods per plant	Control	10.84	2.63	36.25	33.62	5.26	48.53	0.61
		Stress	6.41	1.88	15.00	13.13	2.25	35.05	0.62
SP	Number of seeds per plant	Control	24.90	3.75	107.10	103.35	14.20	57.01	0.62
		Stress	14.20	3.13	36.20	33.08	5.44	38.27	0.57
HSW	100 seed weight/grams	Control	67.85	19.59	144.23	124.64	22.62	33.34	0.66
		Stress	60.88	17.22	123.46	106.24	16.47	27.06	0.78
PY	Plot yield/kg	Control	0.16	0.02	0.44	0.42	0.07	46.75	0.42
		Stress	0.08	0.02	0.24	0.23	0.03	38.71	0.32
PRO	Free proline content/µmol g−1	Control	2.22	0.71	14.75	14.04	0.85	38.10	0.30
		Stress	6.40	1.00	73.99	72.99	8.31	28.37	0.21
TSS	Total content of soluble sugars/µmol g−1	Control	1119.91	332.70	2602.00	2269.30	336.37	30.36	0.28
		Stress	1391.26	523.00	2886.95	2363.95	394.73	28.37	0.53
SPAD1	Chlorophyll content, beginning of stress	Control	37.62	24.80	48.00	23.20	4.05	10.77	0.75
		Stress	39.40	24.80	50.60	25.80	3.97	10.09	0.71
SPAD2	Chlorophyll content, 4 weeks after stress	Control	42.79	13.50	61.60	48.10	7.14	16.69	0.62
		Stress	23.70	9.40	56.40	47.00	8.61	36.33	0.69

heritability were detected. Under drought stress a lower coefficient of variation was observed in most traits with exception of (Table 1).

Correlation of traits

To understand the relationship among the traits, we performed a correlation matrix for control and stress values using the Pearson correlation method. Under control conditions (Figure 1), significant positive correlations were observed among most of the traits, with correlation coefficient (r) values ranging from 0.10 to 0.92. By contrast, SPAD1 and PRO were negatively correlated with TSS (-0.09, -0.13, respectively), HSW was negatively correlated with SPAD1 and PRO (-0.10, -0.12, respectively) and PH, PP and SP with HSW (-0.31, -0.36 and -0.39, respectively). TSS revealed a close to neutral correlation with most of the physiological, morphological and phenological traits except MAT (0.30) and SPAD2 (-0.31), where highly significant positive and negative associations, respectively, were observed (Figure 1).

Similarly, under drought stress conditions most of the phenological and yield related traits (MAT, PH, PP, SP, HSW and PY), were strongly associated, showing positive correlations with Pearson's correlation coefficients ranging between 0.12 and 0.85 (Figure 2). PP and SP showed a significant negative correlation with HSW (-0.32 and -0.34, respectively). SPAD1 maintained a neutral or significantly negative correlation with all the traits while SPAD2 was positively correlated with all of the characters except with HSW (-0.19). SPAD1, PRO and TSS showed negative or close to neutral correlation among them and with the rest of yield related traits. The only exceptions were the significant positive correlation of PRO with PH, PP, SPAD2 and TSS (0.14, 0.10, 0.13, 0.14), and the positive association of TSS with MAT (0.22).

Correlations in control and drought treatments provide useful information on the effect of the physiological parameters (SPAD, PRO and TSS) on the yield related traits studied. In both conditions, SPAD2 (the main indicator for drought stress induced leaf senescence), was significantly correlated with PP, SP and PY, with higher correlations observed under drought stress with PP and SP (0.32 and 0.37, respectively). Lower, but still significant correlations were also detected between PP, SP and PY with PRO in control conditions, while under drought stress only a slight correlation between PP and PRO (0.10) was detected. TSS was not significantly correlated with any of the yield related traits in control conditions whereas in stress conditions, highly significant negative correlations with PP, SP and PY (-0.14, -0.21 and -0.12) were observed.

Genetic structure

To examine divergence of the faba bean collection during evolution, a Bayesian based clustering analysis was performed using FastSTRUCTURE and the 21,915 selected SNPs. According to the K genetic clusters, the most likely number of inferred members was three

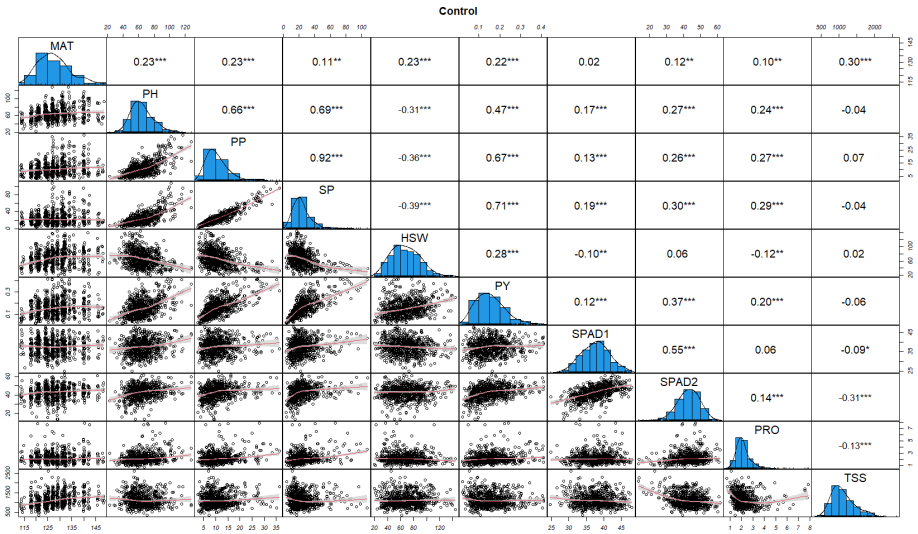


FIGURE 1
Distributions of phenotypic frequency and correlations between 10 morphological, phenological and physiological traits in control conditions. The frequency distribution of each trait is shown on a central diagonal in the form of a histogram. Scatter plots of correlations between every pair of traits are shown in the areas below the diagonal, and numerical Pearson's correlation coefficients (r), between every pair of traits are shown in the areas above the diagonal. The red line in the scatter plots represents the slope of the correlations. The x- and y- axes are the values of the measurements (PH in cm, HSW in grams and PY in kg). *, ** and *** indicate significance at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively.

with $K \geq 0.50$. Besides, we performed PCA using the first two principal components, PC1 (variance explained, 5.3%) and PC2 (variance explained, 3.3%), which are divided into three groups with slight degrees of introgression between them during cultivation (Figure 3). Clade P1 comprised only oriental accessions (10) from China, Nepal and Japan. Clade P2 was the most numerous and consisted of 73 accessions with a wide range of geographic origins spread over four continents: Europe (25), Africa (7), Asia (18) and South America (1), while the remaining 22 accessions are of unknown origin. The third clade (P3) mainly consisted of European accessions (7) together with one from Canada and another from Egypt. Accession EUC_VF_192 was admixed (Supplementary Table 1).

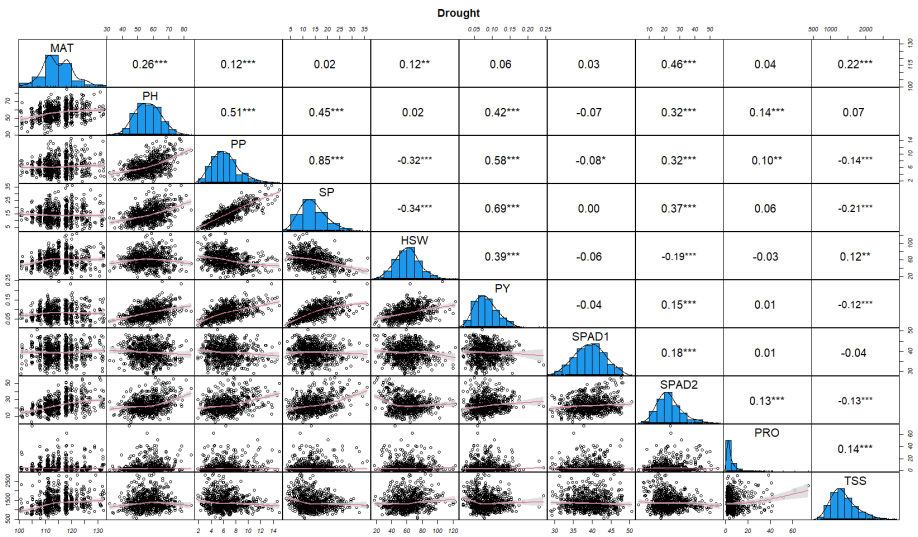


FIGURE 2
Distributions of phenotypic frequency and correlations between 10 morphological, phenological and physiological traits in drought stress conditions. The frequency distribution of each trait is shown on a central diagonal in the form of a histogram. Scatter plots of correlations between every pair of traits are shown in the areas below the diagonal, and numerical Pearson's correlation coefficients (r), between every pair of traits are shown in the areas above the diagonal. The red line in the scatter plots represents the slope of the correlations. The x- and y- axes are the values of the measurements (PH in cm, HSW in grams and PY in kg). *, ** and *** indicate significance at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively.

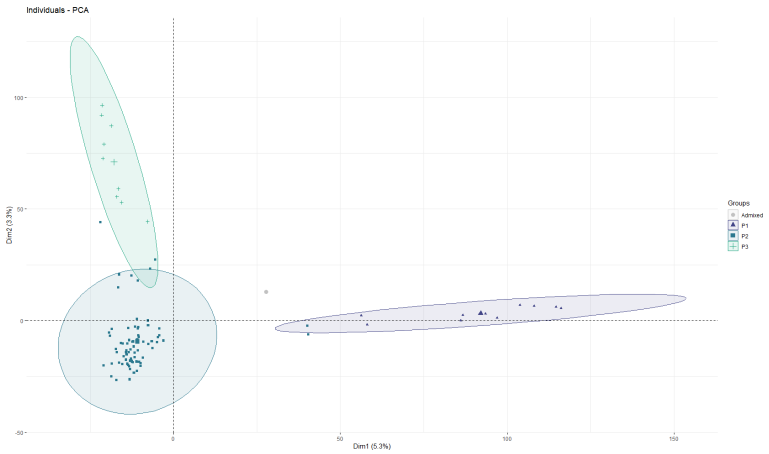


FIGURE 3
Principal component analysis (PCA) of the 93 faba bean accessions. Each dot represents an accession. The horizontal and vertical coordinates represent the first two principal components of analysis (PC1 and PC2), accounting for 5.3% and 3.3% of the total variation, respectively.

Principal component analysis biplot

The contribution of the various traits to the overall variation in the dataset was investigated by PCA (Figure 4). The first three principal components (PCs) with eigenvalues > 1 accounted for 74.4% of the total variation (Supplementary Figure 1). Since the first two PCs showed the highest percentage of variance (62.8%), the PCA biplot was constructed only with PC1 and PC2, showing a clear separation of control vs. drought stress data points along the main axis (Figure 4). PC1 explained 49.1% of the total variability among traits or individuals and was mostly associated with SP, PP, PH, PY, SPAD2 and MAT (Figure 4). PC2 accounted for an additional 13.7% of the total variability among traits and appeared to be related with HSW PRO and TSS (Figure 3). PC3, PC4 and PC5 explained only 11.66%, 9.5% and 7.6%, respectively, of the phenotypic variation (Supplementary Figure 1). PC2 was highly associated with HSW, PRO and TSS. PC3 was strongly associated with SPAD1 and PC with PRO (Supplementary Figure 2). The biplot vectors showed that the morphological, phenological and yield-related traits (PP, SP, PH, PY, MAT) show a strong positive correlation between each other and with the physiological trait SPAD2, while HSW has a less strong correlation and instead shows a strong negative correlation with PRO and SPAD1 (Figure 4).

Genome-wide association mapping

To investigate genetic variants governing drought tolerance in faba bean, 10 morphological, phenological and physiological traits (MAT, PH, PP, SP, HSW, PY, SPAD1, SPAD2, PRO and TSS) were subjected to GWAS analysis using 21,915 SNPs. A total of 74 marker trait associations (MTAs) were identified, revealing candidate loci for

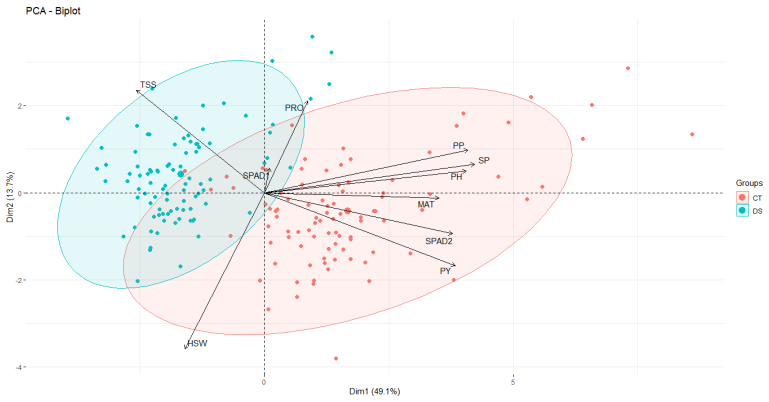


FIGURE 4
PCA biplot showing the clustering of 93 faba bean accessions grown under control and drought stress conditions based on the variance in 10 morpho-physiological and biochemical traits. The traits are maturity date (MAT), plant height (PH), number of pods per plant (PP), number of seeds per plant (SP), 100 seed weight (HSW) and plot yield (PY), free proline content (PRO), total content of soluble sugars (TSS) chlorophyll content (SPAD1 and SPAD2). The first two components explained 49.3% and 18% of the variances, respectively. The magnitude of the vectors (arrows) shows the strength of their contribution to each PC. Vectors pointing in similar directions indicate positively correlated variables, vectors pointing in opposite directions indicate negatively correlated variables, and vectors at proximately right angles indicate low or no correlation. Colored concentration ellipses (size determined by a 0.95-probability level) show the observations grouped by treatment (control or drought conditions). Individuals on the same side as a given variable should be interpreted as having a high contribution on it.

10.3389/fpls.2023.1091875

TABLE 2 List of candidate genes related to control assay traits.

Trait[a]	Axiom_Vfaba ID[b]	%R2	Chrom_Vf	MT_Ortholog	Gene Annotation	Mt locations[c]
PH	AX-416733403	20.1	VF1L	LOC11420042 - MTR_2g034600	Transcription termination factor MTERF5, chloroplastic	Chr2: 13,305,552 - 13,309,126
	AX-416779256	18.2	VF3	LOC25484925 - MTR_1g094180	Uncharacterized LOC25484925	Chr1: 46,026,865 - 46,030,860
	AX-181204108	8.5	VF2	LOC11434242 - MTR_3g101290	Transcription factor MYB73	Chr3: 50,051,392 - 50,052,915
	AX-416746074	6.9	VF3	LOC25501550 - MTR_8g071050	Uncharacterized LOC25501550	Chr8: 32,595,784 - 32,600,071
	AX-416824745	4.7	VF2	LOC11418507 - MTR_3g093440	SET and MYND domain-containing protein 4	Chr3: 46,160,188 - 46,169,983
	AX-416725335	4.5	VF3	LOC11409185 - MTR_1g017770	U-box domain-containing protein 30	Chr1: 5,096,589 - 5,098,303
	AX-181165425	4.1	VF4	LOC25493784 - MTR_4g107280	Serine/threonine-protein kinase AtPK2/AtPK19	Chr4: 52,305,859 - 52,309, 642
	AX-181460581	3.5	VF5	LOC11429781 - MTR_7g113830	Transcription factor bHLH96	Chr7: 53,954,046 - 53,955,883
PP	*AX-416730999*	22.1	VF2	–	No significant similarity found	–
	AX-416814849	15.3	VF1S	LOC11411992- MTR_2g081930	Organic cation/carnitine transporter 4	Chr2: 40,519,275 - 40,525,105
	AX-416811873	9.8	VF2	–	No significant similarity found	–
	AX-416736864	9.6	VF3	LOC11433286 - MTR_1g072570	Protein HUA2-LIKE 3	Chr1: 36,125,447 - 36,149,302
	AX-416742185	8.0	VF2	LOC11418689 - MTR_4g007220	Probable fucosyltransferase 7	Chr4: 1,030,501 - 1,033,327
	AX-181168375	0.1	VF5	LOC11436945 - MTR_7g078070	Bifunctional L-3-cyanoalanine synthase/cysteine synthase 1, mitochondrial	Chr7: 36,224,960 - 36,229,173
	AX-416733741	0.1	VF1L	LOC11436596 - MTR_5g090550	TLC domain-containing protein 4	Chr5: 40,621,005 - 40,626,857
	AX-416746789	-0.1	VF2	LOC11406652 - MTR_3g065480	Uncharacterized LOC11406652	Chr3: 33,106,129 - 33,112,502
SP	*AX-416730999*	37.2	VF2	–	No significant similarity found	–
	AX-181196762	13.8	VF1L	LOC11407110 - MTR_5g072620	Phenylacetaldehyde reductase	Chr5: 32,215,050 - 32,219,289
	AX-416803016	12.0	VF2	LOC11417256 - MTR_3g073530	RHOMBOID-like protein 8	Chr3: 36,651,502 - 36,655,433
	AX-181481926	10.0	VF1S	LOC25495903 - MTR_6g033275	Uncharacterized LOC25495903	Chr6: 11,097,750 - 11,104,911
	AX-416766053	6.4	VF3	LOC11424057 - MTR_1g056550	Syntaxin-121	Chr1: 28,618,423 - 28,622,050
	AX-416809778	5.1	VF2	LOC25487330 - MTR_2g079050	Proline-rich extensin-like protein EPR1	Chr2: 39,201,706 - 39,205,770
HSW	**AX-181155165**	21.3	VF2	LOC11414533 - MTR_4g014600	Protein TIC 56, chloroplastic	Chr4: 4,407,419 - 4,414,567
	AX-181178618	17.3	VF2	LOC11423990 - MTR_3g096160	BTB/POZ domain-containing protein At2g13690	Chr3: 47,404,439 - 47,407,638
	AX-416734460	8.9	VF1S	LOC11433077 - MTR_2g103380	Uncharacterized protein At1g15400	Chr2: 50,770,172 - 50,771,135
	AX-416741157	2.3	VF1L	LOC11430458 - MTR_5g083560	Cation/H(+) antiporter 14	Chr5: 37,260,367 - 37,261,590

(Continued)

10.3389/fpls.2023.1091875

TABLE 2 Continued

Trait[a]	Axiom_Vfaba ID[b]	%R2	Chrom_Vf	MT_Ortholog	Gene Annotation	Mt locations[c]
	AX-181483294	1.9	VF5	LOC11435913 - MTR_7g086510	Cationic amino acid transporter 4, vacuolar	Chr7: 40,354,298 - 40,361,570
	AX-416771643	0.9	VF1L	LOC11426550 - MTR_5g082490	Protein yippee-like	Chr5: 36,649,198 - 36,652,967
	AX-416761080	0.8	VF6	LOC25492837 - MTR_4g074390	Pentatricopeptide repeat-containing protein At1g09820	Chr4: 36,315,927 - 36,318,984
PY	AX-181486157	29.2	VF2	LOC11438062 - MTR_3g062440	Pathogenesis-related genes transcriptional activator PTI6	Chr3: 31,638,312 - 31, 639,244
	AX-416726542	12.3	VF4	LOC25493803 - MTR_4g107500	ATPase 11, plasma membrane-type	Chr4: 52,517,283 - 52,523,566
	AX-181457993	12.0	VF6	LOC11438539 - MTR_8g093850	Uncharacterized LOC11438539	Chr8
	AX-416803843	9.6	VF1L	LOC11439965 - MTR_5g008680	Peptidyl-prolyl cis-trans isomerase FKBP53	Chr5: 1,881,201 - 1,886,862
	AX-416794085	8.9	VF1S	LOC25479984 - MTR_0049s0070	Protein NSP-INTERACTING KINASE 2	MtrunA 17r5.0-ANR-Scaffold
	AX-181496354	8.7	VF4	LOC25483593 - MTR_1g054710	rRNA biogenesis protein RRP5	Chr1: 27,658,013 - 27,684,808
	AX-416740528	3.4	VF3	LOC25484291 - MTR_1g070140	RING-H2 finger protein ATL16	Chr1: 34,725,951 - 34,727,735
	AX-181155156	-0.2	VF4	LOC11438091 - MTR_4g108270	Probable serine/threonine-protein kinase PBL3	Chr4: 52,898,547 - 52,903,260
SPAD1	AX-416773777	15.3	VF1S	LOC11406053 - MTR_2g086780	OVARIAN TUMOR DOMAIN-containing deubiquitinating enzyme 6	Chr2: 42,569,722 - 42,577,518
	AX-416813816	12.5	VF3	LOC25483937 - MTR_1g062760	Ras-related protein RABC2a, mRNA	Chr1: 31,429,552 - 31,432,831
	AX-416806007	10.4	VF1S	LOC25495703 - MTR_6g022710	Probable carboxylesterase 11	Chr1: 8,147,008 - 8,153,086
	AX-416770296	9.4	VF5	LOC11434746 - MTR_7g10068	Probable DNA helicase MCM8	Chr7: 47,402,118 - 47,411,018
	AX-181469161	8.7	VF6	LOC11415976 - MTR_4g068190	Peptidyl-prolyl cis-trans isomerase FKBP16-3, chloroplastic, mRNA	Chr4: 33,706,200 - 33,709,953
	AX-181162616	3.3	VF4	LOC11408450 - MTR_8g058330	Protein transport protein Sec61 subunit alpha	Chr8: 23,516,578 - 23,520,693
	AX-181157586	3.4	VF2	LOC11438520 - MTR_3g085280	Uncharacterized LOC11438520	Chr3: 41,983,181 - 41,987,055
	AX-416766853	2.2	VF4	LOC11439881 - MTR_8g035560	Receptor-like serine/threonine-protein kinase At2g45590	Chr8: 13,391,605 - 13,394,516
PRO	AX-416786927	22.3	VF5	LOC11423151 - MTR_7g112460	Topless-related protein 3	Chr7: 53,265,254 - 53,275,452
	AX-181473167	21.7	VF5	LOC11433065 - MTR_7g104890	Signal recognition particle 54 kDa protein, chloroplastic	Chr7: 49,538,278 - 49,547,049
	AX-181187546	17.7	VF5	LOC11435620 - MTR_7g086430	Pobable zinc metalloprotease EGY1, chloroplastic	Chr7: 40,311,949 - 40,318,392
	AX-181153857	9.5	VF1L	LOC11415575 - MTR_5g070860	AT2G18410-like protein mRNA, partial cds	Chr5: 31,253,236 - 31-258,745
	AX-416734875	3.9	VF1L	LOC11429004 - MTR_5g013970	Phosphoacetylglucosamine mutase, transcript variant X2, mRNA	Chr5: 4,614,512 - 4,620,938
	AX-416766779	0.9	VF4	LOC11420237 - MTR_4g119900	Probable histone-arginine methyltransferase 1.3	Chr4: 57,819,179 - 57,827,376

(Continued)

TABLE 2 Continued

Trait[a]	Axiom_Vfaba ID[b]	%R2	Chrom_Vf	MT_Ortholog	Gene Annotation	Mt locations[c]
	AX-181173312	0.5	VF2	LOC25489275 - MTR_3g067650	Probable serine/threonine-protein kinase PBL7	Chr3: 33,731,346 - 33,735,869

ID of the associated single nucleotide polymorphisms (SNPs) in the Vicia faba Axiom, percentage of phenotypic variation explained (%R2), location in the faba bean chromosomes and orthologous genes, annotation and location in Medicago.
([a]): Plant height (PH), number of pods per plant (PP), number of seeds per plant (SP), 100 seed weight (HSW), plot yield (PY), chlorophyll content (SPAD1), free proline content (PRO), and total content of soluble sugars (TSS).
([b]): In bold, loci associated with HSW, SPAD1 and PRO in both conditions. In red, significant loci associated to the traits that did not reach the Bonferroni threshold (p) > 5.83. In italics, common loci associated with PP and SP traits.
(c): Gene locations were determined using the Genome Data Viewer (GDV).

each trait across different water regimes (Tables 2 and 3). The manhattan and their corresponding quantile-quantile (Q-Q) plots run with the MLMM method are shown in Figures 5 and 6. Q-Q plots revealed that the -log10 (p-values) for the different traits evaluated under each water regime condition conformed to normal distribution.

Under control conditions we detected 52 significant SNPs spread along the genome, although 17 did not reach the Bonferroni threshold -log10(p) > 5.83. Most of the significant markers, however, clustered in chromosome 1 and 2 (Table 2). Eight loci were associated with PH, PP, PY and SPAD1, accounting together for 70.5%, 64.9%, 84% and

TABLE 3 List of candidate genes related to drought resistance traits.

Trait[a]	Axiom_Vfaba ID[b]	%R2	Chrom_Vf	MT_Ortholog	Gene Annotation	Mt locations[c]
PH	AX-181182113	15.4	VF5	LOC11435290 - MTR_2g104100	Formamidopyrimidine-DNA glycosylase	Chr2: 51,106,899 - 51,114,875
	AX-416815601	14.9	VF2	LOC11431814 - MTR_3g101520	B3 domain-containing protein At3g19184	Chr3: 50,178,330 - 50,183,223
	AX-416734003	14.6	VF2	LOC25489500 - MTR_3g075100	Telomere length regulation protein TEL2 homolog	Chr3: 37,598,521 - 37,609,730
	AX-181183872	11.7	VF4	LOC25501139 - MTR_8g044260	DExH-box ATP-dependent RNA helicase DExH3	Chr8:16,827,348 - 16,850,409
	AX-181483303	8.4	VF0	LOC11434981 - MTR_1g045510	28 kDa ribonucleoprotein, chloroplastic	Chr1: 17,459,568 - 17,462,374
	AX-181167806	5.9	VF6	LOC25494322 - MTR_4g133952	Nuclear transcription factor Y subunit B-3	Chr4:64,266,289 - 64,269,332
	AX-416758364	4.9	VF3	LOC11432831 - MTR_1g009620	Exocyst complex component EXO70I	Chr1: 1,426,882 - 1,430,478
	AX-181481274	2.8	VF1L	LOC11408589 - MTR_5g024350	Glutamate receptor 3.6	Chr5: 9,819,043 - 9,825,020
PP	**AX-416814849**	20.6	VF1S	LOC11411992 - MTR_2g081930	Organic cation/carnitine transporter 4	Chr2: 40,519,275 - 40,525,105
	AX-416774842	12.4	VF4	LOC11445678 - MTR_8g022980	AT-rich interactive domain-containing protein 4	Chr8: 8,135,620 - 8,147,219
	AX-416785709	12.0	VF2	–	No significant similarity found	–
	AX-416750038	11.6	VF1L	LOC25486287- MTR_2g029340	Protein ROOT HAIR DEFECTIVE 3 homolog 2	Chr2:11,013,670 - 11,027,992
	AX-416815627	7.1	VF1S	LOC25496020 - MTR_6g034195	Plastidial pyruvate kinase 2	Chr6: 12,142,869 - 12,151,041
	AX-416817671	3.2	VF5	LOC25499196 - MTR_7g096080	Tryptophan synthase alpha chain	Chr7: 45,376,885 - 45,380,449
	AX-181486430	2.7	VF2	LOC25491358 - MTR_4g011600	Molybdate transporter 2	Chr4: 3,007,294 - 3,009,858
HSW	**AX-181155165**	20.3	VF2	LOC11414533 - MTR_4g014600	Protein TIC 56, chloroplastic	Chr4: 4,407,419 - 4,414,567
	AX-416774496	14.5	VF2	LOC11422606 - MTR_3g070390	Nuclear pore complex protein NUP88	Chr3: 35,050,851 - 35,063,461

(Continued)

10.3389/fpls.2023.1091875

TABLE 3 Continued

Trait[a]	Axiom_Vfaba ID[b]	%R2	Chrom_Vf	MT_Ortholog	Gene Annotation	Mt locations[c]
	AX-181178618	13.4	VF2	LOC11423990 - MTR_3g096160	BTB/POZ domain-containing protein At2g13690	Chr3: 47,404,439 - 47,407,638
	AX-181484267	9.7	VF5	LOC11431692 - MTR_7g052640	Putative lipid-transfer protein DIR1	Chr7: 24,973,213 - 24,973,970
	AX-416801263	6.5	VF3	LOC25495073 - MTR_6g003960	Transcription termination factor MTEF1	Chr6: 32,604 - 41,403
	AX-416740666	3.9	VF4	LOC11446464 - MTR_7g093530	Probable xyloglucan endotransglucosylase/ hydrolase protein 23	Chr7: 44,048,945 - 44,050,479
SPAD2	AX-416771580	26.0	VF1S	LOC11422670 - MTR_2g097800	Uncharacterized LOC11422670	Chr2: 47,890,406 - 47,896,247
	AX-416751689	20.4	VF3	LOC25484157 - MTR_1g069165	CLIP-associated protein	Chr1: 33,665,872 - 33,684,619
	AX-416739735	11.8	VF0	LOC11412171 - MTR_5g021260	Probable ubiquitin-conjugating enzyme E2 16	Chr5: 8,210,511 - 8,216,281
	AX-416800399	10.0	VF1S	LOC11415177 - MTR_6g065110	Probable carboxylesterase 18	Chr1: 31,948,768 - 31,950,002
	AX-416757391	4.0	VF5	LOC11421925 - MTR_1g115950	Uncharacterized LOC11421925	Chr1: 56,096,020 - 56,103,099
	AX-181178687	3.2	VF1L	LOC11411195 - MTR_5g080880	Uncharacterized protein At2g39795, mitochondrial	Chr5: 35,837,184 - 35,842,681
PRO	AX-416814537	30.1	VF2	LOC25490802 - MTR_3g462820	Beta-glucosidase BoGH3B	Chr3: 28,335,676 - 28,340,509,
	AX-181191699	13.0	VF2	LOC11433803 - MTR_3g089510	Mitogen-activated protein kinase 20	Chr3: 20,834,197 - 20, 842,245

ID of the associated single nucleotide polymorphisms (SNPs) in the *Vicia faba* Axiom, percentage of phenotypic variation explained (%R2), location in the faba bean chromosomes and orthologous genes, annotation and location in Medicago.

[a])Plant height (PH), number of pods per plant (PP), 100 seed weight (HSW), chlorophyll content (SPAD2), free proline content (PRO).

([b])In bold, loci associated with HSW in both conditions. In red, significant loci associated to the traits that did not reach the Bonferroni threshold (p) > 5.83.

([c])Gene locations were determined using the Genome Data Viewer (GDV).

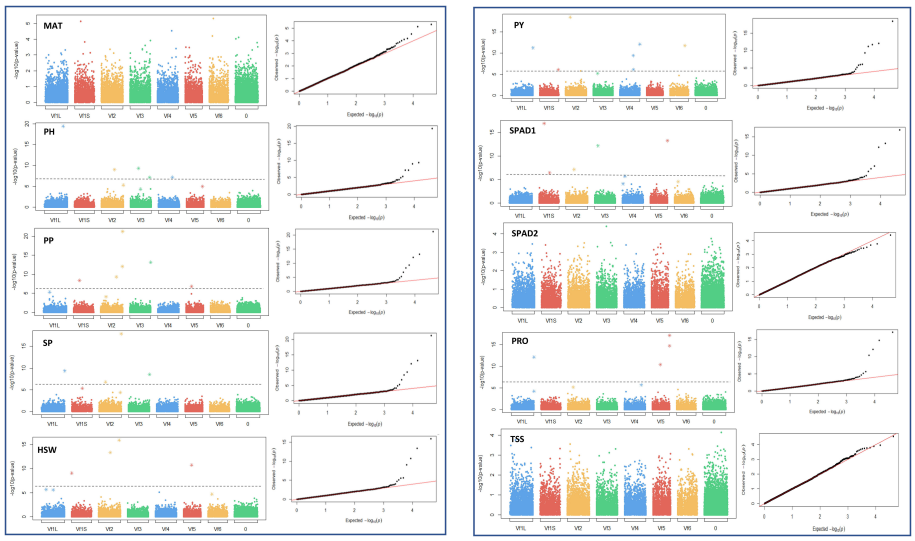


FIGURE 5
Manhattan plots and quantile-quantile (Q-Q) plots of the GWAS results for the 10 traits studied in control conditions. MAT, maturity date; PH, plant height; PP, number of pods per plant; SP, number of seeds per plant; HSW, 100 seed weight; PY, plot yield; SPAD1 and SPAD2, chlorophyll content at the beginning of the stress treatment and about 4 weeks after onset of drought stress, respectively. PRO, free proline content; TSS, total content of soluble sugars. Bonferroni threshold (-log10 (p) > 5.87), is represented by a continuous grey line. X-axis represents the six faba bean chromosomes. The biggest metacentric chromosome I is divided in two corresponding to the large (L) and short (S) arms. Chromosome 0 stands for unknown locations.

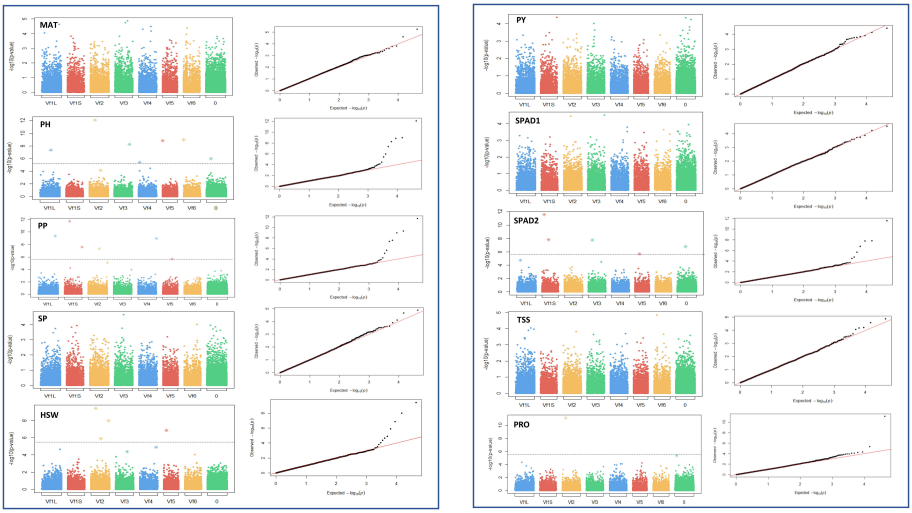


FIGURE 6
Manhattan plots and quantile-quantile (Q-Q) plots of the GWAS results for the 10 traits studied in drought stress conditions. MAT, maturity date; PH, plant height; PP, number of pods per plant; SP, number of seeds per plant; HSW, 100 seed weight; PY, plot yield; SPAD1 and SPAD2, chlorophyll content at the beginning of the stress treatment and about 4 weeks after onset of drought stress, respectively. PRO, free proline content; TSS, total content of soluble sugars. Bonferroni threshold (-log10 (p) > 5.87), is represented by a continuous grey line. X-axis represents the six faba bean chromosomes. The biggest metacentric chromosome I is divided in two corresponding to the large (L) and short (S) arms. Chromosome 0 stands for unknown locations.

65.3% of the respective trait variation. Likewise, seven significant SNPs captured 53.4% of the HSW and 76.4% of the PRO variation. Finally, the six markers associated with SP provided the highest contribution to the phenotypic variance (84.6%). No significant associations were detected for MAT, SPAD2 and TSS. The SNP AX-416730999 has a common association with PP and SP. No such colocalization of SNP markers with multiple traits was observed under drought conditions.

Under drought stress, a total of 29 loci were significantly associated with the traits although 8 did not reach the Bonferroni threshold -log10(p) > 5.83 (Table 3). Although distributed across the six faba bean chromosomes, 16 of them (55%) colocalized mostly in chromosomes 1 and 2 while three of them could not be assigned. Six markers were HSW and SPAD2 associated, jointly explaining 68.4% and 75.4% of the trait variation, respectively. Eight SNPs were associated with PH and seven with PP explaining, respectively, 78.6%, 69.6% of the phenotypic variation. No SNPs associated with MAT, SP and PY were found. The GWAS analysis did not identify significant SNP markers for SPAD1 and TSS, but PRO showed association with two SNPs, one of them explaining the highest percentages of the trait variation in drought conditions (30.1%). Three pleiotropic loci (noted in bold in Tables 2 and 3) were associated with HSW and PP in both water regime conditions.

To further understand the genetic basis of faba bean drought stress-related traits, the sequences flanking SNPs associated with significant traits were subjected to a BLAST search to identify the orthologous sequences in *M. truncatula* or in other model legumes. From the 29 candidate genes (Table 3), 28 were functionally annotated while one of them did not show a significant sequence similarity with *Medicago*. For the sake of brevity, we will mainly comment on the putative genes explaining around 10% of the trait variation.

Starting with the phenological and yield related traits, eight significant SNPs were found for plant height (PH). Four of these genes, annotated as formamidopyrimidine-DNA glycosylase, B3 domain-containing protein At3g19184, Telomere length regulation protein TEL2 and DExH-box ATP-dependent RNA helicase DExH3 (Table 3), accounted for 15.4%, 14.9%, 14.6% and 11.7%, respectively, of the phenotypic variation. Concerning pods per plant (PP), the most significant gene explaining 20.6% of the variation is annotated as organic cation/carnitine transporter 4. Another putative candidate gene explained 12.4% of the trait variation and corresponds to a AT-rich interactive domain-containing protein 4 and a Protein ROOT HAIR DEFECTIVE 3 homolog 2 accounted for 11.6% of the variation. Finally, one uncharacterized SNP explained 12% of the trait. Six putative candidates were associated with hundred seed weight (HSW). The first of these, Protein TIC 56, which contributes to 20.3% of the variation, participates at the inner chloroplast envelope membrane to form a channel for plastid protein import (Kikuchi et al., 2013). Next, a nuclear pore complex protein NUP88, explained 14.5% of the variation, while another gene annotated as a BTB/POZ domain-containing protein explained 13.4% of the trait.

Concerning the physiological traits, six SNPs were significantly associated with SPAD2 and two of them, the uncharacterized LOC11422670 and a CLIP-associated protein, explained the highest percentage of variation (26% and 20.4%, respectively). Besides, a ubiquitin-conjugating enzyme E2 16 together with a probable carboxylesterase 18 were responsible for 11.8 and 10% of the SPAD2 variation. For proline content (PRO), two SNPs annotated as beta-glucosidase BoGH3B and mitogen-activated protein kinase 20 accounted for a significant percentage of the trait variation (30.1% and 13%, respectively).

Discussion

Drought stress represents a major threat to plant growth and development. Since faba bean is generally grown under rainfed conditions, it often experiences water stress at the terminal growth phase of the crop. Considering the present scenario of climate change, enhancing faba bean productivity through improved drought tolerance is a prioritary goal in breeding efforts.

Phenotyping for drought tolerance is costly and time consuming, but secondary characters with high heritability that are correlated with yield under drought conditions can be used for indirect selection (Ziyomo and Bernardo, 2013). In the present study, 100 faba bean accessions from different origins were evaluated for two years under control and drought conditions, for 10 secondary traits associated to phenology, physiology and grain yield. These traits have been used in different legume studies for efficient assessment of stress tolerance (Nadeem et al., 2019; Valdisser et al., 2020; Ravelombola et al., 2021; Wu et al., 2021). Considering both years and conditions, the yield related traits (PP, SP, HSW, PY) were highly correlated with MAT, PH and SPAD2 and could thus be used as efficient secondary traits for drought tolerance in faba bean improvement programs. Accordingly, the vectors among these traits in Figure 4 had small angles confirming the positive correlation. The selection of faba bean accessions from different origins, with sufficient genetic variation and weak population structure revealed a large variation in these traits, suggesting that the panel is genetically diverse and could be advantageous for GWAS implementation.

GWAS has become a critical tool for detecting genetic variants underlying complex traits. The large number of SNPs obtained with the 50K SNP array from Affymetrix (O'Sullivan et al., 2019; Khazaei et al., 2021) has provided an extensive genome coverage to differentiate germplasm accessions and to carry out high-resolution association mapping. Using 21,915 SNPs we detected here a total of 52 significant SNPs in irrigation (control) and 29 in drought conditions, distributed across the six faba bean chromosomes, which collectively explained a high percentage of the total phenotypic variation. Three of these SNPs were associated with the traits evaluated both under control and drought conditions (Tables 2 and 3) and should thus not be relevant to drought stress. In control conditions the SNPs associated with morphological, physiological and yield related traits explained from 53.4% in the case of HSW to 84.6% in SP while under drought stress the R2 values ranged from 43.1% (PRO) to 78.6% (PH). No significantly associated SNPs were detected for MAT, SPAD2 and TSS in control conditions or for MAT, SP, PY, SPAD1 and TSS under stress. Interestingly, two significant SNPs accounting for high percentages of the trait variation in PP and SPAD2, correspond to uncharacterized proteins indicating the presence of novel genes associated with drought tolerance in faba bean.

To progress in our understanding and possible functions of significant genes, we investigated their involvement in water stress-responses reported from other crop species. BLAST search analysis showed that most of the significant SNP markers identified in the present study aligned with candidates, known to be involved in responses to drought stress in different crops (Table 3).

Four main drought stress response candidates were identified for PH. The first corresponds to a formamidopyrimidine-DNA glycosylase reported to initiate base excision repair at damaged sites in response to abiotic stresses (Chen et al., 2012; Wallace, 2014). The expression of many drought-induced genes is regulated at the transcriptional level and this activity can be extended to the second candidate, a B3 domain-containing protein At3g19184 candidate since in maize, the B3 domain-containing transcription factor Viviparous1 (Vp1) was induced by drought stress (Cao et al., 2007). The third major candidate identified in our study is a homolog of the telomere length regulation protein TEL2, a key regulator of cell proliferation and genome maintenance. TEL2 complexes interacts with and promotes protein kinases stability by controlling telomerase length as well as the DNA damage response (Smogorzewska and de Lange, 2004; Smogorzewska and de Lange, 2004; Sugimoto, 2018). For example, a telomere length regulation protein TEL2 homolog in rice was differentially expressed in response to salinity stress (Cotsaftis et al., 2011). Finally, the fourth candidate gene encodes a DExH-box ATP-dependent RNA helicase, which in plants has a critical role in a variety of RNA-mediated regulation of cell proliferation and abiotic stress responses (Liu and Imai, 2018). In addition, four other candidates genes detected in our work encode, respectively, a 28 kDa ribonucleoprotein, a which was recently reported in chickpea response to biotic and abiotic stresses (Vessal et al., 2020); a nuclear Y-B transcription factor that has proven to regulate resistance to drought stress in *Arabidopsis*, maize and soybean (Nelson et al., 2007; Sun et al., 2022); an exocyst complex component reported as a drought and salt tolerance regulator in grapevine (Wang et al., 2023), and a glutamate receptor with a signalling role in responses to abiotic stresses such as salt, cold, heat, and drought, of Arabidopsis, faba bean and rice (Lu et al., 2014; Yoshida et al., 2016; Qiu et al., 2019).

Concerning the number of pods per plant (PP), an organic cation/carnitine transporter 4, an AT-rich interactive domain-containing protein (ARID domain) and a protein ROOT HAIR DEFECTIVE 3 (RHD3) homolog 2 were the most significant annotated genes identified. In *Arabidopsis*, several organic cation transporters were up-regulated during drought stress suggesting a specific role in plant adaptation to environmental stress (Küfner and Koch, 2008). Likewise, the ARID domain containing proteins are transcription factors implicated in a wide variety of roles, including chromatin remodelling, transcription, and cell growth (Wilsker et al., 2002). In a proteomics study of sugarcane response, (Salvato et al., 2019) showed that different types of transcriptional regulators, including ARID domains proteins were differentially accumulated in response to drought stress. RHD3 was also required for regulation of cell expansion and root hair development. Thus, Wong et al. (2018) reported that the ectopic expression of a *Musa acuminata* RHD3 gene enhanced drought tolerance in *Arabidopsis*. Moreover, a root transcriptomic analysis of contrasting *Gossypium herbaceum* genotypes revealed a higher expression of RHD3 genes in tolerant lines, highlighting the key involvement of these genes in root length development and plasticity under drought stress conditions (Ranjan et al., 2012).

Three main QTLs were associated with hundred seed weight (HSW): Protein TIC 56 participates at the inner chloroplast envelope membrane to form a channel for plastid protein import (Kikuchi et al., 2013), a BTB/POZ domain-containing protein with potential roles in developmental programs such as promotion of leaf and floral meristem fate and determinacy, as well as in defence and abiotic stress response (Guan et al., 2018). Both genes (Protein TIC 56 and BTB/

POZ domain) were significant both under control and drought conditions. The next, candidate is the nuclear pore complex (NPC) protein NUP88. Diverse mechanisms have been proposed to explain the role of NPC family components in responses to different stresses such as cold, abscisic acid (ABA), drought, and biotic stress (Yang et al., 2017). Further candidates associated with HSW in drought conditions include the lipid-transfer protein (LTP) DIR1, the transcription termination factor MTEF1 and a xyloglucan endotransglucosylase/hydrolase protein. LTPs are thought to be involved in plant defense responses (Safi et al., 2015) and their expression is induced by biotic and abiotic stresses, including disease, salinity, temperature and drought (Safi et al., 2015; Akhiyarova et al., 2021; Duo et al., 2021; Zhao et al., 2021). It is also well established that drought tolerance is regulated by the mitochondrial transcription termination factors (MTERFs). A recent analysis of *mterf* mutants supports aa role for plant MTERFs in abiotic stress response (Quesada, 2016). Similarly, the xyloglucan endotransglucosylase/hydrolases are inducible by a broad spectrum of abiotic stresses and have been shown to enhanced tolerance to salt and drought stresses in tomato (Choi et al., 2011).

Concerning the physiological trait SPAD2, only three of the six significant SNPs were annotated, indicating the presence of three novel candidates associated with drought tolerance in faba bean. The first of the annotated candidates, a CLIP-associated protein (CLASPs), correspond to an evolutionarily conserved family of regulatory factors that control microtubule dynamics and the organization of microtubule networks. Although little is known about their function in plants, in Arabidopsis, CLASP is involved in both cell division and cell expansion by linking microtubules and auxin transport (Ambrose et al., 2007). The second candidate gene encodes an ubiquitin-conjugating enzyme that has shown to enhance drought and salt tolerance in *Arabidopsis* and melon, as well as in different legume crops such as soybean, peanut or mung bean (Zhou et al., 2010; Baloglu and Patir, 2014; Wang et al., 2016; Chen et al., 2020). Finally, carboxylesterases are known to play important roles in plant growth, development and resistance to stresses (Prinsi et al., 2018; Arisha et al., 2020; Rui et al., 2022).

The two candidates related to proline content (PRO) were the beta-glucosidase BoGH3B and the mitogen-activated protein kinase 20. Plant β-glucosidases are involved in cell wall biogenesis which protects plants against external stresses (Moradi Tarnabi et al., 2020). Different b-glucosidase homologs were shown to be involved in the response to dehydration and NaCl stress in *Arabidopsis* (Xu et al., 2012) and drought stress in soybean roots (Wang et al., 2017). Finally, mitogen-activated protein kinase (MAPK) genes are involved in many cell activities including growth, differentiation and proliferation, as well as environmental stress responses. MAPKs activation is a common defense response of plants to a range of abiotic stressors (Komis et al., 2018; Muhammad et al., 2019; Kim et al., 2021). Because drought stress leads simultaneously to osmotic and oxidative stress (Zhu, 2002), osmotic stress activates several protein kinases including MAPKs, which mediate osmotic homeostasis and/or detoxification responses.

Osmotic adaptation is a major component of drought resistance in different crops (Sánchez et al., 1998; Bajji et al., 2001). Proline and soluble sugar accumulation are common physiological responses in many plants during water-deficit stress, to protect cellular components and to restore the osmotic balance (Chen and Murata, 2002; Guo et al., 2018). Accordingly, PRO and TSS increased under drought stress (Table 1). Severe drought stress also inhibits the photosynthesis of plants by causing a decrease in chlorophyll content (Ommen et al., 1999). Our results reveal that four weeks after the onset of stress, the mean chlorophyll content (SPAD2) was highly reduced mainly due to damage in chloroplasts caused by reactive oxygen species (Smirnoff, 1995).

The wide range of candidates functionally annotated and significantly associated with drought stress component traits evidences that drought responses are complex and that each induction phase may be controlled by different signalling mechanisms and transcription factors (Shinozaki et al., 2003) classified the products of stress-inducible genes identified in microarray experiments into two groups, one includes molecules such as late embryogenesis abundant (LEA) proteins, osmotin, key enzymes for osmolyte biosynthesis, water channel proteins, sugar and proline transporters, and various proteases, while the second group consists of regulators of intracellular signalling and stress-inducible gene expression (e.g. protein kinases such as MAP kinases, phosphatases, phospholipid metabolic enzymes, and various types of transcription factors). A mitogen-activated protein kinase and a plastidial pyruvate kinase were associated with PRO and PP respectively, while two transcription factors were significantly associated with PP and HSW. Receptor kinases are considered as key regulators of plant architecture and growth, but they also function in defence and stress responses (Marshall et al., 2012). In fact, some serine/threonine-protein kinases are known to play a role in signal transduction and were shown to improve drought tolerance in Arabidopsis, rice, soja and bamboo (Xie et al., 2014; Liu et al., 2022). On the other hand, it is well known that transcription factors synchronise signal transduction and expression of drought tolerance regulatory genes, enabling plants to withstand stress conditions (Joshi et al., 2016; Hrmova and Hussain, 2021). For these reasons, they are considered as potential candidates with broad applications in crop breeding. These results show that the approach applied to this faba bean collection could lead to the efficient identification of candidate genes that are relevant to faba bean drought tolerance.

In summary, our study demonstrates the feasibility of GWAS analysis with a diverse germplasm collection and a high-density array chip, for the identification of drought tolerance-related traits in faba bean. Under stress conditions, 29 SNP markers that were significantly correlated to these traits have been identified, mostly clustered in chromosomes 1 and 2. Interestingly, all of them were directly or indirectly involved in responses to drought stress, thus establishing a solid foundation for further research. The identification of a number of stress-inducible signal transducers provides valuable information on the putative faba bean response mechanisms against drought stress. Nevertheless, a validation of the identified markers in a larger size or bi-parental population, using tissue and stage specific gene expression data from RNA-Seq, would be reasonable before embarking on a broad breeding program. The results from this study will contribute to a better understanding of the genetic architecture governing drought tolerance in faba bean and provide a foundation for marker-assisted breeding in this crop.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

AT and NG selected and multiplied the seeds for the experiments and devised the project. CB planned and carried out the drought evaluations and measurements. MP designed and supervised the statistical analysis. NG processed the experimental data and performed the GWAS analysis. AT and NG wrote the manuscript in consultation with MP and CB. All authors contributed to the article and approved the submitted version.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312 (EUCLEG Project) and RTA2020-007 co-financed by ERDF. We thank the seed companies and institutes that provided us with faba bean seeds of their accessions: INRA, Dijon (Jean-Bernard Magnin-Robert), ICARDA (Fouad Maalouf) and IFAPA. The technical staff of IFAPA, centro Alameda del Obispo, is thanked for their deep investment to multiply accessions and collect leaves for genotyping. The authors are grateful to Dr. A. Di Pietro for critical reading of the manuscript.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1091875/full#supplementary-material

References

Abid, G., Muhovski, Y., Mingeot, D., Watillon, B., Toussaint, A., Mergeai, G., et al. (2015). Identification and characterization of drought stress responsive genes in faba bean (Vicia faba l.) by suppression subtractive hybridization. *Plant Cell Tiss. Organ Cult.* 121, 367–379. doi: 10.1007/s11240-014-0707-x

Akhiyarova, G. R., Ivanov, R. S., Ivanov, I. I., Finkina, E. I., Melnikova, D. N., Bogdanov, I. V., et al. (2021). Effects of salinity and abscisic acid on lipid transfer protein accumulation, suberin deposition and hydraulic conductance in pea roots. *Membranes (Basel)* 11, 762. doi: 10.3390/membranes11100762

Alderfasi, A. A., and Alghamdi, S. S. (2010). Integrated water supply with nutrient requirements on growth, photosynthesis productivity, chemical status and seed yield of faba bean. *American-Eurasian J. Agron.* 3, 8–17.

Ali, M. B. M., Welna, G. C., Sallam, A., Martsch, R., Balko, C., Gebser, B., et al. (2016). Association analyses to genetically improve drought and freezing tolerance of faba bean (*Vicia faba* l.). *Crop Sci.* 56, 1036–1048. doi: 10.2135/cropsci2015.08.0503

Ambrose, J. C., Shoji, T., Kotzer, A. M., Pighin, J. A., and Wasteneys, G. O. (2007). The arabidopsis CLASP gene encodes a microtubule-associated protein involved in cell expansion and division. *Plant Cell* 19, 2763–2775. doi: 10.1105/tpc.107.053777

Amede, T., Schubert, S., and Stahr, K. (2003). Mechanisms of drought resistance in grain legumes I: Osmotic adjustment. *SEJS* 26 (1), 37–46. doi: 10.4314/sinet.v26i1.18198

Ammar, M. H., Anwar, F., El-Harty, E. H., Migdadi, H. M., Abdel-Khalik, S. M., Al-Faifi, S. A., et al. (2015). Physiological and yield responses of faba bean (*Vicia faba* l.) to drought stress in managed and open field environments. *J. Agro. Crop Sci.* 201, 280–287. doi: 10.1111/jac.12112

Annicchiarico, P., and Iannucci, A. (2008). Breeding strategy for faba bean in southern Europe based on cultivar responses across climatically contrasting environments. *Crop Sci.* 48, 983. doi: 10.2135/cropsci2007.09.0501

Arisha, M. H., Ahmad, M. Q., Tang, W., Liu, Y., Yan, H., Kou, M., et al. (2020). RNA-Sequencing analysis revealed genes associated drought stress responses of different durations in hexaploid sweet potato. *Sci. Rep.* 10, 12573. doi: 10.1038/s41598-020-69232-3

Bajji, M., Lutts, S., and Kinet, J. M. (2001). Water deficit effects on solute contribution to osmotic adjustment as a function of leaf ageing in three durum wheat (Triticum durum desf.) cultivars performing differently in arid conditions. *Plant Sci.* 160, 669–681. doi: 10.1016/S0168-9452(00)00443-X

Balko, C. (2005). "Physiological parameters of drought tolerance in relation to yield and yield stability in faba beans," in *The 2nd international conference on integrated approaches to sustain and improve plant protection under drought stress* (Rome: INTERDROUGHT-II).

Balko, C., Stelling, D., Larher, F., Sedding, S., Kittlittz, E., and Jürgens, H. U. (1995). *Investigations into selection for drought tolerance in vicia faba l.* Ed. E. Belhassen (Montpelier: ENSA-INRA).

Balko, C., Torres, A. M., and Gutierrez, N. (2003). *Variability in drought stress response in a panel of 100 faba bean genotypes (submitted).*

Baloglu, M. C., and Patir, M. G. (2014). Molecular characterization, 3D model analysis, and expression pattern of the CmUBC gene encoding the melon ubiquitin-conjugating enzyme under drought and salt stress conditions. *Biochem. Genet.* 52, 90–105. doi: 10.1007/s10528-013-9630-9

Bates, L. S., Waldren, R. P., and Teare, I. D. (1973). Rapid determination of free proline for water-stress studies. *Plant Soil* 39, 205–207. doi: 10.1007/BF00018060

Bonnafous, F., Duhnen, A., Gody, L., Guillaume, O., Mangin, B., Pegot-Espagnet, P., et al. (2019). *Mlmm.gwas: Pipeline for GWAS using MLMM.* https://CRAN.Rproject.org/package=mlmm.gwas. r package version 1.0.6.

Cantet, R. J. C., Birchmeier, A. N., Canaza Cayo, A. W., and Fioretti, C. (2005). Semiparametric animal models *via* penalized splines as alternatives to models with contemporary groups. *J. Anim. Sci.* 83, 2482–2494. doi: 10.2527/2005.83112482x

Cao, X., Costa, L. M., Biderre-Petit, C., Kbhaya, B., Dey, N., Perez, P., et al. (2007). Abscisic acid and stress signals induce Viviparous1 expression in seed and vegetative tissues of maize. *Plant Physiol.* 143, 720–731. doi: 10.1104/pp.106.091454

Cappa, E. P., and Cantet, R. J. C. (2007). Bayesian Estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi: 10.1139/X07-116

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. C. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11, 120. doi: 10.1007/s11295-015-0917-3

Chen, H., Chu, P., Zhou, Y., Li, Y., Liu, J., Ding, Y., et al. (2012). Overexpression of AtOGG1, a DNA glycosylase/AP lyase, enhances seed longevity and abiotic stress tolerance in arabidopsis. *J. Exp. Bot.* 63, 4107–4121. doi: 10.1093/jxb/ers093

Chen, T. H. H., and Murata, N. (2002). Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. *Curr. Opin. Plant Biol.* 5, 250–257. doi: 10.1016/S1369-5266(02)00255-8

Chen, K., Tang, W.-S., Zhou, Y.-B., Xu, Z.-S., Chen, J., Ma, Y.-Z., et al. (2020). Overexpression of gmubc9 gene enhances plant drought resistance and affects flowering time *via* histone H2B monoubiquitination. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.555794

Choi, J. Y., Seo, Y. S., Kim, S. J., Kim, W. T., and Shin, J. S. (2011). Constitutive expression of CaXTH3, a hot pepper xyloglucan endotransglucosylase/hydrolase, enhanced tolerance to salt and drought stresses without phenotypic defects in tomato plants (Solanum lycopersicum cv. dotaerang). *Plant Cell Rep.* 30, 867–877. doi: 10.1007/s00299-010-0989-3

Choudhary, S., Isobe, S., and Chahota, R. K. (2022). Elucidation of drought tolerance potential of horsegram (Macrotyloma uniflorum var.) germplasm using genome wide association studies. *Gene* 819, 146241. doi: 10.1016/j.gene.2022.146241

Cotsaftis, O., Plett, D., Johnson, A. A. T., Walia, H., Wilson, C., Ismail, A. M., et al. (2011). Root-specific transcript profiling of contrasting rice genotypes in response to salinity stress. *Mol. Plant* 4, 25–41. doi: 10.1093/mp/ssq056

Daryanto, S., Wang, L., and Jacinthe, P.-A. (2015). Global synthesis of drought effects on food legume production. *PloS One* 10, e0127401. doi: 10.1371/journal.pone.0127401

Dossa, K., Li, D., Zhou, R., Yu, J., Wang, L., Zhang, Y., et al. (2019). The genetic basis of drought tolerance in the high oil crop sesamum indicum. *Plant Biotechnol. J.* 17, 1788–1803. doi: 10.1111/pbi.13100

Dramadri, I. O., Nkalubo, S. T., Kramer, D. M., and Kelly, J. D. (2021). Genome wide association analysis of drought adaptive traits in common bean. *Crop Sci* 61, 3232–3253. doi: 10.1002/csc2.20484

Duo, J., Xiong, H., Wu, X., Li, Y., Si, J., Zhang, C., et al. (2021). Genome-wide identification and expression profile under abiotic stress of the barley non-specific lipid transfer protein gene family and its qingke orthologues. *BMC Genomics* 22, 674. doi: 10.1186/s12864-021-07958-8

FAOSTAT (2020) *FAOSTAT* (FAOSTAT). Available at: http://www.fao.org/faostat/en/#data (Accessed June 17, 2021).

Guan, L., Haider, M. S., Khan, N., Nasim, M., Jiu, S., Fiaz, M., et al. (2018). Transcriptome sequence analysis elaborates a complex defensive mechanism of grapevine (Vitis vinifera l.) in response to salt stress. *Int. J. Mol. Sci.* 19, 4010. doi: 10.3390/ijms19124019

Guo, R., Shi, L., Jiao, Y., Li, M., Zhong, X., Gu, F., et al. (2018). Metabolic responses to drought stress in the tissues of drought-tolerant and drought-sensitive wheat genotype seedlings. *AoB Plants* 10, ply016. doi: 10.1093/aobpla/ply016

Hoyos-Villegas, V., Song, Q., and Kelly, J. D. (2017). Genome-wide association analysis for drought tolerance and associated traits in common bean. *Plant Genome* 10, 1–17. doi: 10.3835/plantgenome2015.12.0122

Hrmova, M., and Hussain, S. S. (2021). Plant transcription factors involved in drought and associated stresses. *Int. J. Mol. Sci.* 22 (11), 5662. doi: 10.3390/ijms22115662

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednar, P., et al. (2022). The giant diploid faba genome unlocks variation in a global protein crop. *BioRxiv*. doi: 10.1101/2022.09.23.509015

Joshi, R., Wani, S. H., Singh, B., Bohra, A., Dar, Z. A., Lone, A. A., et al. (2016). Transcription factors and plants response to drought stress: current understanding and future directions. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01029

Kassambara, A., and Mundt, F. (2020). *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7*. https://CRAN.R-project.org/package=factoextra.

Kaur, S., Kimber, R. B. E., Cogan, N. O. I., Materne, M., Forster, J. W., and Paull, J. G. (2014). SNP discovery and high-density genetic mapping in faba bean (Vicia faba l.) permits identification of QTLs for ascochyta blight resistance. *Plant Sci.* 217–218, 47–55. doi: 10.1016/j.plantsci.2013.11.014

Khan, H., Link, W., Hocking, T. J., and Stoddard, F. L. (2007). Evaluation of physiological traits for improving drought tolerance in faba bean (Vicia faba l.). *Plant Soil* 292, 205–217. doi: 10.1007/s11104-007-9217-5

Khan, H. R., Paull, J. G., Siddique, K. H. M., and Stoddard, F. L. (2010). Faba bean breeding for drought-affected environments: A physiological and agronomic perspective. *Field Crops Res.* 115, 279–286. doi: 10.1016/j.fcr.2009.09.003

Khazaei, H., O'Sullivan, D. M., Sillanpää, M. J., and Stoddard, F. L. (2014). Use of synteny to identify candidate genes underlying QTL controlling stomatal traits in faba bean (Vicia faba l.). *Theor. Appl. Genet.* 127, 2371–2385. doi: 10.1007/s00122-014-2383-y

Khazaei, H., O'Sullivan, D. M., Stoddard, F. L., Adhikari, K. N., Paull, J. G., Schulman, A. H., et al. (2021). Recent advances in faba bean genetic and genomic tools for crop improvement. *Legume Sci.* 3, e75. doi: 10.1002/leg3.75

Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., et al. (2013). Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339, 571–574. doi: 10.1126/science.1229262

Kim, M., Jeong, S., Lim, C. W., and Lee, S. C. (2021). Mitogen-activated protein kinase CaDIMK1 functions as a positive regulator of drought stress response and abscisic acid signaling in capsicum annuum. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.646707

Komis, G., Šamajová, O., Ovečka, M., and Šamaj, J. (2018). Cell and developmental biology of plant mitogen-activated protein kinases. *Annu. Rev. Plant Biol.* 69, 237–265. doi: 10.1146/annurev-arplant-042817-040314

Küfner, I., and Koch, W. (2008). Stress regulated members of the plant organic cation transporter family are localized to the vacuolar membrane. *BMC Res. Notes* 1, 43. doi: 10.1186/1756-0500-1-43

Lafitte, R., Blum, A., and Atlin, G. (2003)Using secundary traits to help identify drought-tolerant genotypes. In: *Breeding rice for drought-prone enviroments* (Phillipines: International Rice Research Institute). Available at: https://www.researchgate.net/publication/259885067_Using_secondary_traits_to_help_identify_drought-tolerant_genotypes (Accessed May 30, 2022).

Liu, Y., and Imai, R. (2018). Function of plant DExD/H-box RNA helicases associated with ribosomal RNA biogenesis. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00125

Liu, R., Vasupalli, N., Hou, D., Stalin, A., Wei, H., Zhang, H., et al (2022). Genome-wide identification and evolution of WNK kinases in Bambusoideae and transcriptional profiling during abiotic stress in Phyllostachys edulis. *PeerJ* 10, e12718. doi: 10.7717/peerj.12718

Lu, G., Wang, X., Liu, J., Yu, K., Gao, Y., Liu, H., et al. (2014). Application of T-DNA activation tagging to identify glutamate receptor-like genes that enhance drought tolerance in plants. *Plant Cell Rep.* 33, 617–631. doi: 10.1007/s00299-014-1586-7

Marshall, A., Aalen, R. B., Audenaert, D., Beeckman, T., Broadley, M. R., Butenko, M. A., et al. (2012). Tackling drought stress: receptor-like kinases present new approaches. *Plant Cell* 24, 2262–2278. doi: 10.1105/tpc.112.096677

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. doi: 10.1093/genetics/157.4.1819

Moradi Tarnabi, Z., Iranbakhsh, A., Mehregan, I., and Ahmadvand, R. (2020). Impact of arbuscular mycorrhizal fungi (AMF) on gene expression of some cell wall and membrane elements of wheat (Triticum aestivum l.) under water deficit using transcriptome analysis. *Physiol. Mol. Biol. Plants* 26, 143–162. doi: 10.1007/s12298-019-00727-8

Muhammad, T., Zhang, J., Ma, Y., Li, Y., Zhang, F., Zhang, Y., et al. (2019). Overexpression of a mitogen-activated protein kinase SlMAPK3 positively regulates tomato tolerance to cadmium and drought stress. *Molecules* 24 (3), 556. doi: 10.3390/molecules24030556

Muktadir, M. A., Adhikari, K. N., Merchant, A., Belachew, K. Y., Vandenberg, A., Stoddard, F. L., et al. (2020). Physiological and biochemical basis of faba bean breeding for drought adaptation–a review. *Agronomy* 10, 1345. doi: 10.3390/agronomy10091345

Muñoz, F., and Sanchez, L. (2022)BreedR. In: *Statistical methods for forest genetic resources analysts*. Available at: https://github.com/famuvie/breedR (Accessed March 9, 2022).

Nadeem, M., Li, J., Yahya, M., Sher, A., Ma, C., Wang, X., et al. (2019). Research progress and perspective on drought stress in legumes: A review. *Int. J. Mol. Sci.* 20 (10), 1–32. doi: 10.3390/ijms20102541

Nelson, D. E., Repetti, P. P., Adams, T. R., Creelman, R. A., Wu, J., Warner, D. C., et al. (2007). Plant nuclear factor y (NF-y) b subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16450–16455. doi: 10.1073/pnas.0707193104

Nemecek, T., von Richthofen, J.-S., Dubois, G., Casta, P., Charles, R., and Pahl, H. (2008). Environmental impacts of introducing grain legumes into European crop rotations. *Eur. J. Agron.* 28, 380–393. doi: 10.1016/j.eja.2007.11.004

Ocaña, S., Seoane, P., Bautista, R., Palomino, C., Claros, G. M., Torres, A. M., et al. (2015). Large-Scale transcriptome analysis in faba bean (Vicia faba l.) under ascochyta fabae infection. *PloS One* 10, e0135143. doi: 10.1371/journal.pone.0135143

Ommen, O. E., Donnelly, A., Vanhoutvin, S., van Oijen, M., and Manderscheid, R. (1999). Chlorophyll content of spring wheat flag leaves grown under elevated CO2 concentrations and other environmental stresses within the 'ESPACE-wheat' project. *Eur J Agron.* 10, 197–203. doi: 10.1016/S1161-0301(99)00011-8

O'Sullivan, D. M., Angra, D., Harvie, T., Tagkouli, V., and Warsame, A. (2019). "A genetic toolbox for vicia faba improvement," in *: International conference on legume genetics and genomics* (Dijon: ICLGG).

Prinsi, B., Negri, A. S., Failla, O., Scienza, A., and Espen, L. (2018). Root proteomic and metabolic analyses reveal specific responses to drought stress in differently tolerant grapevine rootstocks. *BMC Plant Biol.* 18, 126. doi: 10.1186/s12870-018-1343-0

Qiu, X.-M., Sun, Y.-Y., Ye, X.-Y., and Li, Z.-G. (2019). Signaling role of glutamate in plants. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01743

Quesada, V. (2016). The roles of mitochondrial transcription termination factors (MTERFs) in plants. *Physiol. Plant* 157, 389–399. doi: 10.1111/ppl.12416

Raj, A., Stephens, M., and Pritchard, J. K. (2014). astSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ranjan, A., Pandey, N., Lakhwani, D., Dubey, N. K., Pathre, U. V., and Sawant, S. V. (2012). Comparative transcriptomic analysis of roots of contrasting gossypium herbaceum genotypes revealing adaptation to drought. *BMC Genomics* 13, 680. doi: 10.1186/1471-2164-13-680

Ravelombola, W., Shi, A., and Huynh, B.-L. (2021). Loci discovery, network-guided approach, and genomic prediction for drought tolerance index in a multi-parent advanced generation intercross (MAGIC) cowpea population. *Hortic. Res.* 8, 24. doi: 10.1038/s41438-021-00462-w

Ricciardi, L., Polignano, G. B., and De Giovanni, C. (2001). Genotypic response of faba bean to water stress. *Euphytica.* 118, 39–46. doi: 10.1023/a:1004078017159

Richards, R. A. (2006). Physiological traits used in the breeding of new cultivars for water-scarce environments. *Agric. Water Manage.* 80, 197–211. doi: 10.1016/j.agwat.2005.07.013

Rui, C., Peng, F., Fan, Y., Zhang, Y., Zhang, Z., Xu, N., et al. (2022). Genome-wide expression analysis of carboxylesterase (CXE) gene family implies GBCXE49 functional responding to alkaline stress in cotton. *BMC Plant Biol.* 22, 194. doi: 10.1186/s12870-022-03579-9

Safi, H., Saibi, W., Alaoui, M. M., Hmyene, A., Masmoudi, K., Hanin, M., et al. (2015). A wheat lipid transfer protein (TdLTP4) promotes tolerance to abiotic and biotic stress in arabidopsis thaliana. *Plant Physiol. Biochem.* 89, 64–75. doi: 10.1016/j.plaphy.2015.02.008

Salvato, F., Loziuk, P., Kiyota, E., Daneluzzi, G. S., Araújo, P., Muddiman, D. C., et al. (2019). Label-free quantitative proteomics of enriched nuclei from sugarcane (Saccharum ssp) stems in response to drought stress. *Proteomics* 19, e01900004. doi: 10.1002/pmic.201900004

Sánchez, F. J., Manzanares, M., de Andres, E. F., Tenorio, J. L., and Ayerbe, L. (1998). Turgor maintenance, osmotic adjustment and soluble sugar and proline accumulation in 49 pea cultivars in response to water stress. *Field Crops Res.* 59, 225–235. doi: 10.1016/S0378-4290(98)00125-7

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

Shinozaki, K., Yamaguchi-Shinozaki, K., and Seki, M. (2003). Regulatory network of gene expression in the drought and cold stress responses. *Curr. Opin. Plant Biol.* 6, 410–417. doi: 10.1016/S1369-5266(03)00092-X

Siddiqui, M. H., Al-Khaishany, M. Y., Al-Qutami, M. A., Al-Whaibi, M. H., Grover, A., Ali, H. M., et al. (2015). Response of different genotypes of faba bean plant to drought stress. *Int. J. Mol. Sci.* 16, 10214–10227. doi: 10.3390/ijms160510214

Skovbjerg, C. K. K., Angra, D., Robertson-Shersby-Harvie, T., Kreplak, J., Ecke, W., Windhorst, A., et al. (2022). Genetic analysis of global faba bean germplasm maps agronomic traits and identifies strong selection signatures for geographical origin. *BioRxiv*. doi: 10.1101/2022.07.18.500421

Smirnoff, N. (1995). "Antioxidant systems and plant response to the environment," in *Environment and plant metabolism: Flexibility and acclimation*. Ed. V. Smirnoff (Oxford: BIOS Scientific Publishers), 217–243.

Smogorzewska, A., and de Lange, T. (2004). Regulation of telomerase by telomeric proteins. *Annu. Rev. Biochem.* 73, 177–208. doi: 10.1146/annurev.biochem.73.071403.160049

Stoddard, F. L., Balko, C., Erskine, W., Khan, H. R., Link, W., and Sarker, A. (2006). Screening techniques and sources of resistance to abiotic stresses in cool-season food legumes. *Euphytica* 147, 167–186. doi: 10.1007/s10681-006-4723-8

Sugimoto, K. (2018). Branching the Tel2 pathway for exact fit on phosphatidylinositol 3-kinase-related kinases. *Curr. Genet.* 64, 965–970. doi: 10.1007/s00294-018-0817-9

Sun, M., Li, Y., Zheng, J., Wu, D., Li, C., Li, Z., et al. (2022). A nuclear factor y-b transcription factor, GmNFYB17, regulates resistance to drought stress in soybean. *Int. J. Mol. Sci.* 23(13):7242. doi: 10.3390/ijms23137242

Valdisser, P. A. M. R., Müller, B. S. F., de Almeida Filho, J. E., Morais Júnior, O. P., Guimarães, C. M., Borba, T. C. O., et al. (2020). Genome-wide association studies detect multiple QTLs for productivity in mesoamerican diversity panel of common bean under drought stress. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.574674

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

Vessal, S., Arefian, M., and Siddique, K. H. M. (2020). Proteomic responses to progressive dehydration stress in leaves of chickpea seedlings. *BMC Genomics* 21, 523. doi: 10.1186/s12864-020-06930-2

Wallace, S. S. (2014). Base excision repair: a critical player in many games. *DNA Repair (Amst)* 19, 14–26. doi: 10.1016/j.dnarep.2014.03.030

Wang, X., Khodadadi, E., Fakheri, B., and Komatsu, S. (2017). Organ-specific proteomics of soybean seedlings under flooding and drought stresses. *J. Proteomics* 162, 62–72. doi: 10.1016/j.jprot.2017.04.012

Wang, N., Liu, Y., Cong, Y., Wang, T., Zhong, X., Yang, S., et al. (2016). Genome-wide identification of soybean U-box E3 ubiquitin ligases and roles of GmPUB8 in negative regulation of drought stress response in arabidopsis. *Plant Cell Physiol.* 57, 1189–1209. doi: 10.1093/pcp/pcw068

Wang, L., Zhang, X., Tang, Y., Zhao, T., Huang, C., Li, Y., et al. (2023). Exocyst subunit VviExo70B is degraded by ubiquitin ligase VviPUB19 and they regulate drought and salt tolerance in grapevine. *Environ. Exp. Bot.* 206, 105175. doi: 10.1016/j.envexpbot.2022.105175

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., et al. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (Vicia faba l.). *Plant Biotechnol. J.* 14, 177–185. doi: 10.1111/pbi.12371

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genet. Res.* 75, 249–252. doi: 10.1017/S0016672399004462

Wilsker, D., Patsialou, A., Dallas, P. B., and Moran, E. (2002). ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. *Cell Growth Differ.* 13, 95–106.

Wong, G. R., Mazumdar, P., Lau, S.-E., and Harikrishna, J. A. (2018). Ectopic expression of a musa acuminata root hair defective 3 (MaRHD3) in arabidopsis enhances drought tolerance. *J. Plant Physiol.* 231, 219–233. doi: 10.1016/j.jplph.2018.09.018

Wu, X., Sun, T., Xu, W., Sun, Y., Wang, B., Wang, Y., et al. (2021). Unraveling the genetic architecture of two complex, stomata-related drought-responsive traits by high-throughput physiological phenotyping and GWAS in cowpea (Vigna. unguiculata l. walp). *Front. Genet.* 12. doi: 10.3389/fgene.2021.743758

Xie, M., Wu, D., Duan, G., Wang, L., He, R., Li, X., et al (2014). AtWNK9 is regulated by ABA and dehydration and is involved in drought tolerance in Arabidopsis. *Plant Physiol. Biochem.* 77, 73–83. doi: 10.1016/j.plaphy.2014.01.022

Xu, Z.-Y., Lee, K. H., Dong, T., Jeong, J. C., Jin, J. B., Kanno, Y., et al. (2012). A vacuolar β-glucosidase homolog that possesses glucose-conjugated abscisic acid hydrolyzing activity plays an important role in osmotic stress responses in arabidopsis. *Plant Cell* 24, 2184–2199. doi: 10.1105/tpc.112.095935

Yang, Y., Wang, W., Chu, Z., Zhu, J.-K., and Zhang, H. (2017). Roles of nuclear pores and nucleo-cytoplasmic trafficking in plant stress responses. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00574

Yemm, E. W., and Willis, A. J. (1954). The estimation of carbohydrates in plant extracts by anthrone. *Biochem. J.* 57, 508–514. doi: 10.1042/bj0570508

Yoshida, R., Mori, I. C., Kamizono, N., Shichiri, Y., Shimatani, T., Miyata, F., et al. (2016). Glutamate functions in stomatal closure in arabidopsis and fava bean. *J. Plant Res.* 129, 39–49. doi: 10.1007/s10265-015-0757-0

Zhao, J., Bi, W., Zhao, S., Su, J., Li, M., Ma, L., et al. (2021). Wheat apoplast-localized lipid transfer protein TaLTP3 enhances defense responses against puccinia triticina. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.771806

Zhou, G.-A., Chang, R.-Z., and Qiu, L.-J. (2010). Overexpression of soybean ubiquitin-conjugating enzyme gene GmUBC2 confers enhanced drought and salt tolerance through modulating abiotic stress-responsive gene expression in arabidopsis. *Plant Mol. Biol.* 72, 357–367. doi: 10.1007/s11103-009-9575-x

Zhu, J.-K. (2002). Salt and drought stress signal transduction in plants. *Annu. Rev. Plant Biol.* 53, 247–273. doi: 10.1146/annurev.arplant.53.091401.143329

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome J.* 1, 5. doi: 10.3835/plantgenome2008.02.0089

Ziyomo, C., and Bernardo, R. (2013). Drought tolerance in maize: Indirect selection through secondary traits versus genomewide selection. *Crop Sci.* 53, 1269–1275. doi: 10.2135/cropsci2012.11.0651

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/fpls.2023.1128823

Check for updates

OPEN ACCESS

EDITED BY
Fernando Juan Yuste-Lisbona,
University of Almeria, Spain

REVIEWED BY
Lina Ahmed,
INRAE Nouvelle Aquitaine Poitiers, France
Petr Smýkal,
Palacký University in Olomouc, Czechia

*CORRESPONDENCE
Roland Kölliker
roland.koelliker@usys.ethz.ch

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 21 December 2022
ACCEPTED 13 February 2023
PUBLISHED 01 March 2023

CITATION
Nay MM, Grieder C, Frey LA, Amdahl H, Radovic J, Jaluvka L, Palmé A, Skøt L, Ruttink T and Kölliker R (2023) Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover. *Front. Plant Sci.* 14:1128823. doi: 10.3389/fpls.2023.1128823

COPYRIGHT
© 2023 Nay, Grieder, Frey, Amdahl, Radovic, Jaluvka, Palmé, Skøt, Ruttink and Kölliker. 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.

Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover

Michelle M. Nay[1], Christoph Grieder[1], Lea A. Frey[2], Helga Amdahl[3], Jasmina Radovic[4], Libor Jaluvka[5], Anna Palmé[6], Leif Skøt[7], Tom Ruttink[8,9] and Roland Kölliker[2]*

[1]Fodder Plant Breeding, Division of Plant Breeding, Agroscope, Zurich, Switzerland, [2]Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland, [3]Graminor Breeding Ltd., Bjørke Forsøksgård, Norway, [4]Institute for Forage Crops (IKBKS), Kruševac, Serbia, [5]DLF Seeds, Hladké Životice, Czechia, [6]The Nordic Genetic Resource Centre, Plant Section, Alnarp, Sweden, [7]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom, [8]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food, Melle, Belgium, [9]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

Red clover (*Trifolium pratense* L.) is an outcrossing forage legume that has adapted to a wide range of climatic and growing conditions across Europe. Red clover is valued for its high yield potential and its forage quality. The high amount of genetic diversity present in red clover provides an invaluable, but often poorly characterized resource to improve key traits such as yield, quality, and resistance to biotic and abiotic stresses. In this study, we examined the genetic and phenotypic diversity within a diverse set of 395 diploid red clover accessions *via* genome wide allele frequency fingerprinting and multi-location field trials across Europe. We found that the genetic structure of accessions mostly reflected their geographic origin and only few cases were detected, where breeders integrated foreign genetic resources into their local breeding pools. The mean dry matter yield of the first main harvesting season ranged from 0.74 kg m^{-2} in Serbia and Norway to 1.34 kg m^{-2} in Switzerland. Phenotypic performance of accessions in the multi-location field trials revealed a very strong accession x location interaction. Local adaptation was especially prominent in Nordic red clover accessions that showed a distinct adaptation to the growing conditions and cutting regime of the North. The traits vigor, dry matter yield and plant density were negatively correlated between the trial location in Norway and the locations Great Britain, Switzerland, Czech Republic and Serbia. Notably, breeding material and cultivars generally performed well at the location where they were developed. Our results confirmed that red clover cultivars were bred from regional ecotypes and

show a narrow adaptation to regional conditions. Our study can serve as a valuable basis for identifying interesting materials that express the desired characteristics and contribute to the adaptation of red clover to future climatic conditions.

KEYWORDS

red clover, plant breeding, forage, legume, adaptation

1 Introduction

Red clover (*Trifolium pratense* L.) is the most important perennial forage legume in Northern and Central Europe. It is either grown in pure stands or in mixtures with tall grasses. Due to its ability to fix atmospheric nitrogen, red clover produces protein-rich forage and, when grown in mixtures, can provide additional nitrogen to companion grasses (Broderick, 1995; Taylor and Quesenberry, 1996; Halling et al., 2004; Nyfeler et al., 2011). Grass-legume mixtures have the additional benefit of reducing nitrogen leaching, increasing forage quality, and improving drought resilience (Lüscher et al., 2014; Hofer et al., 2016). Red clover is, therefore, an important component of sustainable grassland-based animal production and can substantially contribute to the protein self-sufficiency of Europe.

Red clover is native to Europe, the Middle East, North Africa and Central Asia, and was introduced to most temperate regions in the world as a forage crop (Taylor and Quesenberry, 1996; Boller et al., 2010). The benefits of red clover were readily appreciated by farmers, when they realized that the replacement of fallow by clover led to higher yields and better nutrient availability in the fields (Weir, 1926; Taylor and Quesenberry, 1996). Consequently, farmers started to harvest red clover seeds on farm for sowing in the crop rotation or for improving existing grasslands. This has led to the development of numerous landraces adapted to the specific management regimes and environmental conditions of the sites they emerged from (Taylor, 2008). Many of these landraces survived as naturalized populations, were integrated in breeders' collections, or were conserved in genebanks. Together with other germplasm collections, these landraces laid the foundation for systematic breeding of red clover, which started in the second half of the 20th century and is nowadays conducted in research organizations or small to large multinational companies throughout Europe.

Red clover is a naturally diploid (2n = 2x = 14), insect pollinated, outbreeding species with a high degree of self-incompatibility. Tetraploid red clover accessions have been produced through colchicine treatment, but despite their higher yield, their market success is hampered by low seed yield (Boller et al., 2010). In order to maintain a certain level of genetic diversity to maximize adaptability to a broad range of environmental conditions and due to its self-incompatibility, red clover is usually bred as population-based cultivars. Breeding schemes are mainly based on recurrent phenotypic selection either to directly develop open pollinated cultivars through population-improvement or to select suitable parental plants for mutual intermating (poly-crossing) and the development of synthetic cultivars (Posselt, 2010). The base populations used for breeding may consist of ecotype or wild populations, landraces, breeding material as well as commercial cultivars and are continuously complemented to ensure sufficient diversity in target traits. Specific breeding goals may differ between individual programs, but general breeding targets include high biomass and seed yield, persistence over the desired cultivation period and disease resistance (Abberton and Marshall, 2005; Taylor, 2008; Riday, 2010). In the northern regions of Europe, particular emphasis is put on improving freezing tolerance and resistance to clover rot, two factors strongly associated with persistence. Clover rot, caused by the ascomycete *Sclerotinia trifoliorum* Eriks., is favored by cooler temperatures, high humidity and long snow cover, and can lead to severe overwintering damage (Saharan and Mehta, 2008). Freezing tolerance is associated with winter survival and is particularly pronounced in wild populations from northern regions (Zanotto et al., 2021). In the warmer climates of Central Europe, persistence is an important breeding target with the specific aim to extend the period of cultivation for red clover cultivars over several growing seasons. Based on landraces developed through decades of on-farm seed production, a particular type of red clover ('Mattenklee') evolved and a number of highly persistent and high yielding cultivars were developed from this type (Kölliker et al., 2003; Boller et al., 2010). In the same region, resistance to southern anthracnose has gained importance as a breeding target. The disease is caused by *Colletotrichum trifolii* Bain & Essary and has benefitted from warmer summer temperatures, which has made it a limiting factor of red clover production in the warmer regions of Europe (Boller et al., 2010). The disease has long been recognized as a major threat in Southern USA, where targeted selection resulted in cultivars largely resistant to southern anthracnose (Taylor, 2008). North American red clover accessions are also often distinguished by their characteristic hairiness that is thought to aid in preventing leafhopper damage (Pieters and Hollowell, 1937). To account for different utilization requirements, different growth types have been developed. Early flowering clovers are capable of vigorous regrowth and remain generative throughout the growing period. Late flowering clover, termed 'single-cut' or 'mammoth' clover in the USA, remain vegetative after the first cut and are generally more

persistent (Bird, 1948). Other types, such as the Swiss 'Mattenklee', combine the rapid regrowth of early flowering red clover with the persistence of the late flowering types (Boller et al., 2010). Thus, the various breeding activities combined with the use of landraces and the introgression of wild populations for specific characteristics resulted in very diverse genetic resources, adapted to vastly different environments – ranging from arctic to mediterranean and continental climates. However, most of these breeding activities were focused on creating cultivars optimized for local or regional needs, and little effort was spent on breeding for a broader range of environments. The latter may be particularly important in the future in view of rapidly changing environmental conditions through climate change and shifts in priorities for land-use.

Plant genetic resources, *i.e.*, the genetic material contained in wild relatives, ecotypes, landraces, and cultivars, are crucial for the continued improvement of modern forage crop cultivars (Boller and Greene, 2010). For red clover, the European Search Catalogue for Plant Genetic Resources (EURISCO; https://eurisco.ipk-gatersleben.de/apex/eurisco_ws/r/eurisco/home) currently lists more than 10,600 red clover accessions and many more are likely to be found in non-European or local germplasm collections. For targeted utilization in breeding programs, detailed phenotypic and genotypic characterization is indispensable, but is often not available for many of the current genetic resources. Molecular genetics or DNA-based markers allow a rapid characterization of genetic diversity in a large number of samples, independent of environmental effects (Bachmann, 1994). PCR-amplification based marker systems targeting either many unspecific loci (*e.g.*, amplified fragment length polymorphism (AFLP) markers; Vos et al., 1995) or single highly variable loci (*e.g.*, simple sequence repeat (SSR) markers; Tautz, 1989) have long been the methods of choice for analyzing genetic diversity in forage crop species where genome sequence information has been scarce. More recently, highly cost-effective genotyping methods based on high-throughput sequencing of restriction site-associated DNA such as RAD-seq (Baird et al., 2008) or genotyping-by-sequencing (GBS; Elshire et al., 2011) have become available. These methods allow an even larger number of samples to be efficiently genotyped at several thousand loci. In red clover, molecular markers have been successfully used to investigate the origin of red clover types and cultivars (Semerikov et al., 2002; Kölliker et al., 2003), the genetic structure of germplasm collections (Herrmann et al., 2005; Dias et al., 2008; Gupta et al., 2017) and the distribution of genetic diversity over a broad geographic range (Jones et al., 2020). As expected for an outbreeding species, a large proportion of the genetic variation is observed between individuals within populations (Kölliker et al., 2003; Dias et al., 2008), making it necessary to analyze multiple individuals to capture the genetic diversity present in a population. Pooling leaves of 20 individual red clover plants prior to DNA extraction has been proven effective to analyze genetic diversity on a population level (Herrmann et al., 2005). In perennial ryegrass, GBS on pooled leaf samples has proven useful to establish genome wide allele frequency fingerprints and to differentiate populations (Byrne et al., 2013). Despite the large number of accessions available in genebanks, most studies have used a rather restricted number of accessions and/or individual plants or focused on specific geographic regions. Jones et al. (2020) investigated genetic variability of 75 red clover accessions using eight to 16 individual plants per accession and more than 8,000 single nucleotide polymorphism (SNP) markers. The accessions were clearly grouped into four groups, corresponding to their geographic origin (Asia, Iberia, UK and Central Europe). The largest study so far investigated genetic diversity of 382 Nordic red clover accessions using 661 SNPs (Osterman et al., 2022).

Although studies on genetic diversity can substantially assist breeding decisions, knowledge about phenotypic diversity is key to efficiently exploit red clover germplasm for cultivar development. While some studies have investigated the variability and inheritance of specific traits such as disease resistance (Frey et al., 2022), seed yield (Herrmann et al., 2005) or flowering time (Jones et al., 2020), only a few studies have investigated phenotypic diversity of several traits within a region or country and even fewer have studied diversity across regions. A study by Zanotto et al. (2021) that was conducted with 48 ecotypes and six cultivars from Norway, Sweden, and Finland found large variation in winter survival and yield between red clover accessions. Wild accessions sometimes outperformed commercial cultivars, indicating their value for improving adaptation to colder climates. Winter survival was found to vary considerably between the four locations and could only partially be predicted by a test under controlled conditions (Zanotto et al., 2021). Furthermore, in a meta-analysis of legume yields in trials conducted in several locations ranging from Southern Germany to Northern Norway, Halling et al. (2004) found large differences in forage yield not only across different environments but also between early, intermediate and late flowering accessions. These findings highlight the need to account for accession x location interactions and to characterize breeding germplasm in their respective environments.

In order to provide a solid foundation for the utilization of red clover genetic resources in breeding programs, we aimed at (i) establishing a representative, diverse collection of red clover accessions, (ii) characterizing the genetic diversity and structure present in this collection, (iii) assessing the phenotypic diversity of agriculturally relevant traits at multiple locations and (iv) getting an insight into the extent of accession x location interaction for future adaptive breeding efforts.

2 Materials and methods

2.1 Red clover diversity panel

In this study, the red clover diversity panel for the project 'Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency – EUCLEG' (www.eucleg.eu) was compiled to cover as much of the available genetic diversity present in diploid red clover in Europe as possible. Members of the project consortium, including breeders, research institutes and genebanks, were asked to contribute with their available material. In addition, other institutes outside the consortium were contacted for possible contributions to compile as many accessions as possible. Accessions were delivered with specific material transfer agreements from each supplier to a central coordinator, who further

distributed seeds under the same agreements to the researchers performing the experiments. Thousand seed weight of each accession was measured by the central coordinator to confirm diploidy. The final panel consists of 395 accessions representing available diversity of diploid red clover in Europe, complemented with a few accessions from the Americas and Oceania (Figure 1, Supplementary Table S1). In addition to country, accessions were grouped by their region of origin: Americas (16 accessions), Oceania (10), Central Europe (51). Eastern Europe (82), Switzerland (81), Southern Europe (13), Northern Europe (139) and others (3, Japan, Iran).

2.2 Genotyping

Genotyping of the 395 accessions and SNP calling was conducted as described in Frey et al. (2022). In brief, plants from all accessions were grown in the greenhouse in 96-compartment trays. Fresh leaves from 200 plants per accession were harvested at the one-leaf stage and pooled for DNA extraction. DNA extraction was done with the QIAGEN DNeasy 96 Plant kit (QIAGEN, Manchester, UK) according to the user manual. After normalizing to 20 ng μl^{-1}, DNA of each accession pool was shipped to LGC Genomics (Berlin, Germany) for pooled-GBS and PE-150 Illumina sequencing. SNP calling and calculations of allele frequencies were done as described in Keep et al. (2020). A detailed description of the parameters used for this study are given in the supplementary methods of Frey et al. (2022). SNPs were only retained if allele frequencies of at least 10 accessions were between 0.05 and 0.95 and if mean allele frequencies across all accessions were between 0.05 and 0.95. SNP positions with more than 5% missing values were discarded. Remaining missing values in the allele frequency matrix were replaced by the mean allele frequency across all accessions at the given SNP position.

Genetic diversity among accessions was investigated using principal component analysis implemented in the function 'prcomp' (R Core Team, 2020). The genetic structure was further studied using discriminant analysis of principal components implemented in the 'adegenet' package version 2.7.3 (Jombart, 2008). The potential number of subpopulations (clusters) was estimated using successive *k*-means clustering. Partition of variance was investigated using permutational analysis of variance implemented in the function 'adonis2' of the 'vegan' package (Oksanen et al., 2022).

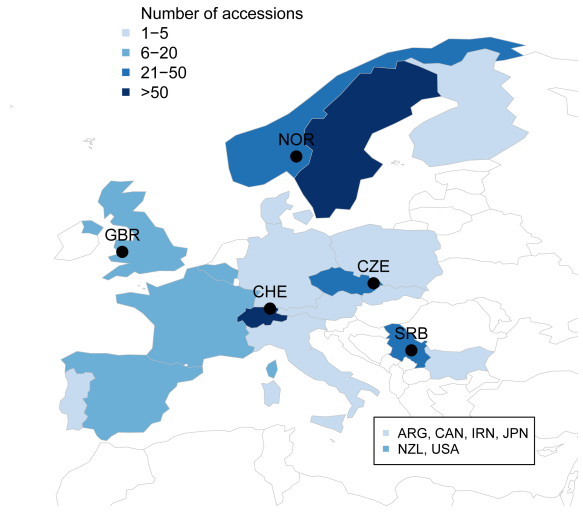


FIGURE 1
Origin and number of the European accessions in the EUCLEG red clover panel. Additional accessions from Argentina (ARG, 5), Canada (CAN, 1), Iran (IRN, 1), Japan (JPN, 2), New Zealand (NZL, 10) and USA (10) are not shown on the map. Locations of the field trials in Norway (NOR), Great Britain (GBR), Switzerland (CHE), Czech Republic (CZE) and Serbia (SRB) are indicated by black circles.

2.3 Field trials

2.3.1 Locations and trial designs

Field plot trials with the full set of all 395 accessions were conducted at Agroscope in Tänikon, Switzerland (CHE 47.480°N, 8.904°E, 535 m.a.s.l.) and at DLF Seeds AS in Hladké Životice, Czech Republic (CZE 49.690°N, 17.960°E, 220 m.a.s.l.). A reduced set of 110 accessions was grown at Graminor in Bjørke, Norway (NOR 60.757°N, 11.203°E, 147 m.a.s.l.) and reduced sets of 100 accessions each were grown at Aberystwyth University in Aberystwyth, Wales (GBR 52.427°N, 4.020°W, 35 m.a.s.l.) and at the Institute for Forage Crops Kruševac Ltd. in Kruševac, Serbia (SRB 43.583°N, 21.206°E, 150 m.a.s.l.). Temperature and precipitation at the trial sites are given in Supplementary Figure S1. To assess flowering time independently of the cutting regime of the main plot, additional observation rows with two replicates were sown at each trial location.

For the main plot trials in CZE, CHE and GBR, partially replicated (Prep) designs were employed, with accessions being grown unreplicated, in two or in six replicates (Table 1). For these trials, no complete blocks (CB) were applied, and each row and column constituted an incomplete block (IB1 and IB2, respectively). The main plot trials in NOR and SRB were designed as alpha designs with two CB containing the full set of accessions (*i.e.* complete replicates) and incomplete blocks in only one direction (IB1) of dimension 1 row by 10 columns (10 plots) in NOR and dimension 1 row by 5 columns (5 plots) in SRB. Observation rows for assessing flowering time were designed with two CB in all locations. Due to their larger size, the observation rows in CHE and CZE further included rows as first dimension of incomplete blocks (IB1) and columns as second dimension of incomplete blocks (IB2).

Due to different growing conditions and standard equipment used by the experimenters in the five locations, trials were set up and nursed according to local standard procedures. Plot size ranged from 5 to 8 m^2 and seeds were sown at a depth of 1 cm, whereby the amount of seeds per plot was adjusted for thousand seed weight and germination rate in order to reach a sowing density of 600 or 800 germinating seeds m^{-2} (Table 1). All trials were to be sown at the beginning of the 2018 field season to establish plant stands with minimal weed infestation for subsequent measurements during 2019 and 2020. Due to unsuitable weather conditions, sowing had to be postponed to 8th of August in GBR and the trial in SRB had to

TABLE 1 Layout of field trials at the different locations.

	Trial location				
Parameter	CHE	CZE	GBR	NOR	SRB
Design	Prep	Prep	Prep	Alpha	Alpha
Number of accessions					
- with 1 plot	350	350	80	–	–
- with 2 plots	35	35	15	109	100
- with 6 or more plots	10	10	5	1[1)]	–
- in total	395	395	100	110	100
Number of plots	490[2)]	480	140	240	200
Trial dimension (rows x columns)	7 x 70	20 x 24	14 x 10	12 x 20	20 x 10
Complete blocks	No	No	No	Yes	Yes[3)]
Incomplete blocks 1st dimension	IB1 (1 x 70)	IB1 (1 x 24)	IB1 (1 x 10)	IB1 (1 x 10)	IB1 (1 x 5)
Incomplete blocks 2nd dimension	IB2 (7 x 1)	IB2 (20 x 1)	IB2 (14 x 1)	–	–
Plot size	1.5 x 5m	1.6 x 5m	1.2 x 5m	1.25 x 4m	1 x 5m
Distance between rows	0.14m	0.13m	n.a.[4)]	0.13m	0.15m
Sowing density (seeds m^{-2})	600	800	600	600	600
Sowing date	2018-04-21	2018-05-30	2018-08-08	2018-06-27	2018-09-17

1) Accession 'Lea' was grown with 22 plots in total (i.e. 11 plots per complete block).
2) 10 plots of the trial in CHE were sown with a control accession to facilitate harvest. No data were recorded for these plots (missing values).
3) No data assessment was possible in second complete block.
4) Plots were sown by manually distributing the seed.

be re-sown on the 17th of September due to drought and flood destroying the first trial sown in spring 2018. Because of additional damage to the field trial in SRB during early 2019, data could only be assessed from the first replicate at this location.

2.3.2 Phenotyping

All traits were assessed on a plot basis during the year of sowing (Y0 [2018]), and the first and second main harvesting years (Y1 [2019] and Y2 [2020], respectively). The cutting regime followed common practice for the respective locations, ranging from 3 (NOR), over 4 (CZE, GBR, SRB), to 5 (CHE) cuts per main harvesting year. Before each cut, different traits were assessed visually: plant vigor (VIG) was rated on a scale from 1 (very weak) to 9 (very vigorous), weed occurrence was rated as the percentage of plot biomass derived from weeds (Weed%) and the resistance to spontaneously occurring diseases was rated on a scale from 1 (very high disease infestation) to 9 (no disease infestation). Diseases observed in the trials were powdery mildew caused by *Microsphaera trifolii* (Grev.) U. Braun, southern anthracnose caused by *Colletotrichum trifolii* Bain & Essary, clover rot caused by *Sclerotinia trifoliorum* Eriks., plant rot caused by different *Fusarium ssp.* and brown spot caused by *Stemphylium sarciniforme* (Cav.) Wiltsh.

For each cut, plots were harvested with a plot harvester and weight of fresh matter was determined using a machine mounted balance. Fresh matter yield (FMY) was determined as fresh matter per plot divided by the plot area. In all trials but CHE, a subsample per plot was dried in an oven at 60°C for 72 hours to constant weight and percentage of dry matter concentration (DMC) was determined as weight of dry subsample divided by weight of fresh subsample multiplied by 100. For the trial in CHE, DMC was determined indirectly using near infrared (NIR) spectroscopy measurements taken on the plot harvester using a PSS 1721 spectrometer equipped with a PSS-H-A03 distance sensor head from Polytec GmbH (Waldbronn, Germany). Dry matter yield (DMY) corrected for weed occurrence was then calculated for each plot as:

$$DMY = FMY \times \frac{DMC}{100} \times \left(1 - \frac{Weed\%}{100}\right) \quad (1)$$

For the first and second cut of Y1, samples for forage quality analysis were taken from each plot at all locations and were dried for 72 hours to constant weight at 50 to 60°C. Dried samples were then shipped to Agroscope for central processing. Grinding of samples was done using a cutting mill (SM200, Retsch, Haan, Germany) with a mesh size of 0.75 mm. NIR spectra on the ground samples were measured in the wavelength range from 800 to 2500 nm using a laboratory spectrometer (NIRFlex N-500, Büchi, Flawil, Switzerland). The two parameters crude protein content (CP) and digestible organic matter (DOM) were then determined for each sample using NIR calibrations developed in-house based on reference values determined in red clover with the respective reference methods. Reference measurements for calibration development of CP (in % of dry matter) were established by

multiplying total nitrogen (% of dry matter) determined following Kjeldahl (1883) by a factor of 6.25. Reference determination for DOM (in mg g^{-1}) was performed using the *in vitro* assay following Tilley and Terry (1963).

In addition to the traits assessed before each cut, plant density was determined visually at the end of Y0 and at the beginning and end of Y1 and Y2. At each date, plant density was rated in percentage of the densest plot observed at the end of Y0. Date of flowering (DOF) was determined in Y1 and Y2 as the number of days after 1st of January when at least five plants per plot started flowering (visible appearance of pink color of the petals). Because the plots were cut before all accessions started flowering, DOF was determined on separate observation rows. In CHE, DOF was assessed in Y2, whereas it was assessed in Y1 for all other locations. For accessions that did not start flowering at all, a value of 222 was assigned.

2.3.3 Statistical analyses

Traits were analyzed for each cut separately. For comparison among locations, additional secondary traits were calculated. Mean vigor rating for Y1 and Y2 (VIG_Y1 and VIG_Y2, respectively) was calculated as simple mean over all cuts and a missing value was assigned if data from one cut was not available. The sum of DMY for Y1 and Y2 (DMY_Y1 and DMY_Y2, respectively) was calculated as sum of DMY over all cuts of the respective year.

Locations were analyzed separately using the following model:

$$y_{imno} = \mu + g_i + b_m + ib1_n + ib2_o + \epsilon_{imno} \quad (2)$$

where y_{imno} represents the observation for trait y on a single plot basis, μ denotes the overall mean, g_i is the effect of accession i, b_m the effect of block m, $ib1_n$ the effect of incomplete block 1 (*i.e.* row) n, and $ib2_o$ the effect of incomplete block 2 (*i.e.* column) o, and ϵ_{imno} is the residual error. If not present in the design of a given location, the respective factor was omitted for analysis (i.e. b_m for CHE, CZE and GBR and $ib2_o$ for NOR). Because data was only available from one complete block (*i.e.* replicate) in SRB, no separate analysis could be performed for this location. In a first model, to estimate the best linear unbiased estimates of accession means (BLUEs), $ib1_n$ and $ib2_o$ were considered as random, while all other effects were considered as fixed. In a second model, to estimate the best linear unbiased predictors of accession means (BLUPs) and the respective variance components, g_i was additionally considered as random. For all subsequent phenotypic analyses, BLUE values were used. Multi-location analyses using data from all five locations were done following the model

$$y_{ijmno} = \mu + g_i + e_j + ge_{ij} + b_m + ib1_n + ib2_o + \epsilon_{ijmno} \quad (3)$$

where y_{ijmno} represents the observation for trait y on a single plot basis, μ denotes the overall mean, g_i is the effect of accession i, e_j the effect of the location j, ge_{ij} the interaction between accession i and location j, b_m the effect of complete block m (if available), $ib1_n$ the effect of incomplete block 1 n, $ib2_o$ the effect of incomplete block 2 o, and ϵ_{ijmno} is the residual error. In a first model, effect $ib1_n$ and $ib2_o$ were considered as random, while all other effects were considered as fixed to estimate BLUEs. To estimate respective variance components, g_i and ge_{ij} were additionally considered as random in a second model. For single and across location analyses, heritability (h^2) values were calculated according to the method of Walsh and Lynch (2018) as the slope from the linear regression of BLUPs on BLUEs of a given trait.

For the principal component analysis (PCA) on phenotypic data, BLUEs of the traits DOF_Y1, CP_Y1.C1, CP_Y1.C2, DMY_Y1.C1, DMY_Y1.C2, DMY_Y1, DMY_Y2, VIG_Y1, VIG_Y2, PD_Y1 were used. Diseases were additionally included if they occurred at a given location. Occurring diseases were southern anthracnose and powdery mildew in CHE, brown spot in SRB, fusarium in CZE and clover rot in NOR. Only accessions with data in at least 50% of the selected traits were included and only traits that have data in at least 80% of the accessions were kept. Hence, DOF, DMY_Y2 and VIG_Y2 were not included for NOR and DMY_Y2 and VIG_Y2 were not included for CZE in the analysis. The remaining missing values were imputed with the mean of the trait.

For estimation of variance components and calculation of mean values, heritability, correlations coefficients, and PCA, DOF values of accessions that did not flower were set to 'not available'. Thereby, an artificial inflation of variance could be omitted. For all graphical representations, the pre-set value of 222 was kept for these accessions.

All statistical analyses of phenotypic data were conducted with R version 4.1.2 within RStudio v4.0.5. (R Core Team, 2020; RStudio Team, 2021) using the functions 'lmer' and 'ranef' of the package 'lme4' and 'emmeans' of the package 'emmeans' for mixed model analyses (Bates et al., 2015). For PCA analysis, the function 'prcomp' of the package 'stats' was used. In addition, the packages 'tidyverse', 'psych', 'ggfortify', 'rworldmap', 'gridExtra', 'cowplot', 'grid' and 'ggpubr' were used for data handling and illustration.

3 Results

3.1 Genetic diversity of the EUCLEG red clover panel

Genotyping-by-sequencing on pooled samples of the 395 accessions resulted in a set of 20,137 quality controlled and filtered SNPs evenly distributed across the red clover genome (Frey et al., 2022). Three samples (TP107, TP133, TP309; Supplementary Table S1) did not meet the quality requirements and were excluded from further analyses, resulting in a reduced dataset of 392 accessions for which GBS data was available.

The first two axes of the PCA together explained 14.5% of the genetic variation and allowed to distinguish some of the accessions based on their region of origin and their type (Figure 2). Principal component 1 (PC1) mainly separated ecotypes from Southern Europe from the remaining accessions, while PC2 allowed to distinguish accessions from Switzerland, Central Europe and Northern Europe. This separation was even more pronounced for PC3, which also allowed to distinguish breeding material and cultivars from the ecotypes and landraces in the Northern European material (data not shown). Permutational multivariate

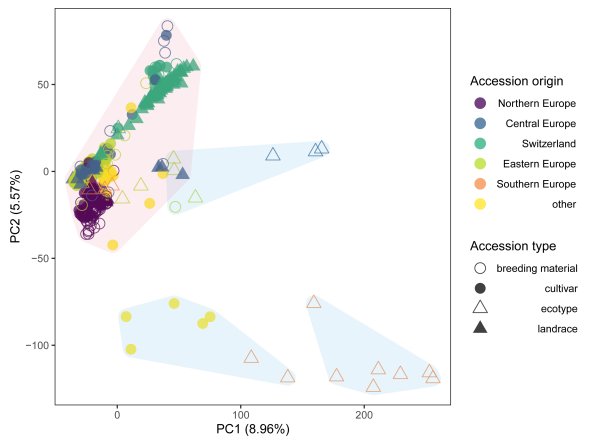


FIGURE 2
Genetic similarity of 392 red clover accessions within the EUCLEG panel revealed by principal component analysis (PCA) of allele frequency data based on 20,137 SNP markers and pooled leaf samples of 200 individuals per accession. Colors indicate the region of origin of the accessions and symbols indicate the accession type. *K*-means clustering identified three major clusters, which combined contained 371 accessions (shaded in red) and three minor clusters each comprising seven accessions (shaded in blue).

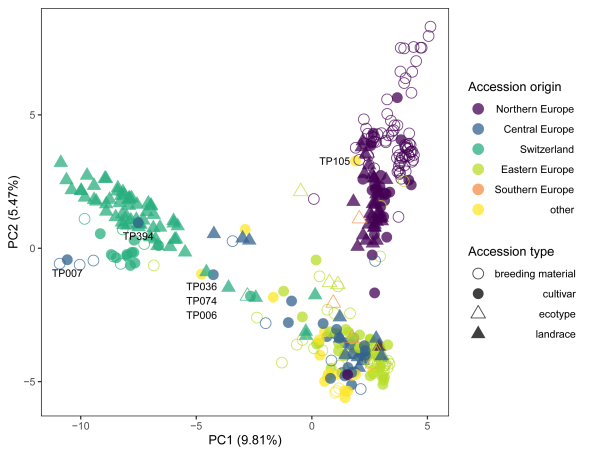


FIGURE 3
Genetic similarity of 371 selected red clover accessions revealed by principal component analysis (PCA) of allele frequency data based on 20,137 SNP markers and pooled leaf samples of 200 individuals per accession. Accessions forming the three major clusters of the complete set identified by *k*-means clustering (Figure 2) were selected. Colors indicate the region of origin of the accessions and symbols indicate the accession type. Accession IDs are given for selected accessions.

analysis of variance identified 16.8% of the variation to differentiate among regions, while 12.4% was attributed to variation between countries of origin and 70.8% to variation among accessions within countries (Table 2).

Successive *k*-means clustering of the PCs and comparing the models using the Bayesian Information Criteria (BIC) indicated the existence of six to 11 clusters (Supplementary Figure S2). Defining the clusters using $k = 6$ (the number of regions in the dataset) revealed three minor clusters that together contained 21 accessions, while the remaining three clusters contained the remaining 371 accessions (Figure 2). Consequently, a second PCA was performed on the subset of the 371 accessions grouped in the three major clusters.

PCA of the 371 accessions grouped in the three main clusters by successive *k*-means clustering further separated accessions according to their origin and accession type (Figure 3). PC1 and PC2 allowed to clearly separate three major groups: accessions from Northern Europe, accessions from Switzerland and accessions from the remaining regions. Particularly, PC2 distinguished Northern European breeding material and cultivars from landraces, while for Swiss accessions, cultivars are separated from landraces (Figure 3).

TABLE 2 Permutational multivariate analysis of variance of 392 red clover accessions from 24 countries and six regions genotyped using 20,137 SNPs.

Source of variation	DF[1]	SS[2]	Variance explained (%)	*F*-value	*P*-value
Region	5	16,423	17.0	17.694	0.091
Origin (region)	17	11,956	12.3	3.789	0.091
Accessions (origin)	369	68,498	70.7		

[1]Degrees of freedom; [2]Sum of squares.

3.2 Field trials

Except for the second field replicate in SRB, all trials were successfully established during the initial year (Y0) and could be used for assessment of yield, quality and disease occurrence data in the subsequent main harvesting years (Supplementary Table S2). In CHE, DMY was highest during the first three cuts in Y1 and strongly decreased thereafter (Figure 4). Due to southern anthracnose infections during cut 4, plant density was reduced allowing weeds to spread. In Y2, DMY in CHE started at a lower level compared to Y1 and was reduced to nearly zero for most plots towards the end of the season with red clover plants being replaced by weeds. In CZE, DMY for the first cut of Y1 (DMY_Y1.C1) was highest among all locations, but strongly decreased with the subsequent cut (Figure 4, Table 3). Along with the observation of severe fusarium rot during the third cut and subsequent infestation with mice, no more yield assessment was possible in CZE and the trial had to be abandoned. In GBR, weed infestation started at a relatively high level at cut 1 in Y1, but continuously decreased until cut 3, staying relatively low for the remaining time of the experiment. DMY thereby increased, reaching highest values for cut 3 in Y1 and for cut 2 in Y2. In NOR, three cuts were performed in both Y1 and Y2 with the third cut having a reduced DMY compared to the first two cuts in both main harvesting years. Occurrence of clover rot was observed before the first cut in Y1. Weed occurrence increased after winter at cut 1 in Y2, but again decreased thereafter. In SRB, plots had to be resown in autumn of Y0 and were cut four times in Y1, with DMY values decreasing with subsequent cuts. In Y2, DMY started at a higher level compared to Y1 and decreased until the third and last cut. Due to a generally low infestation with weeds, no weed infestation was assessed in SRB.

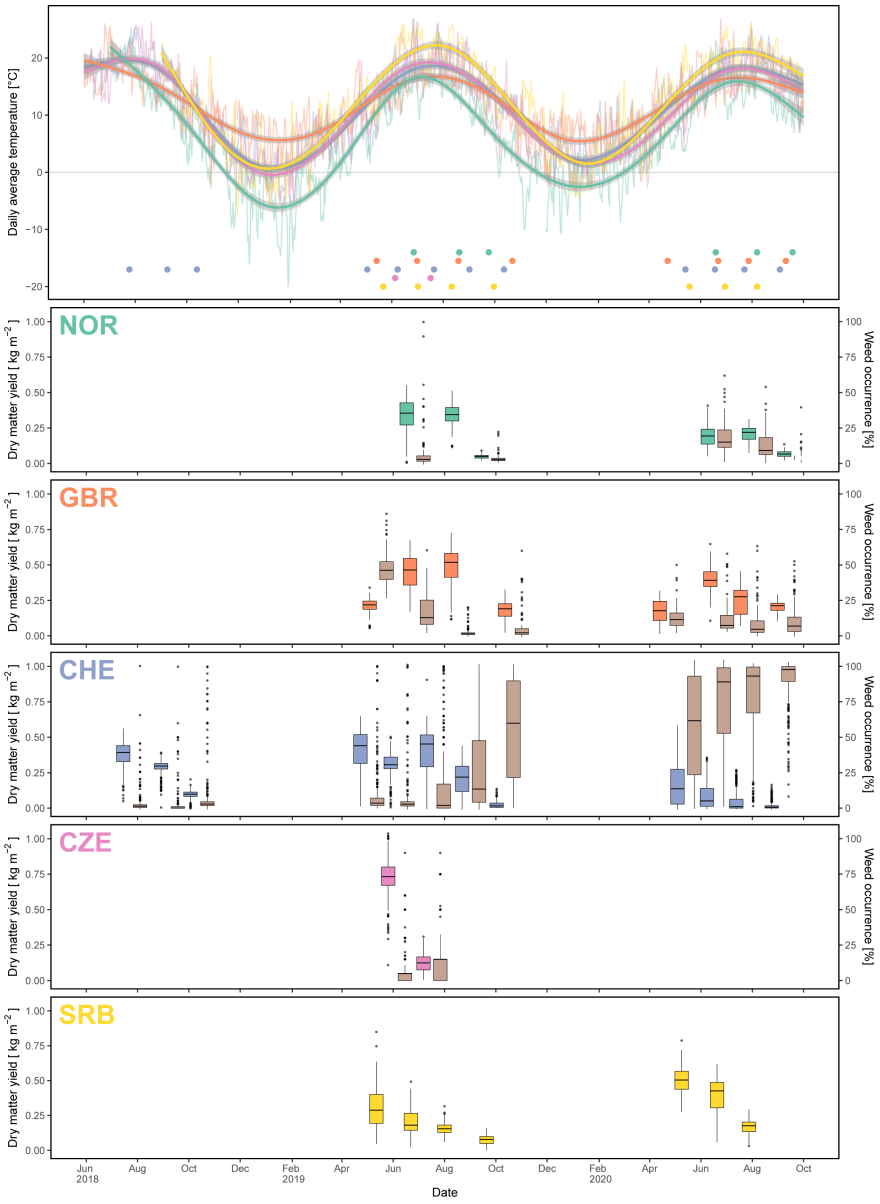


FIGURE 4
Uppermost panel: temperature course during the period of the experiment with thin lines denoting the daily average temperature, thick lines denoting the smoothed temperature and dots showing the cutting dates per location. Subsequent panels: boxplots of dry matter yield (different colour per location) and weed occurrence (brown colour) for each cut and location.

Due to the lack of data from location CZE after cut 2 in Y1 and the heavily reduced plant stands at location CHE after cut 4 in Y1 due to non-adapted (*i.e.*, southern anthracnose susceptible) accessions coming along with a high weed occurrence, subsequent analyses mostly focused on Y1. Given the vast amount of data obtained, a main focus was on the traits date of flowering in Y1 (DOF_Y1), total dry matter yield in Y1 and Y2 (DMY_Y1 and DMY_Y2, respectively), average vigor of Y1 (VIG_Y1) and plant density measured in autumn of Y1 (PD_Y1). In addition, we took a closer look at the crude protein data of the first and second cut in Y1 (CP_Y1.C1, CP_Y2.C1). Furthermore, the natural occurrence of diseases in the different trials is reported and discussed.

3.2.1 Variance components and heritability

Quality parameters of the red clover herbage were measured for the first and second cut of Y1. From single location analysis, most of the variation in CP was explained by the residual variance and only minor effects of accession or field heterogeneity could be observed (Figure 5, Supplementary Table S2). The second quality parameter, DOM, showed a very similar pattern to CP and was also largely dominated by residual variance (Supplementary Table S2). In comparison, variation in DMY of the corresponding cuts (DMY_Y1.C1, DMY_Y1.C2) could, except for location CZE, be explained to a large part by accession. Also, for DOF_Y1, PD_Y1, DMY_Y1, DMY_Y2, and VIG_Y1, a large proportion of the

10.3389/fpls.2023.1128823

TABLE 3 Mean and range of best linear unbiased estimators (BLUE) per accession observed at the five different locations.

	NOR		GBR		CHE		CZE		SRB	
Parameter	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
DMY_Y1 (kg m^{-2})	0.74	0.19-1.01	1.33	0.44-1.83	1.34	0.01-2.03	0.86	0.12-1.26	0.74	0.14-1.44
DMY_Y1.C1 (kg m^{-2})	0.34	0-0.55	0.21	0.05-0.34	0.41	0.01-0.65	0.73	0.11-1.07	0.31	0.05-0.85
DMY_Y1.C2 (kg m^{-2})	0.34	0.12-0.51	0.45	0.17-0.67	0.31	(-0.03)-0.5	0.12	(-0.02)-0.31	0.20	0.02-0.49
DMY_Y2 (kg m^{-2})	0.50	0.18-0.85	1.03	0.41-1.42	0.28	(-0.06)-1.24	NA	NA	1.07	0.4-1.55
DMY_Y2.C1 (kg m^{-2})	0.20	0.05-0.41	0.17	0.01-0.32	0.15	(-0.04)-0.58	NA	NA	0.50	0.28-0.79
DMY_Y2.C2 (kg m^{-2})	0.21	0.07-0.31	0.40	0.11-0.65	0.07	(-0.02)-0.36	NA	NA	0.38	0.06-0.62
CP_Y1.C1 (g kg^{-1})	174	119.73-215	208	147-265	249	194-300	162	93-250	151	105-225
CP_Y1.C2 (g kg^{-1})	172	120-226	173	116-220	247	179-295	200	122-266	153	91-204
DOM_Y1.C1 (mg g^{-1})	584	478-679	731	661-800	682	584-738	624	545-698	525	367-622
DOM_Y1.C2 (mg g^{-1})	576	525-621	647	532-731	638	523-748	524	444-603	531	375-630
DOF_Y1 (d)	181	167-199	150.8	109-222	153.9	121-222	153.2	142-164	132.4	125-142
PD_Y1.beginning (%)	26.1	(-0.1)-66.2	51.5	4.8-96.7	NA	NA	7.54	1.25-9.53	97.6	90-100
PD_Y1.end (%)	53.4	5.08-84.7	57.8	17.9-74.4	41.9	(-7.7)-94.9	3.42	(-0.5)-9.62	91.8	80-100
VIG_Y1 (rating 1-9)	8.36	6.55-9.16	6.53	3.25-8.01	5.63	2.02-8.46	6.11	1.8-8.01	5.74	2.25-8.75
VIG_Y2 (rating 1-9)	7.91	6.85-8.78	6.96	4.48-8.74	2.66	0.69-8.12	NA	NA	5.67	3.4-7.2

Traits are dry matter yield (DMY), crude protein content (CP), digestible organic matter (DOM), date of flowering (DOF), plant density (PD) and plant vigor (VIG). Extensions Y1 and Y2 denote the first and second year of observation, .C1 and .C2 the first and second cut per year of observation, respectively. NA indicates data not available.

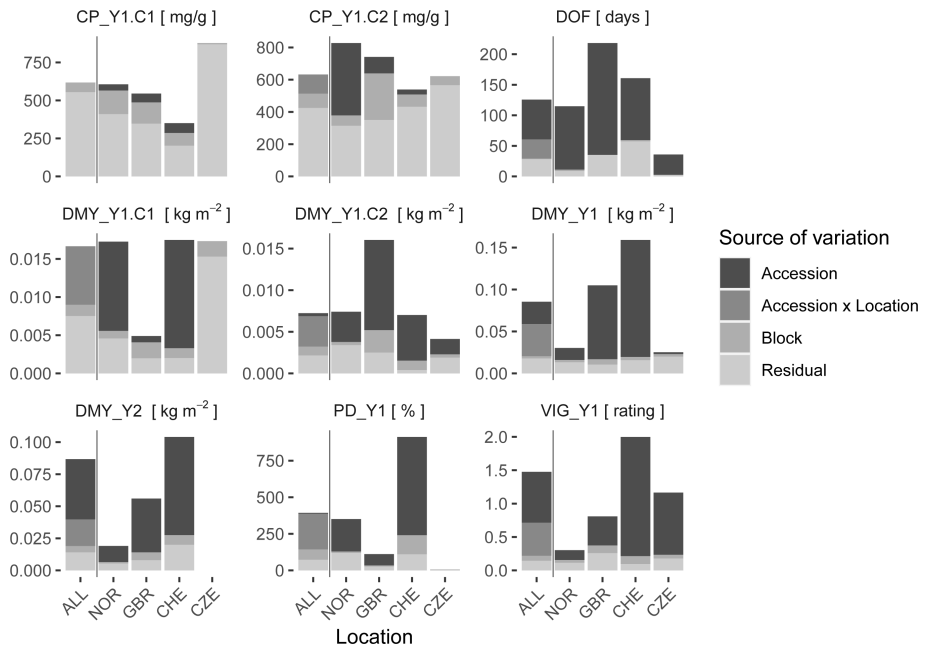


FIGURE 5
Variance components of single locations (NOR, GBR, CHE, CZE) and multi-location models (ALL). The variation attributed to the red clover accession, field heterogeneity as assessed by incomplete blocks (Block), the interaction of accession x location (for multi-location model only) and residuals is shown. Traits are crude protein content (CP), date of flowering (DOF), dry matter yield (DMY), plant density in autumn (PD) and plant vigor (VIG). Extensions _Y1 and _Y2 denote the first and second main year of observation, .C1 and .C2 the first and second cut per year, respectively. If the cut number is not specified, average values for VIG and sum of values for DMY over one year were used.

observed variation could be attributed to the accession, while the variance, due to field variation, indicated by block, and the residual variance were relatively low (Figure 5). When all locations were considered simultaneously in a multi-location analysis, a large accession x location effect was observed for most traits. For CP, values observed for the first cut (CP_Y1.C1) were still dominated by residual variance, while the accession variance observed for the second cut (CP_Y1.C2) was replaced by accession x location interaction variance. For DMY at specific cuts, *e.g.*, DMY_Y1.C1 and DMY_Y1.C2, the accession variance observed at single locations was almost entirely replaced by accession x location interaction variance. For sum of DMY per main harvest year, *e.g.*, DMY_Y1 and DMY_Y2, and average vigor of year 1 (VIG_Y1) some variance attributed to accessions was still observed in the multi-location analysis. The absolute amount of variation varied considerably between trials and traits. For the trial in SRB, no variance components could be estimated, because data could only be obtained from one replicate per accession.

Traits that showed low accession variance were accompanied by low heritability. For CP in both cuts, observed heritability values were very low ($h^2 < 0.25$) in all trials, except for the second cut in NOR (Figure 6). Due to the low genetic component of CP, a further detailed analysis of this trait is not provided. Heritability values for DMY showed a very wide range among locations, being lowest in CZE and highest in CHE. Highest heritability values were observed for DOF with values > 0.90 except for location CHE ($h^2 = 0.72$).

3.2.2 Correlation among locations

The high accession x location interactions observed for different traits (Figure 5), suggest that accessions performed differently at the five locations and that at least one of the trial locations showed a non-positive correlation to the others. Pairwise correlation analyses among locations for traits DMY_Y1, VIG_Y1 and PD_Y1 showed negative correlations of location NOR with all other locations, being significant in some instances (Figure 7). For the same traits, correlation coefficients among the other locations were all positive, being significant in most instances. Strongest correlations were observed between locations CHE and GBR. For DOF, a trait with only moderate accession x location interaction variance, correlation coefficients were positive among all locations, being significant in most instances (Figure 7).

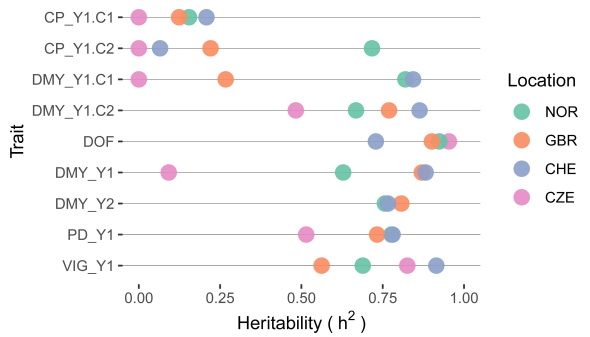


FIGURE 6
Heritability of selected traits within red clover field trials per location. Traits are crude protein content (CP), dry matter yield (DMY), date of flowering (DOF), plant density in autumn (PD) and plant vigor (VIG). Extensions _Y1 and _Y2 denote the first and second main year of observation, .C1 and .C2 the first and second cut per year, respectively. If the cut number is not specified, average values for VIG and sum of values for DMY over one year were used.

3.2.3 PCA of phenotype and disease occurrence

PCA was performed for each trial location using phenotypic data including disease occurrence (Figure 8). Natural disease infections occurred in all trials except location GBR. At location CHE, southern anthracnose was observed at Y0.C2 ($h^2 = 0$) and Y1.C4 ($h^2 = 0.40$), while powdery mildew was observed at Y0.C1 and Y0.C2 with heritability of 0.75 and 0.80, respectively. In location SRB, brown spot was observed in Y2.C2 ($h^2 = 0.67$). In location NOR, clover rot was observed in spring Y1.C1 ($h^2 = 0.59$). In location CZE, fusarium plant rot was observed in Y1.C3 ($h^2 = 0.30$). Correlations among traits are given in Supplementary Figures S3–7.

PCA of phenotypic values showed clustering of accessions according to their origin at each trial location. In the NOR trial, the local accessions (accession origin = Northern Europe) clustered in the upper left side and are distinguished from other accessions by an increased resistance to clover rot and an increased CP_Y1.C2. In addition, the local accessions flowered later and expressed high yields mainly in the first, but not in the second cut (high DMY_Y1.C1 and low DMY_Y1.C2). In GBR, PC1 was strongly determined by yield and PC2 by DOF and CP_Y1.C2. Notably, Northern European accessions and ecotypes showed low yield, while Central and Eastern European as well as Swiss accessions were high yielding. In CHE, PC1 mainly distinguished accessions by their sum of DMY per year (DMY_Y1 and DMY_Y2) and yield of the first cut (DMY_Y1.C1), while PC2 represents variation in disease resistance (powdery mildew and southern anthracnose) as well as DOF_Y1. Thereby, Northern European accessions clustered separately from other materials with lower total yields per season, but higher yield in the second cut and later flowering. In CZE, no clear clustering of accessions by origin was observed. Variation in fusarium resistance Y1.C3 and plant density PD_Y1 both correlated with PC2 and are indicative for reduced plant stands after disease infestation. In SRB, the Eastern European accessions are distinguished from the Central European and Swiss accessions mainly by their increased yield and vigor in Y1. A weak trend of Northern European accessions for later flowering and lower vigor in Y1 was observed (Figure 8).

3.2.4 Breeding material and cultivars adapted to local environmental conditions

To find signs of local adaptation in red clover breeding materials, the performance of accessions belonging to the type 'breeding material' or 'cultivar' and grouped by their geographic origin, were compared in the different trial locations across Europe (Figure 9). On average, the Northern European accessions flowered later in all locations and displayed a clear adaptation to their native northern climate: Northern European accessions showed the highest plant density after Y1 in NOR, while accessions of other origin did not persist very well. However, the Northern European

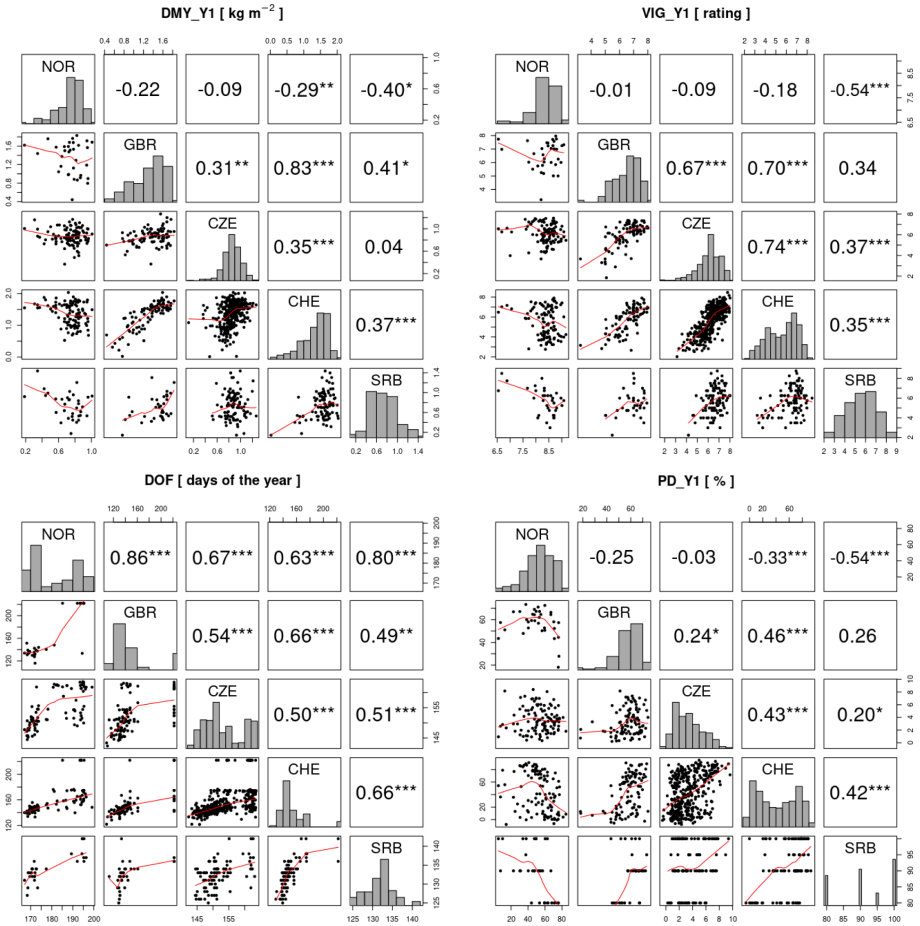


FIGURE 7
Correlation analysis of the traits total dry matter yield in the first year (DMY_Y1), average vigor in the first year (VIG_Y1), date of flowering (DOF) and plant density after the first year (PD_Y1) in the five field trial locations. In the diagonal, histograms of the trait values for each trial location are shown. Above the diagonal, Pearson correlations between trials are shown and significant correlations are indicated by asterisks (*p< 0.05, ** p< 0.01, *** p< 0.001). Below the diagonal, trait values of the two trials are plotted against each other with the red line representing the LOESS (locally estimated scatterplot smoothing) line.

accessions displayed reduced plant densities under non-native conditions and generally did not perform very well concerning yield and vigor. The Central and Eastern European accessions performed comparably well in the trials in CZE, CHE, and SRB for all traits. The Swiss accessions performed slightly better at their origin (CHE) regarding yield and vigor, and much better regarding plant density. However, in the trials in SRB and CZE, the Swiss accessions performed slightly worse compared to the Central and Eastern European accessions regarding yield and vigor. In GBR, the Central European accession performed best, followed by the Swiss and the Eastern European accessions.

3.2.5 Extent of accession x location interaction

Twenty accessions, which are representing all major regions of Europe, were tested in all five trials and their accession x location performance was analyzed (Figure 10). The Northern European accessions uniformly showed distinct patterns compared to the remaining groups. Northern European accessions flowered later than the other accessions and, although vigorous and high yielding in NOR, they did not perform well in the other locations. For the accessions from CHE, GBR, Central and Eastern Europe, such a pattern was not observed and within groups, usually good as well as poorly performing accessions were found.

4 Discussion

This study, to the best of our knowledge, characterised the most comprehensive collection of red clover accessions to date. By evaluating the collection on field plot level at diverse locations, a wide range of environmental conditions across Europe were covered. Combined with population-level allele frequency genotyping, this allowed to relate genetic diversity to phenotypic performance.

The genetic structure of the EUCLEG panel largely reflected the geographical origin of the accessions. Ecotypes from Southern Europe and some cultivars from non-European countries were clearly separated from the remaining accessions (Figure 2). A similar pattern was observed in an analysis of a global red clover collection, where accessions could be grouped into four regional

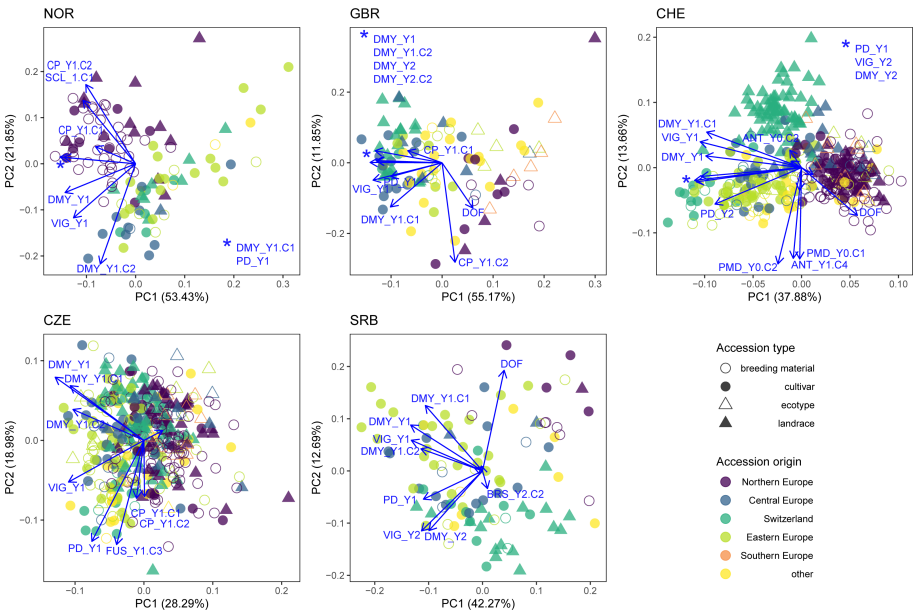


FIGURE 8
Principal component analysis (PCA) for each trial location including phenotypic data (DMY_Y1.C1, DMY_Y1.C2, CP_Y1.C1, CP_Y1.C2, DMY_Y1, DMY_Y2, VIG_Y1, VIG_Y2 and PD_Y1), disease scoring (ANT = southern anthracnose, BRS = brown spot, FUS = fusarium plant rot, PDM = powdery mildew, SCL = clover rot) and where they occurred. Colors indicate the region of origin of the accessions and symbols indicate the accession type.

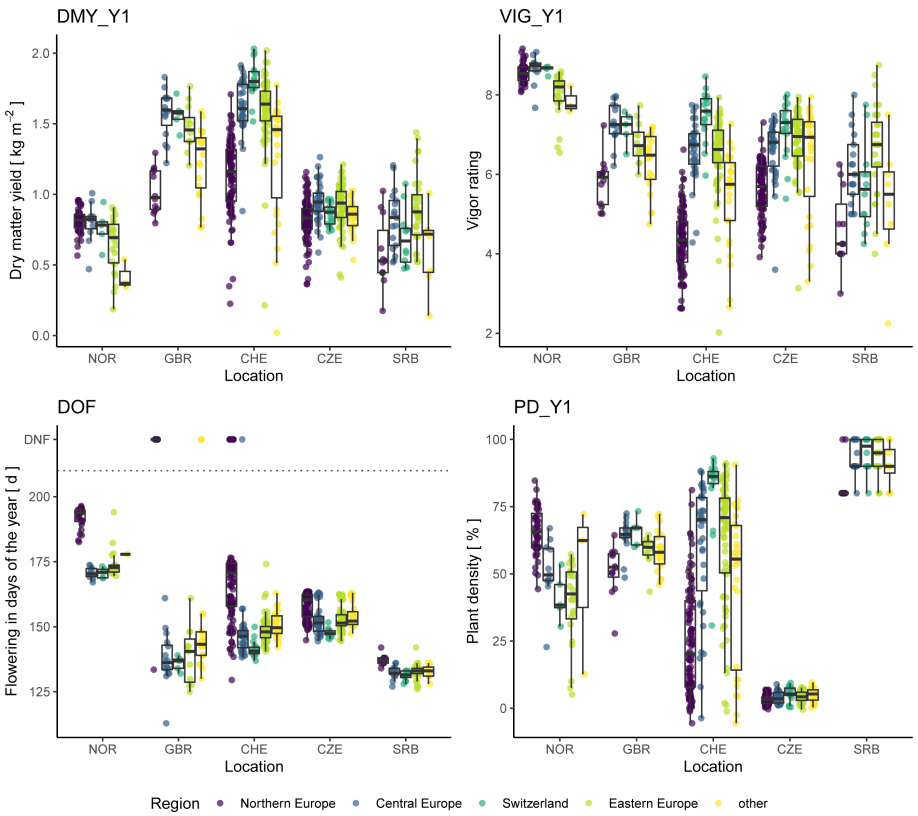


FIGURE 9
Performance of red clover accessions of the type 'cultivar' or 'breeding material' grouped by their region of origin. Traits analyzed are total dry matter yield of the first year (DMY_Y1), average vigor of the first year (VIG_Y1), date of flowering (DOF) as well as plant density after the first year (PD_Y1). In the plot of DOF, accessions that did not flower are displayed with a value of 222 and are labelled accordingly (DNF).

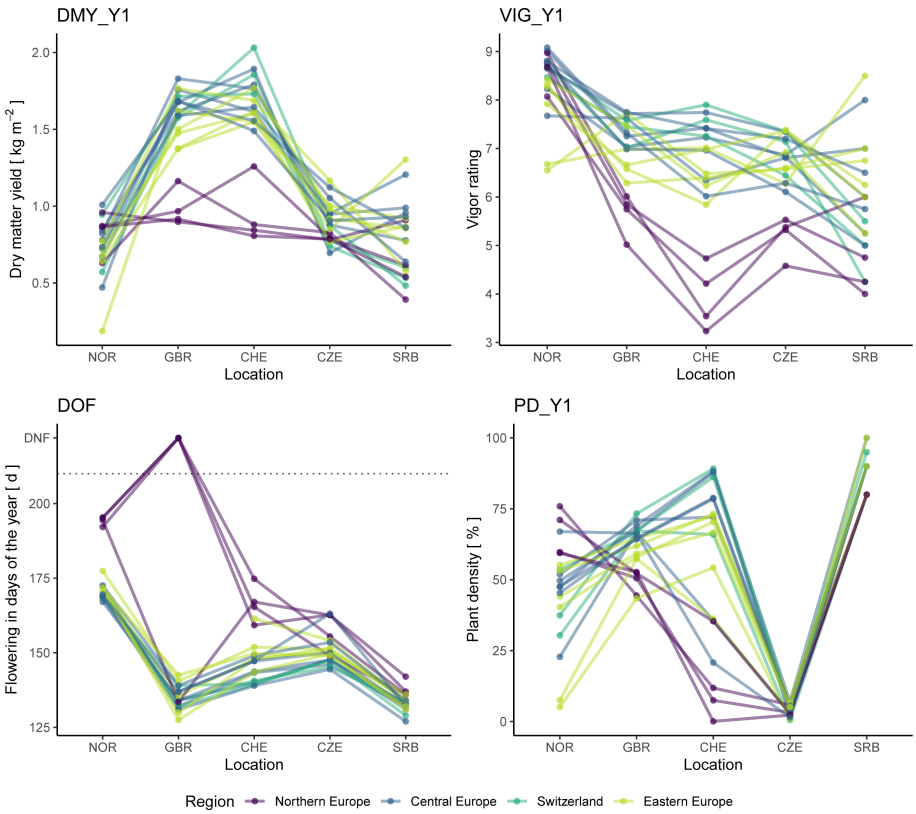


FIGURE 10
Performance of 20 red clover accessions grown in all five field trials. Traits analyzed are total dry matter yield of the first year (DMY_Y1), average vigor of the first year (VIG_Y1), date of flowering (DOF) as well as plant density after the first year (PD_Y1). In the plot of DOF, accessions that did not flower are displayed with a value of 222 and are labelled accordingly (DNF). Accessions are colored according to their region of origin.

groups (Asia, Iberia, UK and Central Europe) based on GBS of single plants (Jones et al., 2020). In the present study, accessions from Central and Eastern Europe clustered together, while accessions of Northern Europe and Switzerland each formed distinct clusters. Within the Northern European and Swiss clusters, sub-clusters were identified that distinguished breeding material from landraces (Figure 3). The genetic similarity of landraces and breeding material is in agreement with the historic origin of breeding programs that mostly started with improving local germplasm (Boller et al., 2010). Still, systematic improvement through breeding for traits like disease resistance or yield has led to populations that are distinctly different from the original landraces as indicated by the separate clustering of breeding material and landraces within the Swiss and Nordic accessions (Figure 3). A separation of ecotypes from cultivars has also been demonstrated recently in a study on 29 Nordic red clover accessions using genetic markers (Osterman et al., 2021).

Closer inspection of the genetic structure of the EUCLEG red clover diversity panel reveals a few particularities that may reflect specific breeding histories. For example, the cultivar 'Merian' (TP394) from Belgium clustered with the Swiss landraces, indicating the latter being the initial source of breeding germplasm used (Figure 3). According to information from the breeder, the cultivar 'Merian' orginates to 75% from ILVO breeding material from before 1970, which may also have included some Swiss landraces (T. Vleugels, personal communication). An interesting cluster of three accessions was observed at the PC1, PC2 coordinates -5, -0.5 (Figure 3): The Swiss landrace 'Ueberstorf' (TP074), as well as the cultivars 'Grasslands Sensation' from New Zealand (TP036) and 'AberClaret' from the United Kingdom formed a distinct group and the Pearson correlations of their allelic profiles ranged from 0.89 (TP036/TP074) to 0.94 (TP006/TP074; data not shown). 'Grasslands Sensation', also known as 'G40' or 'Swiss' was bred based on four Swiss red clover cultivars, *i.e.* 'Renova', 'Mont-Calme', 'Leisi' and 'Changins', all tracing back to Swiss landraces (Kölliker et al., 2003; Claydon et al., 2010). 'AberClaret' and the cultivar 'AberChianti' (TP007; Figure 3), are advertised as high-yielding up to their fourth and fifth harvest year (Marshall et al., 2017), a typical trait of Swiss 'Mattenklee' cultivars and landraces, to which they show considerable similarity in the PCA (Figure 3). The Canadian cultivar 'Altaswede' (TP105) clustered within the Northern European accessions. This is in agreement with it being described as 'single cut' clover (Valle, 1958), the predominant phenotype of red clover in Northern Europe. Indeed, 'Altaswede' was selected by the University of Alberta, Canada, based on Swedish seed stock around 1919 (Aasen and Bjorge, 2009). These examples are clear evidence that red clover breeders sometimes make use of the 'breeder's exemption', *i.e.*, the right to use registered cultivars from the market in their own breeding programs and integrate desired

traits into their own breeding pools. Nevertheless, the distinct grouping of the accessions according to their region of origin suggests that this is, so far, not frequently done and breeders mostly rely on their own, local genetic resources. This is in contrast with many other breeding programs of crops such as wheat or rice, where the incorporation of foreign material is far more widespread (Garrett et al., 2017; Luttringhaus et al., 2020).

While the characterisation of genetic diversity is an important first step in the efficient utilization of genetic resources, targeted incorporation of germplasm into the breeding process requires detailed phenotypic characterisation. Conducting field trials with perennial forage legumes is challenging. First, the establishment phase is critical to ensure reliable data in the following years. Second, pest, disease or weed occurrence can substantially reduce the plant density after each cut or during harsh winters (Boller et al., 1998; Öhberg et al., 2008; Zanotto et al., 2021), having a strong impact on the data assessed in subsequent cuts and growing seasons. Although the multi-site trials in this study were all managed by experienced red clover researchers and field teams, there were some setbacks during the trial period. The trial in CZE had to be abandoned after year one, as the damage caused by mice and *Fusarium* spp. was too severe to get reliable data in subsequent years. Several other trials experienced disease pressures that resulted in the loss of several accessions and missing data in the following cuts. Despite the difficult conditions encountered at some trial locations, heritability values of field trials were, with a few exceptions, at a high level ($h^2 > 0.6$, Figure 6) and similar to heritability values reported in other perennial forage crops (Annicchiarico et al., 1999; Han et al., 2006; Li et al., 2015).

A trait showing low heritability across all trial locations was CP. The reason for this is not entirely clear, but may be associated with the low overall variability for this trait (coefficient of genetic variation (CV_g) = 0.07-0.18, Supplementary Table S2). Only the heritability of CP of the second cut (CP_Y1.C2) for trial site NOR was satisfactory with $h^2 = 0.72$. It was highly correlated with DOF (r = 0.76, Supplementary Figure S3) and negatively correlated to DMY_Y1.C2 (r = -0.28). Hence, accessions that had a high CP in C2 generally flowered later, yielded more in C1 and less in C2 and may have concentrated protein levels in C2 (data not shown). Despite several publications reporting red clover forage quality values (Schubiger et al., 1998; Hoekstra et al., 2018), to our knowledge no heritability values have been reported so far. Thus, it remains unclear whether the heritability is truly low or, whether there was not enough variation in the accessions studied. The latter has previously been reported for alfalfa, where limited variability led to low narrow sense heritability of CP (Guines et al., 2002). In addition, the CP levels of red clover are already high compared to forage grasses (Schubiger et al., 1998) and therefore not a primary breeding goal. In alfalfa, leaf-stem ratio influences forage quality and is an important breeding aim (Annicchiarico, 2015).

During the experiment, the plant density of several red clover accessions declined at specific locations. In NOR, the plant density in autumn of Y1 was on average 39% for Eastern European accessions compared to 63% for native Nordic accessions (data not shown). In comparison, in SRB the plant density was on average 83% for Nordic accessions compared to over 90% for accessions from other regions. The overall high plant density in SRB after Y1 (Figure 9) is likely caused by the late sowing (autumn compared to spring for other trials) and the mild winter following sowing. In the Swiss trial, which experienced high southern anthracnose disease pressure, average plant densities after Y1 were highest for local, American and Eastern European accessions (62%, 57% and 59%, respectively) and lowest for Northern European accessions (21%), which is in agreement with the level of southern anthracnose resistance observed for these accessions in artificial inoculation experiments (Frey et al., 2022). The low plant density after Y1 resulted in non-satisfactory data quality in the second year for many accessions. Often, accessions that failed to persist were of foreign origin and not adapted to the conditions of the trial location, which indicates a strong effect of geographical origin on performance. Indeed, the clustering by geographical origin of accessions was also apparent when phenotypic data per trial was analysed (Figure 8). The genetic distinctness of these populations is, thus, manifested in their phenotypic performance. It also indicates a degree of local adaptation among accessions originating from similar environments.

Almost 80 years ago, Pieters and Hollowell (1937) reported that "claims of superiority are made for all of these regional strains, and the evidence of comparative trials shows that in most cases each such regional variety is superior to others in the environment where it was developed". The present work provides strong support for this hypothesis with trial locations covering a wide range of eco-environmental conditions relevant for European red clover cultivation areas. Breeding material and cultivars from Northern Europe showed best performance at the NOR location regarding DMY and vigor (Figure 9). This trend of local adaptation with accessions performing well at the location where they were bred was also observed for other trial locations and is in agreement with previous findings (Valle, 1958; Rosso and Pagano, 2005). In the CHE trial, Swiss breeding materials and cultivars showed best performance regarding DMY_Y1 and VIG_Y1, while in the GBR trial Central European accessions performed best. In CZE and SRB, Eastern European breeding materials performed best (Figure 9). As highlighted above, performance of red clover is often driven by persistence, and the plant density of poorly adapted accessions declines throughout the growing season and/or winter. While red clover has a natural senescence, its persistence can be improved by improving resistance for relevant diseases that occur in the growing location (Taylor, 2008). Depending on the region, different traits are relevant for persistence and hence have resulted in accessions adapted to the location of origin.

The five trial locations cover a wide range of latitudes and climate zones, and large accession x location interactions were observed. From correlation analysis among locations (Figure 7), we could show that especially the Northern European trial location (NOR) differed from the remaining sites. Negative correlation coefficients indicated the presence of cross-over interactions, *i.e.* that ranking of accessions is reversed in NOR compared to other locations. This observation clearly highlights the need for locally adapted breeding programs. We observed a strong adaptation of Northern European accessions to the prevailing conditions and/or management. According to Boller et al. (2010), European cultivars of red clover can be grouped by flowering time: early-flowering

types adapted to southern latitudes are capable of rapid regrowth and continuous flowering, while late-flowering types adapted to northern latitudes remain vegetative after the first cut. This may explain the poor performance of Northern European accessions of red clover in the more southern locations, where the cutting regime is optimized for fast regrowing types. At these locations, the first cut is most likely too early for Northern European accessions, which show lower yield as they are not able to resume generative growth afterwards. In their native habitat of NOR, the first cut happened 1-1.5 months later than in the remaining trial locations and allowed the Northern European accessions to reach their full yield potential.

Due to the natural expansion across Eurasia and Northern Africa and the recent domestication at several locations (Kjærgaard, 2003), red clover has evolved as a very diverse crop. The outcrossing nature of red clover has, in addition to the large between-accession diversity, resulted in an even higher within-accession diversity (Kölliker et al., 2003; Dias et al., 2008; Pagnotta et al., 2010; Collins et al., 2012). This allows red clover accessions to rapidly adapt to new challenges. Being a perennial species, adaptation in red clover has even been detected during continuous cultivation without allowing reproduction. Collins et al. (2012) found considerable genetic shifts within accessions grown at locations different than their origin after three to five years of continuous cultivation. Especially the Swedish environment was very selective and composite accessions became more similar to the Swedish accessions than to their original accession (Collins et al., 2012). Not only different environments but also changing environmental conditions between years may contribute to allelic diversity, as was shown in natural populations of perennial ryegrass (Keep et al., 2021).

A peculiarity of forage crops is the co-existence of wild relatives and cultivated forms in many landscapes which carries the potential for intercrossing. Diversity of wild red clover accessions has been shown not only to be influenced by the geographic location, but also by the proximity to sites where landraces have been grown for decades (Herrmann et al., 2005). This indicates potential geneflow between wild and cultivated forms. It also demonstrates the potential of utilizing wild germplasm for improving forage crop cultivars. Potentially useful adaptation to environmental conditions has also been demonstrated for wild alfalfa and other *Medicago* ssp., where candidate genes associated with adaptation to environmental gradients have been identified (Blanco-Pastor et al., 2021). Thus, variability in environmental conditions seems to drive genetic diversity in forage crop accessions and cultivars, which is a prerequisite for adaptation to changing climatic conditions.

As indicated by the large accession x location interactions observed in our study, breeding a 'one-fits-all' cultivar seems to be very unrealistic, due to the very variable environmental conditions demanding different adaptive traits from the genetic material as discussed above. However, there are trial locations that were more similar to each other based on assessed trait values and cultivars being able to perform well in such regions are more likely to be found. Growing a reduced set of diverse accessions at more locations throughout Europe might help to define such macro-environments representing the different red clover growing conditions. Knowledge of available genetic resources and their characteristics is key for breeding red clover well adapted to environmental conditions. This will become even more important in view of changing climatic conditions. For example, with increasing temperatures diseases adapted to warmer climates like southern anthracnose, which has already extended its local distribution (Jacob et al., 2016; Hartmann et al., 2022), will most likely continue to spread to Northern latitudes. Materials resistant to this disease might be a valuable source to complement the genetics from other regions not yet affected. The present work will serve as a valuable basis to identify interesting materials expressing desired characteristics. In conclusion, the detailed description of the EUCLEG red clover diversity panel provides an important source for further targeted improvement of red clover. The knowledge gained about the importance of local adaptation and genetic differentiation will help to design novel breeding strategies that consider broader scale adaptation of future cultivars. Additionally, the large amount of SNP markers provided for the EUCLEG red clover diversity panel, together with phenotypic data assessed in the presented field trials or other complementary phenotypic studies, may enable the implementation of genome wide association studies to unravel the genetic architecture of agronomically important traits.

Data availability statement

The datasets presented in this study can be found at https://zenodo.org/deposit/7034131.

Author contributions

Study design and curation of the accession collection: CG, RK, and AP. Generation of phenotypic data: CG, HA, JR, LJ, and LS. Generation of genotypic data: LS, TR, and LF. Data analysis and interpretation: MN, RK, CG, and LF. Drafting of the manuscript: MN, RK, and CG. All authors contributed to the article and approved the submitted version.

Funding

The research for this paper was financially supported by the European Union's Horizon 2020 Program for Research & Innovation under grant agreement no. 727312 (project: 'EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency').

Acknowledgments

We thank the EUCLEG consortium for the valuable training and exchange sessions and all institutions providing red clover seeds, including: AgResearch (NZL), Agricultural Res. Ltd. (CZE), Agroscope (CHE), Boreal (FIN), DLF Seeds (CZE), DSV (DEU), Graminor (NOR), HBLFA (AUT), Hokkaido Ag. Res. (JPN), IBERS

(GBR), IFVCNS (SRB), ILVO (BEL), INRAE (FRA), Lantmännen (SWE), NordGen (SWE), PGG Wrightson (NZL), RAGT2n (FRA), and USDA (USA). In addition, we would like to thank the technical staff at Aberystwyth University, Agroscope, DLF Seeds, Graminor, and Institute for Forage Crops Kruševac for their management of the field trials. This manuscript was made available as a preprint on bioRxiv https://doi.org/10.1101/2022.12.19.520744

Conflict of interest

Author HA was employed by the company Graminor Breeding Ltd and author LJ was employed by the company DLF Seeds.

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.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1128823/full#supplementary-material

References

Aasen, A., and Bjorge, M. (2009). *Alberta Forage manual. 2nd Edition* (Edmonton, Alberta, Canada: Alberta Agriculture and Rural Development Information Managment Services).

Abberton, M. T., and Marshall, A. H. (2005). Progress in breeding perennial clovers for temperate agriculture. *J. Agric. Sci.* 143 (2-3), 117–135. doi: 10.1017/s0021859605005101

Annicchiarico, P. (2015). Alfalfa forage yield and leaf/stem ratio: narrow-sense heritability, genetic correlation, and parent selection procedures. *Euphytica* 205 (2), 409–420. doi: 10.1007/s10681-015-1399-y

Annicchiarico, P., Piano, E., and Rhodes, I. (1999). Heritability of, and genetic correlations among, forage and seed yield traits in ladino white clover. *Plant Breed.* 118 (4), 341–346. doi: 10.1046/j.1439-0523.1999.00387.x

Bachmann, K. (1994). Molecular markers in plant ecology. *New Phytol.* 126 (3), 403–418. doi: 10.1111/j.1469-8137.1994.tb04242.x

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3 (10), e3376. doi: 10.1371/journal.pone.0003376

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67 (1), 1–48. doi: 10.18637/jss.v067.i01

Bird, J. N. (1948). Early and late types of red clover. *Sci. Agric.* 28 (10), 444–453. doi: 10.4141/sa-1948-0062

Blanco-Pastor, J. L., Liberal, I. M., Sakiroglu, M., Wei, Y., Brummer, E. C., Andrew, R. L., et al. (2021). Annual and perennial medicago show signatures of parallel adaptation to climate and soil in highly conserved genes. *Mol. Ecol.* 30 (18), 4448–4465. doi: 10.1111/mec.16061

Boller, B., Bigler, P., Bucanovic, I., and Bänziger, I. (1998). "Southern anthracnose - a new threat for red clover persistence in cooler regions?," in *Breeding for a multifunctional agriculture. proceedings of the 21st meeting of the fodder crops and amenity grasses section of EUCARPIA*. Eds. B. Boller and F. J. Stadelmann (Ittingen, Switzerland: EUCARPIA), 195–198.

Boller, B., and Greene, S. L. (2010). "Genetic resources," in *Fodder crops and amenity grasses*. Eds. B. Boller, U. K. Posselt and F. Veronesi (New York Dordrecht Heidelberg London: Springer Science & Business Media), 13–37.

Boller, B., Schubiger, F. X., and Kölliker, R. (2010). "Red clover," in *Fodder crops and amenity grasses*. Eds. B. Boller, U. K. Posselt and F. Veronesi (New York Dordrecht Heidelberg London: Springer Science & Business Media), 439–455.

Broderick, G. A. (1995). Desirable characteristics of forage legumes for improving protein utilization in ruminants. *J. Anim. Sci.* 73 (9), 2760–2773. doi: 10.2527/1995.7392760x

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations *via* genotyping by sequencing. *PloS One* 8 (3), e57438. doi: 10.1371/journal.pone.0057438

Claydon, R. B., Rumball, W., and Miller, J. E. (2010). 'Grasslands sensation' red clover (*Trifolium pratense* l.). *N. Z. J. Agric. Res.* 46 (4), 355–357. doi: 10.1080/00288233.2003.9513564

Collins, R. P., Helgadottir, A., Frankow-Lindberg, B. E., Skøt, L., Jones, C., and Skøt, K. P. (2012). Temporal changes in population genetic diversity and structure in red and white clover grown in three contrasting environments in northern Europe. *Ann. Bot.* 110 (6), 1341–1350. doi: 10.1093/aob/mcs058

Dias, P. M. B., Julier, B., Sampoux, J.-P., Barre, P., and Dall'Agnol, M. (2008). Genetic diversity in red clover (*Trifolium pratense* l.) revealed by morphological and microsatellite (SSR) markers. *Euphytica* 160 (2), 189–205. doi: 10.1007/s10681-007-9534-z

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 (5), e19379. doi: 10.1371/journal.pone.0019379

Frey, L. A., Vleugels, T., Ruttink, T., Schubiger, F. X., Pegard, M., Skøt, L., et al. (2022). Phenotypic variation and quantitative trait loci for resistance to southern anthracnose and clover rot in red clover. *Theor. Appl. Genet.* 135, 4337–4349. doi: 10.1007/s00122-022-04223-8

Garrett, K. A., Andersen, K. F., Asche, F., Bowden, R. L., Forbes, G. A., Kulakow, P. A., et al. (2017). Resistance genes in global crop breeding networks. *Phytopathology* 107 (10), 1268–1278. doi: 10.1094/PHYTO-03-17-0082-FI

Guines, F., Julier, B., Ecalle, C., and Hughye, C. (2002). Genetic control of quality traits of lucerne (*Medicago sativa* l.). *Aust. J. Agric. Res.* 53, 401–407. doi: 10.1071/AR01104

Gupta, M., Sharma, V., Singh, S. K., Chahota, R. K., and Sharma, T. R. (2017). Analysis of genetic diversity and structure in a genebank collection of red clover (*Trifolium pratense* l.) using SSR markers. *Plant Genet. Res.* 15 (4), 376–379. doi: 10.1017/S1479262116000034

Halling, M. A., Topp, C. F. E., and Doyle, C. J. (2004). Aspects of the productivity of forage legumes in northern Europe. *Grass Forage Sci.* 59 (4), 331–344. doi: 10.1111/j.1365-2494.2004.00435.x

Han, Y., Bonos, S. A., Clarke, B. B., and Meyer, W. A. (2006). Inheritance of resistance to gray leaf spot disease in perennial ryegrass. *Crop Sci.* 46 (3), 1143–1148. doi: 10.2135/cropsci2005.07-0217

Hartmann, S., Schubiger, F. X., Grieder, C., and Wosnitza, A. (2022). A decade of variety testing for resistance of red clover to southern anthracnose (*Colletotrichum trifolii* bain et essary) at the Bavarian state research center for agriculture (LfL). *Agriculture* 12 (2), 249. doi: 10.3390/agriculture12020249

Herrmann, D., Boller, B., Widmer, F., and Kölliker, R. (2005). Optimization of bulked AFLP analysis and its application for exploring diversity of natural and cultivated populations of red clover. *Genome* 48 (3), 474–486. doi: 10.1139/G05-011

Hoekstra, N. J., De Deyn, G. B., Xu, Y., Prinsen, R., and Van Eekeren, N. (2018). Red clover varieties of mattenklee type have higher production, protein yield and persistence than ackerklee types in grass–clover mixtures. *Grass Forage Sci.* 73 (2), 297–308. doi: 10.1111/gfs.12307

Hofer, D., Suter, M., Haughey, E., Finn, J. A., Hoekstra, N. J., Buchmann, N., et al. (2016). Yield of temperate forage grassland species is either largely resistant or resilient to experimental summer drought. *J. Appl. Ecol.* 53 (4), 1023–1034. doi: 10.1111/1365-2664.12694

Jacob, I., Hartmann, S., and Struck, C. (2016). Response of different fodder legume species to *Colletotrichum trifolii*. *Crop Pasture Sci.* 67, 1110–1115. doi: 10.1071/cp16162

Jombart, T. (2008). Adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics* 24 (11), 1403–1405. doi: 10.1093/bioinformatics/btn129

Jones, C., De Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., et al. (2020). Population structure and genetic diversity in red clover (*Trifolium pratense* l.) germplasm. *Sci. Rep.* 10 (1), 8364. doi: 10.1038/s41598-020-64989-z

Keep, T., Rouet, S., Blanco-Pastor, J. L., Barre, P., Ruttink, T., Dehmer, K. J., et al. (2021). Inter-annual and spatial climatic variability have led to a balance between local fluctuating selection and wide-range directional selection in a perennial grass species. *Ann. Appl. Biol.* 128 (3), 357–369. doi: 10.1093/aob/mcab057

Keep, T., Sampoux, J. P., Blanco-Pastor, J. L., Dehmer, K. J., Hegarty, M. J., Ledauphin, T., et al. (2020). High-throughput genome-wide genotyping to optimize the use of natural genetic resources in the grassland species perennial ryegrass (*Lolium perenne* l.). *G3* 10 (9), 3347–3364. doi: 10.1534/g3.120.401491

Kjærgaard, T. (2003). A plant that changed the world: The rise and fall of clover 1000-2000. *Landsc. Res.* 28 (1), 41–49. doi: 10.1080/01426390306531

Kjeldahl, J. (1883). Neue methode zur bestimmung des stickstoffs in organischen körpern. *Z. Anal. Chem.* 22, 366–382. doi: 10.1007/BF01338151

Kölliker, R., Herrmann, D., Boller, B., and Widmer, F. (2003). Swiss Mattenklee landraces, a distinct and diverse genetic resource of red clover (*Trifolium pratense* l.). *Theor. Appl. Genet.* 107 (2), 306–315. doi: 10.1007/s00122-003-1248-6

Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., et al. (2015). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8 (2), 1–10. doi: 10.3835/plantgenome2014.12.0090

Lüscher, A., Mueller-Harvey, I., Soussana, J. F., Rees, R. M., and Peyraud, J. L. (2014). Potential of legume-based grassland-livestock systems in Europe: a review. *Grass Forage Sci.* 69 (2), 206–228. doi: 10.1111/gfs.12124

Luttringhaus, S., Gornott, C., Wittkop, B., Noleppa, S., and Lotze-Campen, H. (2020). The economic impact of exchanging breeding material: Assessing winter wheat production in Germany. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.601013

Marshall, A. H., Havard, D., and Dafis, A. (2017) *Advances in red clover breeding gives farmers vital homegrown alternative to bought-in protein* (Aberystwyth University). Available at: https://www.aber.ac.uk/en/news/archive/2017/07/title-204466-en.html (Accessed 29.09.2022).

Nyfeler, D., Huguenin-Elie, O., Suter, M., Frossard, E., and Lüscher, A. (2011). Grass-legume mixtures can yield more nitrogen than legume pure stands due to mutual stimulation of nitrogen uptake from symbiotic and non-symbiotic sources. *Agriculture Ecosyst. Environ.* 140 (1-2), 155–163. doi: 10.1016/j.agee.2010.11.022

Öhberg, H., Ruth, P., and Bang, U. (2008). Differential responses of red clover cultivars to *Sclerotinia trifoliorum* under diverse natural climatic conditions. *Plant Pathol.* 57, 459–466. doi: 10.1111/j.1365-3059.2007.01822.x

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., et al. (2022) *Vegan: Community ecology package*. Available at: https://github.com/vegandevs/vegan.

Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2021). Insights into the genetic diversity of nordic red clover (*Trifolium pratense*) revealed by SeqSNP-based genic markers. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.748750

Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2022). Discovering candidate SNPs for resilience breeding of red clover. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.997860

Pagnotta, M. A., Annicchiarico, P., Farina, A., and Proietti, S. (2010). Characterizing the molecular and morphophysiological diversity of Italian red clover. *Euphytica* 179 (3), 393–404. doi: 10.1007/s10681-010-0333-6

Pieters, A. J., and Hollowell, E. A. (1937). "Clover improvement," in *Yearbook of the United States Department of Agriculture* (Washington DC, USA: US Department of Agriculture), 1190–1214.

Posselt, U. K. (2010). "Breeding methods in cross-pollinated species," in *Fodder crops and amenity grasses*. Eds. B. Boller, U. K. Posselt and F. Veronesi (New York Dordrecht Heidelberg London: Springer Science & Business Media), 39–88.

R Core Team (2020). *R: A language and environment for statistical computing v.4.1.2* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Riday, H. (2010). Progress made in improving red clover (*Trifolium pratense* l.) through breeding. *Int. J. Plant Breed.* 4 (1), 22–29.

Rosso, B. S., and Pagano, E. M. (2005). Evaluation of introduced and naturalised populations of red clover(*Trifolium pratense* l.) at pergamino EEA-INTA, Argentina. *Genet. Resour. Crop Evol.* 52 (5), 507–511. doi: 10.1007/s10722-005-0777-z

RStudio Team (2021). *RStudio: Integrated development for r v4.0.5* (Boston, MA: RStudio, PBC). Available at: http://www.rstudio.com/.

Saharan, G. S., and Mehta, N. (2008). *Sclerotinia diseases of crop plants: biology, ecology and disease management* (New York Dordrecht Heidelberg London: Springer Science & Business Media).

Schubiger, F. X., Bosshard, H. R., and Lehmann, J. (1998). Futterwert von rotklee. *Agrarforschung Schweiz* 5 (4), 181–184.

Semerikov, V. L., Belyaev, A. Y., and Lascoux, M. (2002). The origin of Russian cultivars of red clover (*Trifolium pratense* l.) and their genetic relationships to wild populations in the urals. *Theor. Appl. Genet.* 106 (1), 127–132. doi: 10.1007/s00122-002-1033-y

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* 17, 6463–6471. doi: 10.1093/nar/17.16.6463

Taylor, N. L. (2008). A century of clover breeding developments in the united states. *Crop Sci.* 48 (1), 1–13. doi: 10.2135/cropsci2007.08.0446

Taylor, N. L., and Quesenberry, K. H. (1996). *Red clover science* Vol. 28 (New York Dordrecht Heidelberg London: Springer Science & Business).

Tilley, J. M. A., and Terry, R. A. (1963). A two-stage technique for the *in vitro* digestion of forage crops. *Grass Forage Sci.* 18 (2), 104–111. doi: 10.1111/j.1365-2494.1963.tb00335.x

Valle, O. (1958). Experiences with Canadian altaswede and Swedish commercial red clover in Finland. *Agric. Food Sci.* 30 (1), 293–299. doi: 10.23986/afsci.71463

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van de Lee, T., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23 (21), 4407–4414. doi: 10.1093/nar/23.21.4407

Walsh, B., and Lynch, M. (2018). *Evolution and selection of quantitative traits* Vol. 1 (Oxford, UK: Oxford University Press).

Weir, W. W. (1926). *Soil productivity as affected by crop rotation. farmer's bulletin no. 1475* (Washington, DC, USA: US Department of Agriculture).

Zanotto, S., Palmé, A., Helgadóttir, Á., Daugstad, K., Isolahti, M., Öhlund, L., et al. (2021). Trait characterization of genetic resources reveals useful variation for the improvement of cultivated Nordic red clover. *J. Agron. Crop Sci.* 207 (3), 492–503. doi: 10.1111/jac.12487

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 May 2023
DOI 10.3389/fpls.2023.1189662

Check for updates

OPEN ACCESS

EDITED BY
Yusheng Zhao,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

REVIEWED BY
Sanja LJ. Vasiljević,
Institute of Field and Vegetable Crops, Serbia
Siran Wang,
Nanjing Agricultural University, China

*CORRESPONDENCE
Åshild Ergon
ashild.ergon@nmbu.no

RECEIVED 19 March 2023
ACCEPTED 13 April 2023
PUBLISHED 10 May 2023

CITATION
Zanotto S, Ruttink T, Pégard M, Skøt L, Grieder C, Kölliker R and Ergon Å (2023) A genome-wide association study of freezing tolerance in red clover (*Trifolium pratense* L.) germplasm of European origin. *Front. Plant Sci.* 14:1189662. doi: 10.3389/fpls.2023.1189662

COPYRIGHT
© 2023 Zanotto, Ruttink, Pégard, Skøt, Grieder, Kölliker and Ergon. 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.

A genome-wide association study of freezing tolerance in red clover (*Trifolium pratense* L.) germplasm of European origin

Stefano Zanotto[1], Tom Ruttink[2], Marie Pégard[3], Leif Skøt[4], Christoph Grieder[5], Roland Kölliker[6] and Åshild Ergon[1]*

[1]Faculty of Biosciences, Norwegian University of Life Sciences (NMBU), Ås, Norway, [2]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium, [3]INRAE P3F, Lusignan, France, [4]IBERS, Aberystwyth University, Aberystwyth, United Kingdom, [5]Group of Fodder Plant Breeding, Agroscope, Zurich, Switzerland, [6]Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland

Improvement of persistency is an important breeding goal in red clover (*Trifolium pratense* L.). In areas with cold winters, lack of persistency is often due to poor winter survival, of which low freezing tolerance (FT) is an important component. We conducted a genome wide association study (GWAS) to identify loci associated with freezing tolerance in a collection of 393 red clover accessions, mostly of European origin, and performed analyses of linkage disequilibrium and inbreeding. Accessions were genotyped as pools of individuals using genotyping-by-sequencing (pool-GBS), generating both single nucleotide polymorphism (SNP) and haplotype allele frequency data at accession level. Linkage disequilibrium was determined as a squared partial correlation between the allele frequencies of pairs of SNPs and found to decay at extremely short distances (< 1 kb). The level of inbreeding, inferred from the diagonal elements of a genomic relationship matrix, varied considerably between different groups of accessions, with the strongest inbreeding found among ecotypes from Iberia and Great Britain, and the least found among landraces. Considerable variation in FT was found, with LT50-values (temperature at which 50% of the plants are killed) ranging from -6.0°C to -11.5°C. SNP and haplotype-based GWAS identified eight and six loci significantly associated with FT (of which only one was shared), explaining 30% and 26% of the phenotypic variation, respectively. Ten of the loci were found within or at a short distance (<0.5 kb) from genes possibly involved in mechanisms affecting FT. These include a caffeoyl shikimate esterase, an inositol transporter, and other genes involved in signaling, transport, lignin synthesis and amino acid or carbohydrate metabolism. This study paves the way for a better understanding of the genetic control of FT and for the development of molecular tools for the improvement of this trait in red clover through genomics assisted breeding.

KEYWORDS

cold stress, forage legumes, GWAS, haplotypes, linkage disequilibrium, LT50, pool-GBS, winter survival

1 Introduction

Red clover (*Trifolium pratense* L.) is the most important perennial forage legume in Northern Europe (Helgadóttir et al., 2014; Annicchiarico et al., 2015), where it is mainly cultivated in mixture with grasses under a cutting regime for the production of silage. Its capability to fix atmospheric nitrogen through symbiotic nitrogen fixation significantly reduces the requirement for nitrogen fertilization (Jensen et al., 2012; Lüscher et al., 2014; Reckling et al., 2016). Thanks to this, and to the high protein content (Frame et al., 1997), red clover plays an important role in the shift towards a more sustainable agriculture in Europe, which is currently not self-sufficient when it comes to the production of protein for food and feed (de Visser et al., 2014; Voisin et al., 2014). However, the cultivation of red clover at northern latitudes is hampered by its poor field persistency, mainly caused by low winter survival (Abberton and Marshall, 2005). The improvement of persistency is one of the main breeding goals for the species and will result in both higher forage yield and protein content of mixed red clover - grass swards (Marshall et al., 2017).

Freezing temperatures are one of several causes of winter mortality at locations with continental climate characterized by harsh winters (Bélanger et al., 2006). The ability to tolerate low freezing temperatures (freezing tolerance, FT) is largely dependent on a period of cold acclimation (CA) at low above-zero temperatures. FT measured under controlled conditions was successfully used as a determinant for freezing tolerance and field winter survival in winter wheat (*Triticum aestivum;* Gusta et al., 2001), perennial ryegrass (*Lolium perenne;* Waldron et al., 1998; Hulke et al., 2008) and white clover (*T. repens;* Annicchiarico et al., 2001). However, a more prolonged test, LD50 (lethal duration time for 50% kill), was proposed as a better proxy for winter survival of winter canola (*Brassica napus*) in the field than LT50 (Waalen et al., 2011). A previous study in red clover of Nordic origin (Zanotto et al., 2021), found that FT (calculated as LT50 with the same method as used here) was significantly correlated with winter survival at a continental location characterized by low freezing temperatures in the winter, showing that the LT50 method reveals relevant variation. However, the same study also identified a considerable genotype-by-environment (G×E) interaction on winter survival, which underlines the complex nature of this trait. FT can be improved by selection and breeding (Bertrand et al., 2016), but little is known about the genetic control of FT in red clover. Previous studies identified regions in the red clover genome associated with persistency under various biotic and abiotic stresses (Herrmann et al., 2008; Klimenko et al., 2010; Ergon et al., 2019), but none of these studies specifically analyzed the level of FT of the plant material.

Red clover is a natural diploid (2n = 2x = 14) outbreeding species with a genome size of approximately 420 Mb (Sato et al., 2005). A draft genome for the species (309 Mb, of which 164 Mb are placed on chromosomes) was published in 2015 by De Vega et al., facilitating genomic studies. Recently, a new chromosome-scale assembly of 413.5 Mb was published (Bickhart et al., 2022). Red clover is highly self-incompatible, has a very polymorphic genome, and usually a higher genetic diversity within than between populations (Jones et al., 2020). Red clover cultivars are commonly bred as synthetic populations with up to twenty or more parents. A recent study revealed that a high amount of genetic diversity is maintained within Nordic red clover cultivars created by breeding programs (Osterman et al., 2021). Furthermore, there is a large amount of genetic variation for both FT and other traits available among landraces, wild populations and old cultivars in the Nordic Genetic Resource Center (NordGen; Zanotto et al., 2021). Variation in such material can be exploited to improve cultivars for specific target traits (Kouamé and Quesenberry, 1993; Greene et al., 2004; Dias et al., 2008; Annicchiarico and Pagnotta, 2012), including traits related to persistency such as winter survival and freezing tolerance (Zanotto et al., 2021).

Genotyping-by-sequencing (GBS) of DNA pools is a time- and cost-effective method for genotyping a large number of accessions, which is useful when genetic characterization is more relevant at population level than at the constituent individual level, such as in plant breeding of self-incompatible forage species (Byrne et al., 2013). Genome-wide association studies (GWAS) have been successfully used in forage legumes to elucidate the genetic control of complex traits (Biazzi et al., 2017; Inostroza et al., 2018). GWAS analysis may be confounded by population structure and variation in relatedness among accessions (Korte and Farlow, 2013). Therefore, it is important to use GWAS models that can account for this in order to reduce the risk of identifying false positive associations (Liu et al., 2016). In GWAS analysis, haplotype data can be expected to complement single nucleotide polymorphism (SNP) data, as haplotypes may reveal associations that are not detectable when using SNP data only (Hamblin and Jannink, 2011; Bekele et al., 2018; Ergon et al., 2022).

The goal of this work was to characterize the level of FT of a diverse panel of red clover accessions and to identify genomic regions and candidate genes associated with FT through a GWAS.

2 Materials and methods

2.1 Plant material and phenotyping

A total of 393 red clover accessions were phenotyped for FT under controlled experimental conditions. These accessions are cultivars, breeding material, landraces and ecotypes mostly of European origin (Table 1; Supplementary Table 1) collected and characterized within the EUCLEG project[1] (Nay et al., 2023). FT was determined in young plants after a short cold acclimation treatment (two weeks at 3–4°C, 12 h photoperiod and 110 µmol $m^{-2}s^{-1}$ photosynthetic photon flux density (PPFD) and expressed as LT50 (temperature required to kill 50% of the plants), in a similar way as described by Zanotto et al. (2021). The experiment was organized in eight incomplete blocks, each containing half of the accessions. Accessions were randomized among blocks, so that each accession was represented in four of the eight blocks. Due to space limitation, only one block could be freeze-tested at the time, every

1 www.eucleg.eu

10.3389/fpls.2023.1189662

TABLE 1 The 393 red clover (*Trifolium pratense*) accessions used in this study, distributed across countries or regions and population types.

Region	Cultivars	Breeding material	Landraces	Ecotypes
Americas (Argentina, Canada, USA)	11	5		
Belgium	8	4	4	
Central or eastern Europe (Austria, Bulgaria, Poland, Slovakia)	1			4
Czech Republic	26	21		2
Denmark	3			
Finland	1		3	1
France	11			
Germany	4			
Great Britain	11	5	1	3
Italy				4
Norway	5	18	1	
Oceania/Asia (Japan, New Zealand)	12			
Serbia	10	14		5
Iberia (Portugal, Spain)				9
Sweden	15	58	32	
Switzerland	15	5	61	

two weeks. Sowing, pricking and cold acclimation prior to the freezing test was therefore staggered in a sequential order. Each block consisted of twelve sub-blocks with one plant per accession. Three sub-blocks per accession and block were exposed to one of four different testing temperatures within the range from -5°C, to -17°C (temperatures were adjusted after freeze-testing block 1 and 2 in order to align with the variation in FT). In total, across all blocks, each accession was represented by 48 plants tested at various temperatures. Survival data from the four blocks per accession were pooled before LT50 was estimated using the invest function of the "investr" package in R (Greenwell and Kabban Schubert, 2014).

2.2 Genotyping

The 393 accessions were genotyped at the accession level using pool-GBS (Elshire et al., 2011; Byrne et al., 2013) of DNA extracted using the DNAeasy 96 well kit (QIAGEN) from pools of the first emerging leaf of 200 seedlings per accession. GBS was performed by the LGC Group[2], using i) a combination of *Pst*I and *Mse*I for digestion, ii) a methodological adaptation that involves molecular normalization of read depth across loci within samples and size selection (range 100-250 bp, peak around 175 bp), and iii) paired-end sequencing.

SNP calling and allele frequency estimation was done as described in Keep et al. (2020); details regarding the parameters specifically used in the present study are provided in the Supplementary Material 1. One accession (EUC_TP_107) had more than 80% missing values across SNPs and was discarded from further analyses, leaving 392 accessions. Only biallelic SNPs with a read depth between 30 and 500 were considered. Also, SNPs with more than 5% missing values or overall minor allele frequency (MAF) < 0.05 were removed. This resulted in a set of 20,156 SNPs of which 12,777 (63%) were located on chromosomes and 7,379 on unanchored scaffolds. The number of missing values per marker and per individual, as well as the allele frequency distribution, after filtering, is shown in Supplementary Figures 1, 2. Missing data (0.72% of all sample and SNP genotype calls) were imputed by replacing each missing data point with the mean allele frequency across all accessions per SNP.

Haplotype variants within GBS loci were called and their relative frequencies estimated with the SMAP package (Schaumont et al., 2022) as explained in Ergon et al. (2022), using the SNP data set achieved when removing only SNPs with more than 20% missing data as input. In the process, haplotype variants with an overall MAF < 0.05 were discarded. We then defined a haplotype polymorphism (HTP) as a GBS locus that contains two or more haplotype variants. We removed HTPs with missing values in more than 5% of the samples. This resulted in a data set with allele frequencies of 20,745 haplotype variants in 7,477 HTPs, giving an average of 2.8 variants per HTP. The number of missing values per haplotype and per accession, as well as the haplotype allele frequency distribution after filtering, are shown in Supplementary Figures 3, 4. The data set was then imputed for missing values (0.38%) as explained above for the SNP data set. Overall, 13,359 haplotype variants (64%) were located on chromosomes and 7,389 (36%) on unanchored scaffolds, within 4,833 and 2,644 HTP loci, respectively. The set of 4,833 HTP loci located on the chromosomes

2 www.lgcgroup.com

was used to obtain a measure of marker distribution across the genome in our material.

2.3 Population structure, genomic relationship and linkage disequilibrium (LD)

The population structure among the 392 accessions was investigated with a principal component analysis (PCA) based on the reference allele frequencies of the SNPs after filtration and imputation. SNP- and haplotype-based genomic relationship matrices (GRM) were calculated following the method described by Ashraf et al. (2016). The method is based on 'method 1' in VanRaden (2008) and adapted for the use of allele frequency data. The genotype matrix ($\mathbf{F}_{ij}$, with i indexing the samples and j indexing the markers) was centered by the mean allele frequencies ($\overline{F_j}$), and the resulting genotype matrix $M(M_{ij} = F_{ij} - \overline{F_j})$ was used to compute the GRM $\boldsymbol{G}$, as follows

$$G = \frac{\mathrm{MM}'}{\frac{1}{n}\sum_{j=1}^{m}\hat{p}_j(1-\hat{p}_j)}$$

where m represents the number of markers, $\hat{p}_j$ is the estimated allele frequency of the jth marker, and n represents the total ploidy number, which is the sum of the ploidy number of the parents used to generate the populations as demonstrated by Ashraf et al. (2016). We used n = 16, as suggested by Cericola et al. (2018) for synthetic perennial ryegrass populations. GRMs were calculated with higher n values, and a negligible effect was found in the GWAS results using these different GRMs (data not shown), therefore a value of n = 16 was used throughout this study.

Because we were using population allelic frequencies it was not feasible to assess the true LD between SNP markers in the population. Instead, the LD was estimated as the squared partial correlation between the reference allele frequencies of pairs of SNPs r^2_s; (Lin et al., 2012; Mangin et al., 2012). This measure removes the bias on the correlation between reference allele frequencies at different loci that is due to kinship. r^2_s values were calculated in R (R Core Team, 2020), using the "pcor.shrink" function of the "corpcor" package (Schafer et al., 2021).

2.4 GWAS and candidate gene search

Different SNP-based GWAS models were initially performed using the "rrBLUP" package (Endelman, 2011) with and without correction for kinship (as expressed by the GRM) and/or population structure (as expressed by principal component scores), but none of the models resulted in any significant associations. SNP- and haplotype-based GWAS for FT were therefore performed with the Multi Locus Mixed Model (MLMM), which incorporates multiple markers identified as significantly associated with the phenotypic trait as covariates simultaneously (Segura et al., 2012). We included the GRM in the models in order to account for kinship. The analyses were performed in R with the "mlmm.gwas" package (Bonnafous et al., 2019). The model with the lowest extended Bayesian information criterion (e-BIC) value (Chen and Chen, 2008) was selected as the best model, and SNPs and haplotypes with a p-value lower than the Bonferroni threshold were considered as significant. For each of the significant markers (SNPs and haplotypes) a linear regression between the reference allele frequency and FT was calculated to estimate the allele effect on the phenotype (slope) and the phenotypic variance explained (R^2). Furthermore, to characterize the effect of kinship on the R^2, two different regression models were fitted, one without and one with kinship (as a random genetic effect), with the "lmekin" and the "lm" functions of the "coxme" (Therneau, 2020) and "stats" (R Core Team, 2020) packages, respectively. The variance explained by regression models including all the significant markers simultaneously was also calculated, again with and without including the kinship matrix. All the R^2 values were estimated from the analyses output with the function "r.squaredLR" of the package "MuMin" (Barton, 2020). All the statistical analyses were performed in R (R Core Team, 2020).

Given the limited LD found, the nearest gene to significant SNPs and haplotypes identified by the GWAS were considered as candidate genes. They were identified in the red clover genome Tp2.0 (De Vega et al., 2015) with the gbrowse function available within the Legume Information System[3]. Gene-coding sequences (CDS) were used as query in BLASTn searches (blast.ncbi.nlm.nih.gov) in the *Arabidopsis thaliana* and *Medicago truncatula* genome to identify the most similar genes in these model species.

3 Results

3.1 Variation in freezing tolerance

LT50 among the 392 phenotyped accessions ranged between -11.5°C and -6.0°C with an average of -9.1°C. The average standard error among the estimated LT50-values was 0.5°C. The variation in LT50 across different geographical origins is shown in Figure 1. The most freezing tolerant accessions were found among the Czech, Norwegian and Swedish material while the most susceptible accessions were found among the Iberian material (note that higher LT50 corresponds to lower FT). Even though only nine accessions were assigned to the Iberian group, this group had the largest variation in FT.

3.2 Marker density, LD and genomic relationship among accessions

Since several SNP markers are often clustered within the same GBS locus, the number of HTP loci is more representative of the marker density than the number of SNPs or haplotype variants. The average distance between the 4833 HTP loci that mapped to the 164.2 Mb of the genome assembly assigned to chromosomes was 34

3 www.legumeinfo.org

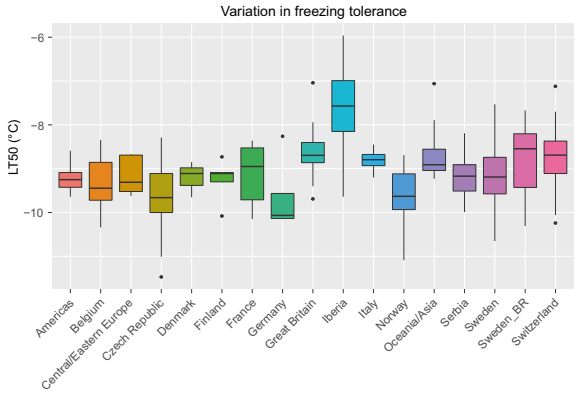


FIGURE 1
Variation in freezing tolerance (LT50, in °C) among 392 red clover accessions from different countries or regions. F2 breeding populations from Sweden (Sweden_BR) were kept as a separate group. See Table 1 for number of accessions belonging to each geographical group.

kb (see Supplementary Figure 5 for an illustration of SNP and HTP density along the chromosomes). The average LD-values (r^2_s) between SNP markers was very low (0.005-0.01) at distances ranging from 1 Mb to 1 kb and increased to 0.16 at distances shorter than 0.5 kb (Table 2; Supplementary Figure 6).

The three first principal components of the PCA (Figure 2) together explained 19% of the genetic variation and indicated that some of the genetic differentiation among the accessions in this study is associated with their origin. Accessions from Iberia were distinct from the other accessions but also diverse. Accessions from some other countries or regions were also relatively diverse (e.g. Great Britain and Czech Republic), while other groups were less diverse (e.g. Switzerland, Norway and Serbia). The genomic relationship values between populations ranged between -0.66 and 6.94 for SNPs and -0.66 and 7.38 for haplotypes, with median values close to zero (-0.03) for both (data not shown). The diagonal elements of the GRMs ranged between 0.25 and 7.84 for the SNPs and between 0.27 and 8.25 for the haplotypes, with median values of 0.92 and 0.97, respectively. The mean diagonal element varied between groups of accessions (Figure 3). Ecotypes from Iberia and Great Britain had the highest values (5-6.5), followed by cultivars from Oceania/Asia and breeding material from Great Britain and Sweden (2.5-3.5), reflecting inbreeding. Almost all landraces had GRM-values below 1, indicating outbreeding, while cultivars from most countries had mean values close to 1 or slightly less.

3.3 SNP- and haplotype-based GWAS

A total of eight and six loci were detected as significantly associated with LT50 by the SNP- and haplotype-based GWAS, respectively (Figure 4). One significant SNP was located within one significant HTP, *i.e.*, a total of thirteen significant GBS loci were identified (Table 3). Eight were located on chromosomes while five were located on unanchored scaffolds. One more haplotype marker on LG1 was just below the Bonferroni threshold of significance, but clearly separated from the expected distribution of p-values in the quantile-quantile (QQ) plot (Figure 4B). This HTP also contained one of the significant SNPs.

When kinship was accounted for, the phenotypic variance explained (R^2) by the GWAS models with the lowest e-BIC values was 0.30 and 0.26 for the SNP- and haplotype-based analysis,

TABLE 2 Linkage disequilibrium (LD) at different distances along the seven red clover chromosomes.

	1Mb	0.5Mb	10kb	5kb	2.5kb	1kb	0.5kb[1]	Mean[2]
Chr1	0.0049 (318)	0.0054 (258)	0.0081 (393)	0.0089 (129)	0.0103 (222)	0.0098 (261)	0.2074 (4446)	0.0048
Chr2	0.0047 (165)	0.0043 (377)	0.0068 (274)	0.0108 (102)	0.0155 (116)	0.0125 (240)	0.1680 (4144)	0.0048
Chr3	0.0043 (333)	0.0049 (251)	0.0070 (351)	0.0077 (287)	0.0087 (310)	0.0171 (278)	0.2018 (4661)	0.0048
Chr4	0.0045 (178)	0.0053 (310)	0.0068 (213)	0.0063 (115)	0.0119 (169)	0.0090 (261)	0.1637 (2810)	0.0050
Chr5	0.0055 (112)	0.0067 (163)	0.0058 (169)	0.0051 (77)	0.0062 (100)	0.0081 (334)	0.1089 (2455)	0.0063
Chr6	0.0048 (166)	0.0054 (113)	0.0061 (283)	0.0073 (154)	0.0100 (142)	0.0103 (249)	0.1385 (2407)	0.0052
Chr7	0.0048 (117)	0.0052 (185)	0.0081 (212)	0.0071 (205)	0.0078 (165)	0.0092 (310)	0.1662 (3161)	0.0049
Mean[3]	0.0048	0.0053	0.0070	0.0076	0.0101	0.0109	0.1649	0.0048

LD was calculated according to Lin et al. (2012) and Mangin et al. (2012), as the average squared partial correlations (r^2_s) between allele frequencies of SNP pairs at given distances ± 0.5 kb. The number of pairs for each chromosome and distance is given in parentheses.
[1]Average LD in the interval 0-0.5 kb distance between pairs of SNP markers. [2] Mean LD across the selected distances for each chromosome.[3] Mean LD across all chromosomes for each selected distance.

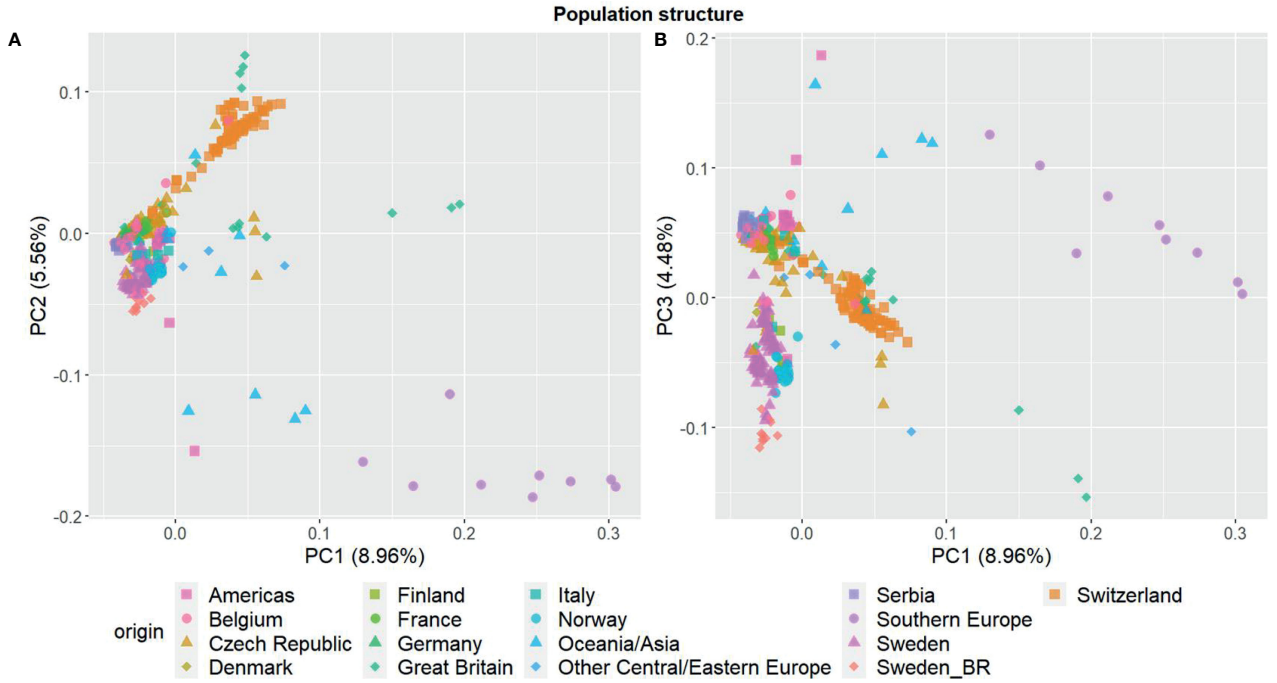


FIGURE 2
Principal component analysis showing the population structure among the 392 red clover (*Trifolium pratense*) accessions based on 20, 156 filtered and imputed SNPs. **(A)** first and second principal component; **(B)** first and third component. Accessions are grouped by country or region of origin. F2 breeding populations from Sweden (Sweden_BR) were kept as a separate group.

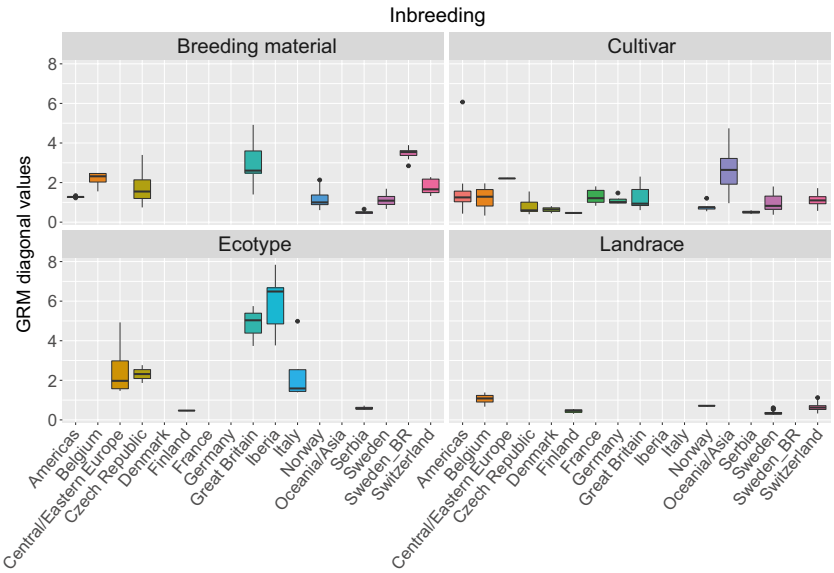


FIGURE 3
Variation in inbreeding expressed as the as values of the diagonal of the genomic relationship matrix (GRM) among 392 red clover accessions from different countries or regions. The GRM was calculated based on SNPs allele frequencies. Values above 1 indicate inbreeding, while values below 1 indicate outbreeding. F2 breeding populations from Sweden (Sweden_BR) were kept as a separate group. See Table 1 for number of accessions belonging to each geographical group.

respectively. When kinship was not accounted for, these values were 0.48 and 0.45 (Table 3A). The values for the individual significant SNPs, when kinship was accounted for ranged from 0.01 to 0.05, while those for the individual haplotypes ranged from 0.02 to 0.05 (Table 3B).

Some of the significant loci were characterized by a relatively strong effect of kinship (Table 3B; Supplementary Figures 7–10). Not surprisingly, these loci were also those whose distribution of allele frequencies was clearly associated with the origin of the plant material (Supplementary Figure 7).

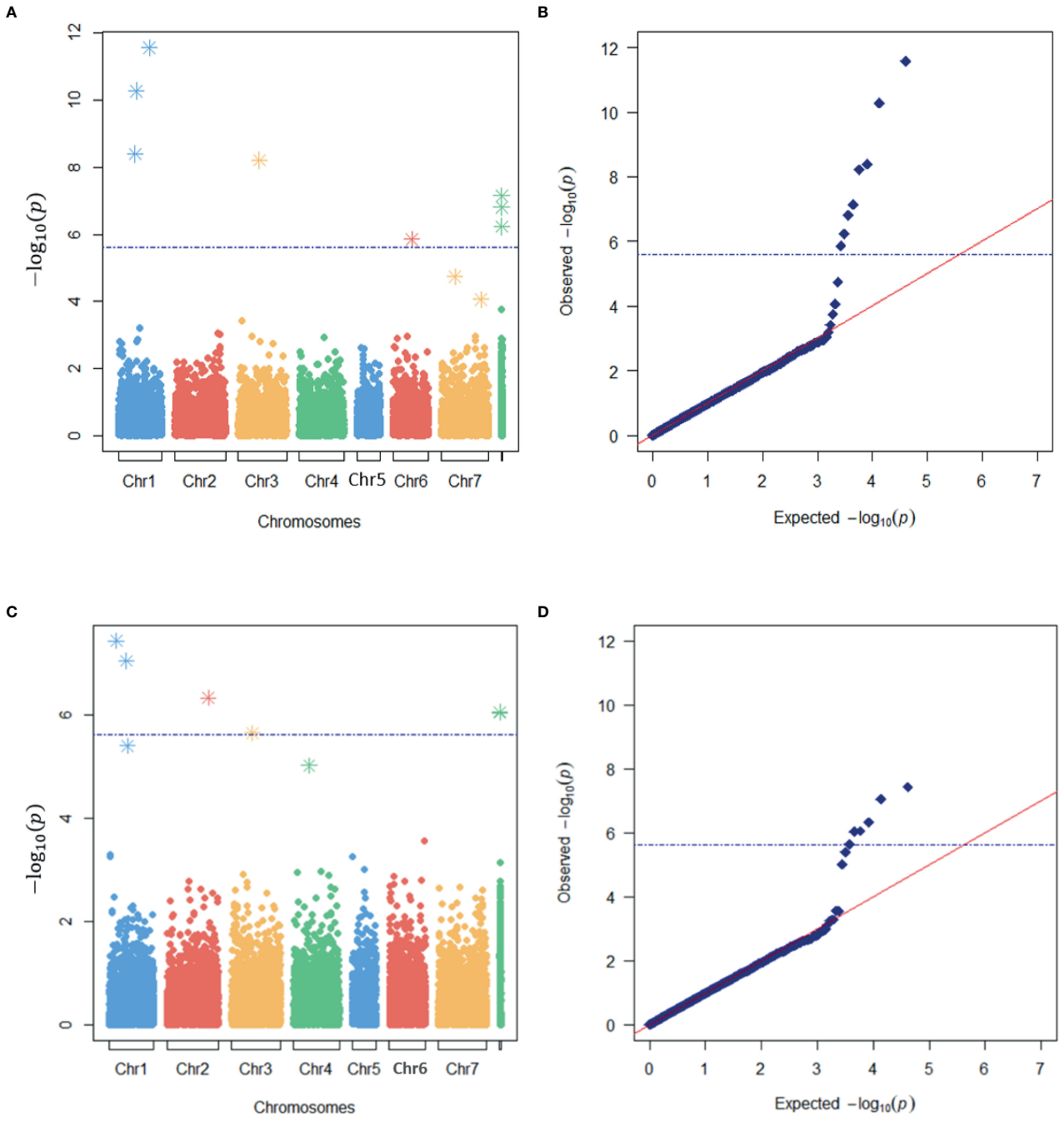


FIGURE 4
Identification of markers associated with freezing tolerance in a genome-wide association study (GWAS) of red clover. Manhattan **(A, C)** and quantile-quantile (QQ) **(B, D)** plots of the expected and observed -log10(p) –values of markers in the SNP-based **(A, B)** and haplotype-based **(C, D)** GWAS models with the lowest e-BIC values. The blue broken horizontal lines represent the Bonferroni threshold while asterisks represent the SNPs or haplotypes used as cofactors in the MLMM analyses.

3.4 Candidate gene search

For nine of the thirteen loci that were significantly associated with FT, the marker was located within a gene, for one locus the marker was located in close proximity (<0.5 kb) of a gene, while for the remaining three loci the marker was located on scaffolds that contained no genes. Functional annotations and the best BLAST hits to the *A. thaliana* or *M. truncatula* genome are reported in Table 4; Supplementary Table 2, respectively.

4 Discussion

4.1 Phenotypic and genotypic variation

There was considerable phenotypic variation in FT among the studied red clover accessions, with LT50 values ranging from –6.0 to –11.5°C. For comparison, 48 Nordic gene bank accessions and 6 Nordic cultivars tested using the same protocol ranged from -6.7 to -13.1°C (Zanotto et al., 2021), while a mixture of northern Fennoscandian

TABLE 3 Phenotypic variance in LT50 explained (R^2) by significant markers identified in GWAS.

A			
Model	R^2 with kinship	R^2 without kinship	Number of loci
SNP-based	0.30	0.48	8
Haplotype-based	0.26	0.45	7[1]
B			
Marker	R^2 with kinship	R^2 without kinship	
SNP-based model			
LG1_10733810	0.05	0.09	
LG1_11608488	0.05	0.19	
scaf_836_20601	0.05	0.10	
scaf_1414_9867	0.04	0.01	
LG1_20019082	0.02	0.08	
scaf_153_142327	0.02	0.00	
LG6_12095124	0.01	0.00	
LG3_13453807	0.01	0.12	
Haplotype-based model			
LG1:11608425-11608558_3[1,2]	0.05	0.19	
LG1:10733751-10733981_31	0.05	0.09	
LG2:26124549-26124674_01	0.04	0.15	
LG1:4339801-4340029_24	0.04	0.03	
LG3:13143867-13144034_06	0.04	0.12	
scaf_670:26744-26998_19	0.03	0.08	
scaf_28637:389-568_03	0.02	0.03	

Regression analysis was performed, using allele frequencies of significant markers as predictors of LT50. The R^2 is given for both SNP- and haplotype-based models that either account for or do not account for kinship among accessions. A, R^2 for the regression models including all markers simultaneously; B, R^2 for the regression models including single markers.
[1]One haplotype had a p-value below the Bonferroni threshold of significance in the GWAS but it was included in further analyses because it was very close to being significant and included one significant SNP. [2]Numbers at the end of each haplotype correspond to the variant number within HTP loci.

TABLE 4 Genes containing significant SNPs and/or haplotypes identified by GWAS of freezing tolerance (FT) in red clover.

SNP and/or haplotype	Annotation	Putative role in FT
LG1_10733810[1] LG1:10733751-10733981	caffeoyl shikimate esterase	Lignin synthesis, secondary cell wall strengthening
LG1_11608488 LG1:11608425-11608558[2]	inositol transporter 1	Transport of compatible solutes, cellular signaling
scaf_836_20601	ATP/DNA binding protein	
LG1:4339801-4340029	cationic peroxidase 1	Oxidative stress response, lignin synthesis
LG3:13143867-13144034	transducin/WD40 repeat-like superfamily protein (NEDD1)	Microtubule organization, cellular signaling and trafficking
LG2:26124549-26124674	chloride channel protein CLC-c	
LG1_20019082	cationic amino acid transporter 1	Amino acid metabolism, accumulation and signaling
scaf_153_142327	NETWORKED 1A	Actin-binding, cellular signaling and trafficking
LG6_12095124	sucrose transport protein SUC8	Sucrose metabolism, accumulation and signaling
LG3_13453807	3-hydroxyisobutyryl-CoA hydrolase 1	Cold stress signaling, valine catabolism

Further details on BLASTn hits in the *A. thaliana* and *M. truncatula* genomes are given in Supplementary Table 2. Three of the scaffolds containing significant markers (not shown) were very short and contained no genes.
[1] SNP located 80 bp outside the gene; [2]haplotype marker had p-value slightly below the Bonferroni threshold of significance.

breeding material given a much longer cold acclimation treatment had an LT50 of –15.1°C (Dalmannsdóttir et al., 2016). Bertrand et al. (2020) found LT50 values of –8 to -10°C in two Canadian cultivars acclimated under semi-natural conditions over the whole autumn and winter, and after recurrent selection for FT it increased to –15°C. The group consisting of nine accessions from the Iberian peninsula contained the most freezing sensitive accessions, but also some that were relatively freezing tolerant. According to the PCA this group was also both distinct from and more diverse than the other groups.

Red clover is known to be a genetically and phenotypically diverse crop (Nay et al., 2023). Ergon et al. (2022), studying GBS loci of a single cultivar, found an average of 3.6 haplotype variants per HTP, both when *Pst*I and *ApeK*I were used as restriction enzymes. In the present study, using similar calling and filtering criteria, we found only 2.8 haplotype variants per HTP on average. The true haplotype diversity is likely to be higher, but as the sequencing depth and consequently the SNP density was much lower in the present study (~20,000 vs. ~100,000 SNPs), and because a haplotype must be present in many accessions to reach a MAF > 0.05, a large number of rare polymorphisms are likely to be either not detected or filtered out. This implies that we here explore the genetic variation defined by the most common markers and are likely to miss variation that is due to less common markers. The extremely low LD that we observed is typical in outbreeding species with a high genetic diversity and in line with those found previously in studies considering wide and diversified plant material of red clover (Jones et al., 2020).

The expected diagonal values of the GRMs (relationship of each accession to itself) are 1.0 plus the inbreeding coefficient (Ashraf et al., 2016). The higher average values for the diagonal element of the GRMs that we found in some groups of accessions may be due to some level of inbreeding within those accessions. Some of the accessions used in this study are cultivars and breeding populations originating from a limited number of parents. Often these parents are chosen from already released cultivars within breeding programs which may share the same allele at several loci, and therefore have a reduced level of diversity within these loci. In this study, this was probably the case for breeding material from Sweden, Belgium and Great Britain and cultivars from Asia which all had mean GRM diagonal values exceeding 2. Another likely cause for the inbreeding that we observe in breeding material is the selection for specific traits, which also results in lower diversity and heterozygosity. Interestingly, high mean GRM diagonal values were also found among ecotypes from Iberia and Great Britain and to a lesser degree among those from Czech Republic, while ecotypes from Serbia and Finland had values well below 1, indicating outbreeding. Ecotypes that are located within restricted and isolated areas may have an increased level of inbreeding while other ecotypes may be part of large outbreeding populations.

4.2 GWAS and the proportion of phenotypic variation explained by significant loci

Population allele frequency data from pools have been successfully used in identifying loci associated with specific traits and for the development of genomic prediction models in perennial and Italian ryegrass (*L. multiflorum*) and red clover (Fè et al., 2015; Fè et al., 2016; Guo et al., 2018; Knorst et al., 2019; Ergon et al., 2019; Keep et al., 2020; Ergon et al., 2022; Frey et al., 2022), but this is to our knowledge the first study that uses GBS-generated allele frequency data from population pools to conduct haplotype-based GWAS in such outbred forage species. Our analyses confirmed that haplotype information can reveal associations that are not detectable by single SNPs and *vice versa*, as previously shown by Lorenz et al. (2010); Hamblin and Jannink (2011); Bekele et al. (2018) and Ergon et al. (2022).

With the strong LD decay observed, we would need at least a million polymorphic loci or so to be able to detect all loci with an effect on FT. Thus, with the ~20 000 SNPs in around 5000 HTP loci in this study, only a fraction of the loci can be expected to be found. It may therefore be surprising that as much as 30% and 26% of the phenotypic variation could be explained by the loci identified in the SNP- and haplotype-based models, respectively. A likely explanation for this is that variation in these loci correlates with variation in a lot of other loci controlling FT. Many other loci are likely to also have been under selection towards better FT in populations frequently exposed to freezing temperatures, and the loci we identified in the current study may therefore also explain some of the phenotypic variance caused by these. Since this study is performed at population level, it will only capture variation between, and not within, accessions.

For the markers LG3_13453807, LG1_11608488 and LG1:11608425-11608558, accessions from Iberia (characterized by low average FT) had a clearly different average allele frequency than those from the other countries, suggesting that the former material has a very low frequency for an allele which likely provides a better level of FT. Interestingly, both GBS loci which were shared across SNPs and haplotype-based GWAS (LG1_11608488 and LG1:11608425-11608558; LG1_10733810 and LG1:10733751-10733981) were among the loci with the highest R^2 from a regression model with LT50, which show that these associations are particularly strong. The same loci also had a large difference between from regressions that did and did not include kinship, being therefore strongly associated with the relatedness among accessions. However, even after accounting for the effect of kinship these loci could explain a significantly large proportion of the phenotypic variation.

4.3 Candidate genes

Because of the very rapid LD decay we considered only the closest gene to significant loci as candidate genes, since the likelihood of markers located further away being in LD with FT-associated alleles is low. Ten of the GBS loci significantly associated with variation of FT that mapped within gene sequences, and two more loci were found very close to a gene (< 0.5kb).

LG1_10733810, one of the three markers which could explain the largest phenotypic variance (5% after correcting for kinship), is located in a gene encoding caffeoyl shikimate esterase. This enzyme is essential for lignin synthesis and secondary cell wall formation in *M. truncatula* (Ha et al., 2016), which may play a role in cold acclimation and freezing tolerance through an effect of cell wall strengthening (Sasidharan et al., 2011). Additionally, some

secondary metabolites synthesized through the shikimate pathway were reported to be involved in cold acclimation in *A. thaliana* and red clover (Schulz et al., 2016; Zanotto et al., 2023). Allele frequencies of caffeoyl shikimate esterase were moderately associated with kinship. The SNP allele associated with improved FT was mainly found among Nordic and Czech accessions while it was almost absent among Iberian accessions.

LG1_11608488 is located in a gene encoding an inositol transporter 1 (INT1) protein. INT1 transports *myo*-inositol from the vacuole into the cytoplasm in *A. thaliana* (Schneider et al., 2008; Strobl et al., 2018). Inositol transporters and *myo*-inositol have important roles in various signaling pathways in plants (Valluru and Van den Ende, 2011; Zhou et al., 2021). The methylated derivatives of myo-inositol (e.g. pinitol) accumulate in different plant species in response to abiotic stress, and is thought to act as an osmo-regulator under abiotic stresses (Sengupta et al., 2008; Sengupta et al., 2012). Pinitol concentration in red clover crowns increases during cold acclimation, and to higher levels in freezing tolerant material of different genetic backgrounds, supporting a role in freezing tolerance in red clover (Bertrand et al., 2020; Zanotto et al., 2023). It also increases in red clover in response to drought (Yates et al, 2014). Allele frequencies of this marker were strongly associated with kinship and markedly different in Iberian vs other groups of accessions. However, after correcting for kinship it still explained 5% of the phenotypic variation in FT in our study.

One candidate gene encodes a peroxidase, which may be involved in both lignin synthesis and/or oxidative stress response. Plants exposed to low temperatures combined with light are subjected to oxidative stress because of an energy imbalance in the photosystem (Huner et al., 1993). Several of the other identified candidate genes encode proteins that are involved in actin binding, microtubule organization, amino acid and carbohydrate metabolism.

4.4 Concluding remarks

Considering the current and future need for a more self-sufficient and sustainable production of plant-based protein in Europe, the availability of red clover cultivars with improved persistency is of primary importance to guarantee higher and more stable yields for this species. We have here described the variation in FT across a broad European collection of red clover accessions and identified loci and possible candidate genes associated with this trait. The genetic polymorphisms that were identified may, after validation in relevant germplasm, be transformed into molecular markers and used in selection for this trait. Furthermore, the integration of these markers with markers related to disease resistance and other traits of interest in a genomic selection model will contribute to a powerful tool to accelerate breeding of improved red clover.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NMBU Open Research Data https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/WF1AGU. The GBS reads are available at NCBI under project number PRJNA842231.

Author contributions

CG and RK were responsible for establishing a comprehensive red clover germplasm collection and for collecting and distributing red clover seeds. LS was responsible for DNA extractions and genotyping. LS and TR designed the genotyping strategy. TR performed SNP and haplotype calling. ÅE designed the phenotyping strategy and SZ performed the phenotyping. SZ performed all analyses of phenotypic and marker data with supervision from MP and ÅE. SZ wrote the paper with supervision from ÅE. All authors contributed to the article and approved the submitted version.

Funding

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement no. 727312 (EUCLEG).

Acknowledgments

We acknowledge the technical assistance from Øyvind Jørgensen, Liv Berge, Martin Paliocha and Sabine van Glabeke. We also thank Philippe Barre, Steven Maenhout, Lea Frey and all the other members of the EUCLEG consortium that helped with productive discussions and suggestions.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1189662/full#supplementary-material

References

Abberton, M. T., and Marshall, A. H. (2005). Progress in breeding perennial clovers for temperate agriculture. *J. Agric. Sci.* 143, 117–135. doi: 10.1017/S0021859605005101

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., and Marshall, A. H. (2015). Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* 34, 327–380. doi: 10.1080/07352689.2014.898462

Annicchiarico, P., Collins, R. P., Fornasier, F., and Rhodes, I. (2001). Variation in cold tolerance and spring growth among Italian white clover populations. *Euphytica* 122, 407–416. doi: 10.1023/A:1012918100232

Annicchiarico, P., and Pagnotta, M. A. (2012). Agronomic value and adaptation across climatically contrasting environments of Italian red clover landraces and natural populations. *Grass Forage Sci.* 67, 597–605. doi: 10.1111/j.1365-2494.2012.00887.x

Ashraf, B. H., Byrne, S., Fé, D., Czaban, A., Asp, T., Pedersen, M. G., et al. (2016). Estimating genomic heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-by-sequencing. *Theor. Appl. Genet.* 129, 45–52. doi: 10.1007/s00122-015-2607-9

Barton, K. (2020) *MuMIn: multi-model inference. r package version 1.43.17*. Available at: https://CRAN.R-project.org/package=MuMIn.

Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J., and Tinker, N. A. (2018). Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* 16, 1452–1463. doi: 10.1111/pbi.12888

Bélanger, G., Castonguay, Y., Bertrand, A., Dhont, C., Rochette, P., Couture, L., et al. (2006). Winter damage to perennial forage crops in eastern Canada: causes, mitigation, and prediction. *Can. J. Plant Sci.* 86, 33–47. doi: 10.4141/P04-171

Bertrand, A., Bipfubusa, M., Castonguay, Y., Rocher, S., Szopinska-Morawska, A., Papadopoulos, Y., et al. (2016). A proteome analysis of freezing tolerance in red clover (*Trifolium pratense* l.). *BMC Plant Biol.* 16, 65. doi: 10.1186/s12870-016-0751-2

Bertrand, A., Rocher, S., Claessens, A., Bipfubusa, M., Papadopoulos, Y., and Castonguay, Y. (2020). Biochemical and molecular responses during overwintering of red clover populations recurrently selected for improved freezing tolerance. *Plant Sci.* 292, 110388. doi: 10.1016/j.plantsci.2019.110388

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., et al. (2017). Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PloS One* 12, e0169234. doi: 10.1371/journal.pone.0169234

Bickhart, D. M., Koch, L. M., Smith, T. P. L., Riday, H., and Sullivan, M. L. (2022). Chromosome-scale assembly of the highly heterozygous genome of red clover (*Trifolium pratense* l.), an allogamous forage crop species. *Gigabyte* 2022, 1–13. doi: 10.46471/gigabyte.42

Bonnafous, F., Duhnen, A., Gody, L., et al. (2019). *mlmm.gwas: pipeline for GWAS using MLMM*.Version R package version 1.0.6. Available at: https://CRAN.R-project.org/package=mlmm.gwas.

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations *via* genotyping by sequencing. *PloS One* 8, e57438. doi: 10.1371/journal.pone.0057438

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., et al. (2018). Optimized use of low-depth genotyping-by-Sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00369

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95, 759–771. doi: 10.1093/biomet/asn034

Dalmannsdóttir, S., Rapacz, M., Jørgensen, M., Østrem, L., Larsen, A., Rødven, R., et al. (2016). Temperature before cold acclimation affects cold tolerance and photoacclimation in Timothy (*Phleum pratense* l.), perennial ryegrass (*Lolium perenne* l.) and red clover (*Trifolium pratense* l.). *J. Agron. Crop Sci.* 202, 320–330. doi: 10.1111/jac.12149

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., et al. (2015). Red clover (*Trifolium pratense* l.) draft genome provides a platform for trait improvement. *Sci. Rep.* 5, 17394. doi: 10.1038/srep17394

de Visser, C. L. M., Schreuder, R., and Stoddard, F. (2014). The EU's dependency on soya bean import for the animal feed industry and potential for EU produced alternatives. *OCL* 21, D407. doi: 10.1051/ocl/2014021

Dias, P. M. B., Julier, B., Sampoux, J.-P., Barre, P., and Dall'Agnol, M. (2008). Genetic diversity in red clover (*Trifolium pratense* l.) revealed by morphological and microsatellite (SSR) markers. *Euphytica* 160, 189–205. doi: 10.1007/s10681-007-9534-z

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

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *Plant Genome* 4, 250–255.

Ergon, Å., Milvang, Ø.W., Skøt, L., and Ruttink, T. (2022). Identification of loci controlling timing of stem elongation in red clover using genotyping by sequencing of pooled phenotypic extremes. *Mol. Genet. Genomics* 297, 1587–1600. doi: 10.1007/s00438-022-01942-x

Ergon, Å., Skøt, L., Sæther, V. E., and Rognli, O. A. (2019). Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (*Trifolium pratense* l.). *Front. Plant Sci.* 10, 1–12. doi: 10.3389/fpls.2019.00718

Fè, D., Ashraf, B. H., Pedersen, M. G., Janss, L., Byrne, S., Roulund, N., et al. (2016). Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *Plant Genome* 9, 1–12. doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B. H., Pedersen, M. G., et al. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics* 16, 921. doi: 10.1186/s12864-015-2163-3

Frame, J., Charlton, J. F. L., and Laidlaw, A. S. (1997). *Temperate forage legumes* (Wallingford, UK: CABI International).

Frey, L. A., Vleugels, T., Ruttink, T., Schubiger, F. X., Pegard, M., Skøt, L., et al. (2022). Phenotypic variation and quantitative trait loci for resistance to southern anthracnose and clover rot in red clover. *Theor. Appl. Genet.* 135, 4337–4349. doi: 10.1101/2022.05.23.493028

Greene, S. L., Gritsenko, M., and Vandemark, G. (2004). Relating morphologic and RAPD marker variation to collection site environment in wild populations of red clover (*Trifolium pratense* l.). *Genet. Resour. Crop Evol.* 51, 643–653. doi: 10.1023/B:GRES.0000024655.48989.ab

Greenwell, B. ,. M., and Kabban Schubert, C. ,. M. (2014). Investr: an r package for inverse estimation. *R J.* 6, 90. doi: 10.32614/RJ-2014-009

Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., et al. (2018). Genomic prediction in tetraploid ryegrass using allele frequencies based on genotyping by sequencing. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01165

Gusta, L. V., O'Connor, B. J., Gao, Y.-P., and Jana, S. (2001). A re-evaluation of controlled freeze-tests and controlled environment hardening conditions to estimate the winter survival potential of hardy winter wheats. *Can. J. Plant Sci.* 81, 241–246. doi: 10.4141/P00-068

Ha, C. M., Escamilla-Trevino, L., Yarce, J. C. S., Kim, H., Ralph, J., Chen, F., et al. (2016). An essential role of caffeoyl shikimate esterase in monolignol biosynthesis in m *edicago truncatula*. *Plant J.* 86, 363–375. doi: 10.1111/tpj.13177

Hamblin, M. T., and Jannink, J.-L. (2011). Factors affecting the power of haplotype markers in association studies. *Plant Genome* 4, 145–153. doi: 10.3835/plantgenome2011.03.0008

Helgadóttir, Á., Frankow-Lindberg, B. E., Seppänen, M. M., Søegaard, K., and Østrem, L. (2014). European Grasslands overview: Nordic region. in Hopkins et al (Eds.), EGF at 50: the future of European grasslands. *Grassland Sci. Europe* 19, 15–28.

Herrmann, D., Boller, B., Studer, B., Widmer, F., and Kölliker, R. (2008). Improving persistence in red clover: insights from QTL analysis and comparative phenotypic evaluation. *Crop Sci.* 48, 269–277. doi: 10.2135/cropsci2007.03.0143

Hulke, B. S., Watkins, E., Wyse, D. L., and Ehlke, N. J. (2008). Freezing tolerance of selected perennial ryegrass (*Lolium perenne* l.) accessions and its association with field winterhardiness and turf traits. *Euphytica* 163, 131–141. doi: 10.1007/s10681-007-9631-z

Huner, N. P. A., Öquist, G., Hurry, V. M., Krol, M., Falk, S., and Griffith, M. (1993). Photosynthesis, photoinhibition and low temperature acclimation in cold tolerant plants. *Photosynthetic Resour.* 37, 19–39. doi: 10.1007/BF02185436

Inostroza, L., Bhakta, M., Acuña, H., Vásquez, C., Ibáñez, J., Tapia, G., et al. (2018). Understanding the complexity of cold tolerance in white clover using temperature gradient locations and a GWAS approach. *Plant Genome* 11, 1–14. doi: 10.3835/plantgenome2017.11.0096

Jensen, E. S., Peoples, M. B., Boddey, R. M., Gresshoff, P. M., Hauggaard-Nielsen, H., J.R. Alves, B., et al. (2012). Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries. *A review. Agron. Sustain. Dev.* 32, 329–364. doi: 10.1007/s13593-011-0056-7

Jones, C., De Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., et al. (2020). Population structure and genetic diversity in red clover (*Trifolium pratense* l.) germplasm. *Sci. Rep.* 10, 8364. doi: 10.1038/s41598-020-64989-z

Keep, T., Sampoux, J.-P., Blanco-Pastor, J. L., Dehmer, K. J., Hegarty, M. J., Ledauphin, T., et al. (2020). High-throughput genome-wide genotyping to optimize the use of natural genetic resources in the grassland species perennial ryegrass (*Lolium perenne* l.). *G3: Genes Genomes Genet.* 10, 3347–3364. doi: 10.1534/g3.120.401491

Klimenko, I., Razgulayeva, N., Gau, M., Okumura, K., Nakaya, A., Tabata, S., et al. (2010). Mapping candidate QTLs related to plant persistency in red clover. *Theor. Appl. Genet.* 120, 1253–1263. doi: 10.1007/s00122-009-1253-5

Knorst, V., Byrne, S., Yates, S., Asp, T., Widmer, F., Studer, B., et al. (2019). Pooled DNA sequencing to identify SNPs associated with a major QTL for bacterial wilt resistance in Italian ryegrass (Lolium multiflorum Lam.). *Theor. Appl. Genet.* 132, 947–958. doi: 10.1007/s00122-018-3250-z

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

Kouamé, C. N., and Quesenberry, K. H. (1993). Cluster analysis of a world collection of red clover germplasm. *Genet. Resour. Crop Evol.* 40, 39–47. doi: 10.1007/BF00053463

Lin, C.-Y., Xing, G., and Xing, C. (2012). Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity* 109, 401–402. doi: 10.1038/hdy.2012.54

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

Lorenz, A. J., Hamblin, M. T., and Jannink, J.-L. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PloS One* 5, e14079. doi: 10.1371/journal.pone.0014079

Lüscher, A., Mueller-Harvey, I., Soussana, J. F., Rees, R. M., and Peyraud, J. L. (2014). Potential of legume-based grassland–livestock systems in Europe: a review. *Grass Forage Sci.* 69, 206–228. doi: 10.1111/gfs.12124

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108, 285–291. doi: 10.1038/hdy.2011.73

Marshall, A. H., Collins, R. P., Vale, J., and Lowe, M. (2017). Improved persistence of red clover (*Trifolium pratense* l.) increases the protein supplied by red clover/grass swards grown over four harvest years. *Eur. J. Agron.* 89, 38–45. doi: 10.1016/j.eja.2017.06.006

Nay, M. M., Grieder, C., Frey, L. A., Amdahl, H., Radovic, J., Jaluvka, L., et al. (2023). Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1128823

Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2021). Insights into the genetic diversity of Nordic red clover (*Trifolium pratense*) revealed by SeqSNP-based genic markers. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.748750

R Core Team (2020). *RStudio: integrated development for r* (Boston, MA: RStudio, PBC). Available at: http://www.rstudio.com/.

Reckling, M., Bergkvist, G., Watson, C. A., Stoddard, F. L., Zander, P. M., Walker, R. L., et al. (2016). Trade-offs between economic and environmental impacts of introducing legumes into cropping systems. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00669

Sasidharan, R., Voesenek, L. A., and Pierik, R. (2011). Cell wall modifying proteins mediate plant acclimatization to biotic and abiotic stresses. *Crit. Rev. Plant Sci.* 30, 548–562. doi: 10.1080/07352689.2011.615706

Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., et al. (2005). Comprehensive structural analysis of the genome of red clover (*Trifolium pratense* l.). *DNA Res.* 12, 301–364. doi: 10.1093/dnares/dsi018

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D., and Strimmer., K. (2021) *Corpcor: efficient estimation of covariance and (Partial) correlation. r package version 1.6.10*. Available at: https://CRAN.R-project.org/package=corpcor.

Schaumont, D., Veeckman, E., der Jeugt, F. V., Haegeman, A., van Glabeke, S., Bawin, Y., et al. (2022). Stack mapping anchor points (SMAP): a versatile suite of tools for read-backed haplotyping. *BioRxiv*. doi: 10.1101/2022.03.10.483555

Schneider, S., Beyhl, D., Hedrich, R., and Sauer, N. (2008). Functional and physiological characterization of arabidopsis INOSITOL TRANSPORTER1, a novel tonoplast-localized transporter for myo-inositol. *Plant Cell* 20, 1073–1087. doi: 10.1105/tpc.107.055632

Schulz, E., Tohge, T., Zuther, E., Fernie, A. R., and Hincha, D. K. (2016). Flavonoids are determinants of freezing tolerance and cold acclimation in *Arabidopsis thaliana*. *Sci. Rep.* 6, 34027. doi: 10.1038/srep34027

Segura, V., Vilhjalmsson, B. J., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830.

Sengupta, S., Mukherjee, S., Goswami, L., Sangma, S., Mukherjee, A., Mukherjee, R., et al. (2012). Manipulation of inositol metabolism for improved plant survival under stress: a "network engineering approach". *J. Plant Biochem. Biotechnol.* 21, 15–23. doi: 10.1007/s13562-012-0132-3

Sengupta, S., Patra, B., Ray, S., and Majumder, A. L. (2008). Inositol methyl tranferase from a halophytic wild rice, *Porteresia coarctata* roxb. (Tateoka): regulation of pinitol synthesis under abiotic stress. *Plant Cell Environ.* 31, 1442–1459. doi: 10.1111/j.1365-3040.2008.01850.x

Strobl, S. M., Kischka, D., Heilmann, I., Mouille, G., and Schneider, S. (2018). The tonoplastic inositol transporter INT1 from *Arabidopsis thaliana* impacts cell elongation in a sucrose-dependent way. *Front. Plant Sci.* 9, 1657.

Therneau, T. M. (2020) *Mixed effects cox models. version r package version 2.2-16URL*. Available at: https://CRAN.R-project.org/package=coxme.

Valluru, R., and Van den Ende, W. (2011). Myo-inositol and beyond – emerging networks under stress. *Plant Sci.* 181, 387–400. doi: 10.1016/j.plantsci.2011.07.009

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

Voisin, A.-S., Guéguen, J., Huyghe, C., Jeuffroy, M.-H., Magrini, M.-B., Meynard, J.-M., et al. (2014). Legumes for feed, food, biomaterials and bioenergy in Europe: a review. *Agron. Sustain. Dev.* 34, 361–380. doi: 10.1007/s13593-013-0189-y

Waalen, W. M., Tanino, K. K., Olsen, J. E., Eltun, R., Rognli, O. A., and Gusta, L. V. (2011). Freezing tolerance of winter canola cultivars is best revealed by a prolonged freeze test. *Crop Sci.* 51, 1988–1996. doi: 10.2135/cropsci2011.02.0098

Waldron, B. L., Ehlke, N. J., Vellekson, D. J., and White, D. B. (1998). Controlled freezing as a indirect selection method for field winterhardiness in turf-type perennial ryegrass. *Crop Sci.* 38, 811–816. doi: 10.2135/cropsci1998.0011183X003800030032x

Yates, S. A., Swain, M. T., Hegarty, M. J., Chernukin, I., Lowe, M., Allison, G. G., et al. (2014). *De novo* Assembly of red clover transcriptome based on RNA-seq data provides insight into drought response, gene discovery and marker identification. *BMC Genomics* 15, 453.

Zanotto, S., Bertrand, A., Purves, R. W., Olsen, J. E., Elessawy, F. M., and Ergon, Å. (2023). Biochemical changes after cold acclimation in Nordic red clover (*Trifolium pratense* L.) accessions with contrasting levels of freezing tolerance. *Manuscript*.

Zanotto, S., Palmé, A., Helgadóttir, Á., Daugstad, K., Isolahti, M., Öhlund, L., et al. (2021). Trait characterization of genetic resources reveals useful variation for the improvement of cultivated Nordic red clover. *J. Agron. Crop Sci.* 207, 492–503. doi: 10.1111/jac.12487

Zhou, Y., Foka, I. C. K., and Li, H. (2021). Tonoplast inositol transporters: roles in plant abiotic stress response and crosstalk with other signals. *J. Plant Physiol.* 271, 153660. doi: 10.20944/preprints202106.0515.v1

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 05 July 2023
DOI 10.3389/fpls.2023.1196134

Check for updates

OPEN ACCESS

EDITED BY
Deyue Yu,
Nanjing Agricultural University, China

REVIEWED BY
Ali M. Missaoui,
University of Georgia, United States
Quanzhen Wang,
Northwest A&F University, China

*CORRESPONDENCE
Marie Pégard
marie.pegard@inrae.fr

RECEIVED 29 March 2023
ACCEPTED 30 May 2023
PUBLISHED 05 July 2023

CITATION
Pégard M, Barre P, Delaunay S, Surault F, Karagić D, Milić D, Zorić M, Ruttink T and Julier B (2023) Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits.
Front. Plant Sci. 14:1196134.
doi: 10.3389/fpls.2023.1196134

COPYRIGHT
© 2023 Pégard, Barre, Delaunay, Surault, Karagić, Milić, Zorić, Ruttink and Julier. 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.

Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits

Marie Pégard[1*], Philippe Barre[1], Sabrina Delaunay[1], Fabien Surault[1], Djura Karagić[2], Dragan Milić[3], Miroslav Zorić[2], Tom Ruttink[4] and Bernadette Julier[1]

[1]INRAE P3F, Lusignan, France, [2]Login EKO doo, Bulevar Zorana Đinđića 125, Novi Beograd, Serbia, [3]International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya, [4]ILVO Plant Science Unit, Melle, Belgium

China's and Europe's dependence on imported protein is a threat to the food self-sufficiency of these regions. It could be solved by growing more legumes, including alfalfa that is the highest protein producer under temperate climate. To create productive and high-value varieties, the use of large genetic diversity combined with genomic evaluation could improve current breeding programs. To study alfalfa diversity, we have used a set of 395 alfalfa accessions (i.e. populations), mainly from Europe, North and South America and China, with fall dormancy ranging from 3 to 7 on a scale of 11. Five breeders provided materials (617 accessions) that were compared to the 400 accessions. All accessions were genotyped using Genotyping-by-Sequencing (GBS) to obtain SNP allele frequency. These genomic data were used to describe genetic diversity and identify genetic groups. The accessions were phenotyped for phenology traits (fall dormancy and flowering date) at two locations (Lusignan in France, Novi Sad in Serbia) from 2018 to 2021. The QTL were detected by a Multi-Locus Mixed Model (mlmm). Subsequently, the quality of the genomic prediction for each trait was assessed. Cross-validation was used to assess the quality of prediction by testing GBLUP, Bayesian Ridge Regression (BRR), and Bayesian Lasso methods. A genetic structure with seven groups was found. Most of these groups were related to the geographical origin of the accessions and showed that European and American material is genetically distinct from Chinese material. Several QTL associated with fall dormancy were found and most of these were linked to genes. In our study, the infinitesimal methods showed a higher prediction quality than the Bayesian Lasso, and the genomic prediction achieved high (>0.75) predicting abilities in some cases. Our results are encouraging for alfalfa breeding by showing that it is possible to achieve high genomic prediction quality.

KEYWORDS

alfalfa, genetic diversity, phenology, GWAS, genomic prediction

1 Introduction

Alfalfa (*Medicago sativa*) is a major legume forage species grown worldwide. Its positive impact on protein autonomy through atmospheric nitrogen fixation and the environmental services it provides at the plot and rotation level have become increasingly important in recent years (Julier et al., 2017). This suggests that alfalfa should be grown on expanding areas (Poux and Aubert, 2018; Martin et al., 2020). Breeding is a lever to improved forage yield and quality as well as tolerance to biotic and abiotic stresses (Rubiales et al., 2021). Two aspects are critical: the use of genetic variation in which the selection is applied and the identification of the best genotypes, which will contribute to create the next generation.

From its origin in the Middle East, the domestication and breeding history of alfalfa in the Western world (Michaud et al., 1988; Lesins and Lesins, 2012) as well as in Asia (Basigalup et al., 2014) have previously been described (Small, 2011). Two main subspecies have been described: ssp *sativa* with purple flowers, a tap-root and coiled pods; and ssp *falcata* with yellow flowers, fasciculate roots and silk-shape pods. Even if the two subspecies can be intercrossed, the cultivated varieties mostly relate to the ssp *sativa* type with various levels of introgression with ssp *falcata*, which has conferred cold resistance and variegated flower colour. Studies based on molecular markers have revealed the genetic relationships among accessions (Li, 2013), the reduction of diversity in cultivated compared to wild populations (Muller et al., 2006) and described the genetic distance between Western and Asian accessions (Qiang et al., 2015). Markers have confirmed the huge within-accession diversity (Flajoulot et al., 2005) already observed with phenotypic traits (Julier et al., 2000). These studies have been conducted on small sets of diversity and/or with small sets of markers, and in several cases, the accessions were represented by a single individual. Thus, the description of alfalfa cultivated material is not yet optimal. At present, the use of within and among-accession diversity in breeding programs may be limited because of restricted access to this diversity and the fear that foreign/distant accessions do not comply with the breeders' ideotype (Annicchiarico et al., 2015a). As with other species, a better knowledge of the genetic diversity of alfalfa could broaden the genetic basis of breeding programs and thus increase the potential for genetic gain.

Alfalfa breeding programs still rely on phenotypic selection in which the genetic value of a plant is evaluated directly on the plant (mass selection) or more accurately through its progeny under field or test conditions. In other species, numerous breeding programs have benefited from the advances in high-throughput genotyping technologies (Rasheed et al., 2017). With a large number of markers, it becomes possible to analyse and manage the genetic diversity, to identify markers involved in trait variation (Genome Wide Association Study) (Flint-Garcia et al., 2003) and to create genomic prediction equations to predict the genetic value with the marker information (Meuwissen et al., 2001). Known as genomic selection (GS), this last method has proved its efficiency in plant breeding (Crossa et al., 2017). A successful implementation of GS in breeders' plant material requires considering certain parameters. Firstly, the linkage disequilibrium and the effective size of the population affect the number of markers needed to reach an accurate prediction: the number of required markers increases if linkage disequilibrium decays at shorter distance (Grattapaglia and Resende, 2011; Wientjes et al., 2013). Secondly, the composition of the population used to train (training population) the prediction model must be considered (Lorenz and Smith, 2015; Tayeh et al., 2015; Pégard et al., 2021). The training population must be representative of the selection candidates and several authors have studied the way in which it can be optimised (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015). Thirdly, the trait genetic architecture will affect the prediction performance of the statistical methods (Wimmer et al., 2013), but this parameter is usually unknown and difficult to assess, requiring testing of multiple methods.

Genotyping alfalfa, an autotetraploid and allogamous species, has taken a leap forward with the use of Genotyping-by-Sequencing methodology (GBS), as described on heterozygous diploid species (Elshire et al., 2011). In the first attempts on alfalfa, marker calling was based on *de novo* assembly of reads without a reference genome (Li et al., 2014; Annicchiarico et al., 2015b; Biazzi et al., 2017) or on a mapping of the reads on the reference genome sequence of the related model species *Medicago truncatula* (Julier et al., 2018). With the recent release of tetraploid alfalfa reference genome sequences (Carrère et al., 2020; Chen et al., 2020; Shen et al., 2020; Long et al., 2022), more reads are expected to be mapped and the markers are physically positioned on the alfalfa genome. This GBS methodology offers a high throughput genotyping tool that is convenient for most of the genetic studies, at the individual level with the allele dosage determination as well as the population level with the allele frequency determination (Julier et al., 2018).

Linkage disequilibrium decays at short distance in allogamous species (Flint-Garcia et al., 2003) and this also applies to alfalfa (Herrmann et al., 2010). The candidate gene approach is appropriate for association mapping (Herrmann et al., 2010), but requires previous knowledge of relevant candidate genes. Conversely, performing a genome-wide association study (GWAS) with reduced representation libraries such as GBS, requires sequencing at many loci that are evenly spread across the genome and at high density. Such QTL have been obtained on diploid alfalfa (Sakiroglu and Brummer, 2017) or cultivated tetraploid alfalfa (Biazzi et al., 2017). Genomic prediction (Meuwissen et al., 2001) has been tested on alfalfa (Annicchiarico et al., 2015b; Li et al., 2015b; Biazzi et al., 2017; Medina et al., 2021; Andrade et al., 2022), showing promising predicting ability around 30% for forage yield and quality traits. Higher predicting ability could probably be obtained by using a larger population size and/or a lower percentage of missing data in the genotyping dataset and/or more markers at more loci.

In alfalfa, as in other species, phenology traits are the major drivers of climate adaptation. Flowers are formed at the leaf axillaries (Teuber and Brick, 1988) and do not hamper stem elongation. In contrast with most cultivated species, flowering date is not a component of forage yield in alfalfa, and it is not even scored by breeders. However, the beginning of flowering stage is used as an indicator for the cutting date since it indicates a good

compromise between forage yield, quality and persistence. On the other hand, fall dormancy, defined as the reduction of growth in fall in response to short day length (Blondon et al., 1967), is a component of fall and spring yield. It is also a main, but not unique, component of winter frost tolerance (Teuber et al., 1998; Brummer et al., 2000; Willame et al., 2002; UPOV, 2005). Each breeding program is usually conducted within a restricted fall dormancy range to release varieties targeting a specific climate.

In this study, we gathered 'cultivated material', comprising old and recent cultivated accessions mainly from Europe, North and South America, and China, and further extended the genetic diversity with breeding material of five major European breeders. From this material, we assessed whether a genetic structure has been created by preferential crosses between materials of specific fall dormancy groups. We studied how diversified breeding material of the five European breeders is, compared to the diversity found in the cultivated material. With the cultivated material, we assessed if a GWAS approach can detect QTL for phenology traits and if the genomic prediction models allow to predict phenology with a good accuracy. In this study, GWAS and GP analyses were conducted at accession (*i.e.* population) level, with SNP frequencies as genotyping data.

2 Material and method

2.1 Plant material

We used 400 cultivated accessions (hereafter named as 'cultivated material') comprised of 378 cultivars and 22 landraces whose fall dormancy score mainly ranged between 3 and 7. Their origin, based on the place they have been collected (landraces) or initially selected and registered (cultivars) was Europe (318 accessions), North America (45 accessions), South America (16 accessions), China (17 accessions), Middle East (3 accessions) and Japan (1 accession). In addition, 617 accessions (hereafter named as 'breeding material') representing advanced breeding material obtained by five European breeders were included: 144 accessions from breeder A, 62 accessions from breeder B, 101 accessions from breeder C, 189 accessions from breeder D and 121 accessions from breeder E. Each breeder chose their material for this study but did not mention if it was used in or representative of the breeding program of the company or institution. All the 1017 accessions were genotyped and used for genetic structure study while only the 400 'cultivated material' accessions were phenotyped for phenology and used for GWAS and GP studies. The origin of the material used is available in the data repository (see the Data Availability Statement section).

2.2 Genotyping

The methodology used for the DNA extraction, the optimization of the GBS methodology and GBS sequencing has previously been reported in Julier et al., 2021. To summarize, each accession was represented by 100 plants, the DNA extraction was performed from a pool of 100 leaflets, each taken on a plant. This protocol has previously been shown to be reliable to estimate the allele frequency of an accession (Julier et al., 2018). The double-digest GBS on alfalfa was conducted with the enzymes PstI-MseI to obtain a sufficient number of loci, while reducing the number of missing values and considering the number of reads per accession.

2.3 Trimming and SNP calling

The reads were preprocessed with the GBprocesS bioinformatics pipeline (Schaumont, 2020). This pipeline includes several steps: demultiplexing, trimming of barcodes and restriction enzyme cutsite remnants, merging of forward and reverse reads, removal of reads with low quality base-calling and internal restriction sites. Subsequently, the reads were mapped onto the reference sequence (Chen et al., 2020) by using the BWA software with the BWA-MEM algorithm and default options. We performed a test on a batch of samples to compare the number of SNPs when the reads were mapped on each of the four homologous chromosomes of the reference genome. The haploid copy of the genome giving the highest number of SNPs was chosen (number 2) as the reference to map the reads for the rest of the accessions. We used SMAP *delineate* (Schaumont et al., 2022) to analyse stacks of GBS reads mapped onto the reference sequence, and found 31 743 loci. A custom pipeline was used to perform the genotype calling. First, for each accession, the number of reads per position and per nucleotide (A, T, G, C) was extracted with the software bamreadcount (Khanna et al., 2022). Per accession, a threshold was applied to keep only the positions with at least 10 reads and at most 1200 reads. A list containing all the positions found across accessions was established which included 22 192 769 positions. For each position, the allele frequency of each nucleotide was calculated as the number of reads for the targeted nucleotide divided by the total number of reads at this position. Two stages of position selection were then carried out to retain the positions with a minor allele frequency greater than or equal to 1% and two alleles, leading to 1 194 485 positions. After this step, each accession was genotyped for the remaining positions by calculating the allele frequency of the alternative allele. In a third selection, we retained 631 816 positions with a minor allele frequency per accession between 5% and 50% in at least 10 accessions. Five accessions with more than 80% of missing data were excluded from the analysis (Figure 1A).

To allow robust analyses, it is necessary to filter out markers with too many missing data. When the percentage of missing data per SNP was plotted against the number of markers, we obtained a sigmoid curve (Figure 1B) on which we have represented different thresholds and the number of markers retained. By applying the thresholds of 0%, 1%, 5%, 20% and 50% missing data, 89 216, 186 190, 227 092, 274 471 and 325 453 SNPs were retained, respectively. In this study, we applied the filter of 5% maximum missing data per position (227 092 SNPs). We calculated the number of markers over a distance of 500 kb. This density of markers is variable along and between the chromosomes (Figure 1C) but the whole genome is covered, except for two zones which are certainly centromeres.

2.4 Population structure

A subset of 89 216 SNPs without missing values was used for a genetic structure analysis. Linkage disequilibrium decay was

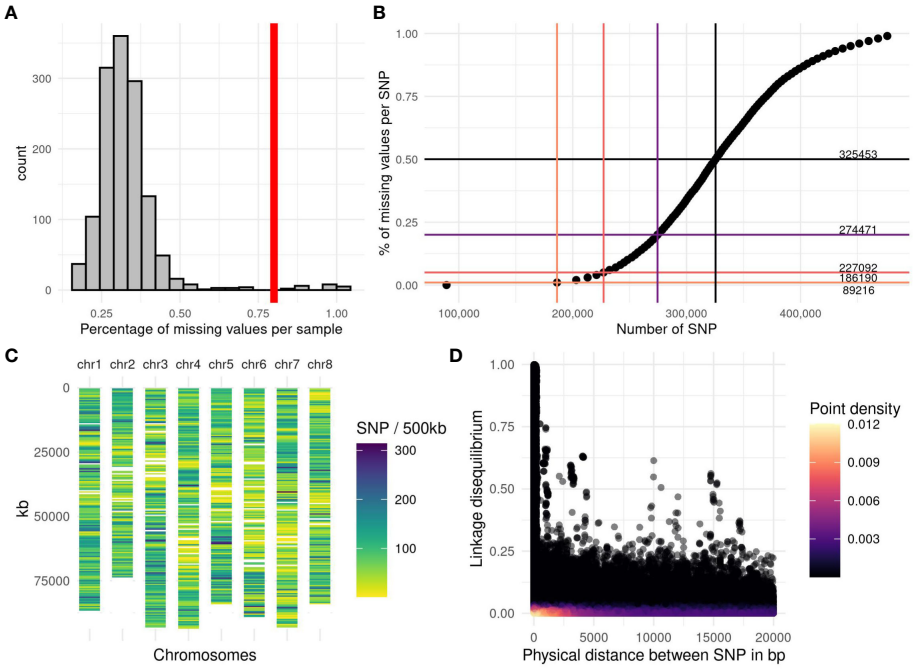


FIGURE 1
Genotyping quality and linkage disequilibrium observed among the accessions. **(A)** Histogram of the percentage of missing values per accession, the vertical red line represents the applied threshold of 80% of missing values per accession. **(B)** Number of SNPs available depending on the percentage of missing values allowed per SNP. The horizontal lines are the thresholds (1%, 5%, 20% and 50%), the vertical bars represent the number of SNPs obtained with the corresponding threshold. The number in the right part of the graph indicates the number of SNPs from 0% of missing value to 50% of missing values. **(C)** SNP density per chromosome along the genome, estimated by the number of markers in a window of 500 kb on the 227 092 SNPs obtained with a threshold of 5% missing value per SNP. **(D)** Linkage disequilibrium (LD) between the 227 092 SNPs, estimated with a squared partial correlation. The LD was plotted for SNP distances of less than 20 000 bp. The purple color scale represents the point density, black for a low density and yellow for the highest density.

calculated based on the squared partial correlation between pairs of SNPs (Lin et al., 2012; Mangin et al., 2012). Genetic groups among the populations were identified with the Discriminant Analysis of Principal Components (DAPC) method (Pritchard et al., 2000; Jombart et al., 2010; Grünwald and Goss, 2011) implemented in the R package adegenet (Jombart, 2008; Jombart and Ahmed, 2011). The genetic groups were identified by using k-means, a clustering algorithm that found a given number (k) of groups by maximizing the variation between groups. The optimal number of groups was the one that provided the lowest Bayesian Information Criterion (BIC). We then ran a Principal Component Analysis (PCA) with the R package 'FactoMineR' (Lê et al., 2008) to analyze the diversity among accessions without prior hypothesis and their inclusion to each group was illustrated. The same R package ('FactoMineR') was used to project the breeders' accessions (as supplementary individuals) into the PCA of the cultivated material. To illustrate the relationship with the genetic structure found with the SNPs and the fall dormancy, a PCA was performed on phenotypes (see next section) and the accessions were colored by group as found by the clustering analysis. We also used the results of the PCA based on the genomic information and colored the accessions depending on the phenotype for each trait. The differentiation between the groups was assessed through the F_{ST} value following the methods of the R package StAMPP (Pembleton et al., 2013) for polyploid species. The group with only two accessions was ignored for the F_{ST} estimation.

2.5 Phenotyping

All the "cultivated material" plant material was established in two locations for phenotyping, the first one in France at the research unit (URP3F) of INRAE (Lusignan: 46° 23' 60" N, 0° 4' 48" E) and the second one in Serbia at the research unit of IFVCNS (Novi Sad: 45° 15' 0" N, 19° 51' 0" E). In Lusignan, the trial was sown on the 10th of May 2018. The trial was damaged by a storm on the 26th of May 2018 and a new trial was sown on the 23rd of August 2018, but some of the cultivated material did not have enough seeds, so only 387 among the original 400 accessions were established. In Novi Sad, the trial was sown on the 21st of May 2018. The trials were composed of 440 plots, 44 columns and 10 rows, in an augmented block design with four incomplete blocks (Federer and Raghavarao, 1975; Lin and Poushinsky, 1985). The Table S1 summarises the dimension and the technical elements of the trials. Five accessions were repeated six times and distributed in the four blocks, 15 other accessions were repeated twice in the trials of Lusignan-May 2018 and Novi Sad, but in Lusignan-August 2018, 28 accessions were repeated twice. The other accessions were present only once and randomly distributed within and between the blocks. Measurements and scorings were performed in each trial in 2018, 2019 and 2020. The trial installed in Lusignan was evaluated during an extra year in 2021. On the trial established in May 2018 in Lusignan, the number of surviving plants in each plot was enough

to record the flowering date in summer 2018. The date of flowering (FD.L) was then converted into a degree.day sum, by adding up the degrees Celsius above zero between the date of sowing and the date of flowering, using the mean daily temperatures at the location and in the year of the trial. The assessment of fall dormancy was carried out by measuring several traits on the trials established in August 2018 in Lusignan (.L) and May 2018 in Novi Sad (.N): plant height (PH19.L, PH20.L, PH21.L, PH19.N, PH20.N) before the last cut in fall 2019 and 2020 and dry matter yield (F-DMY19.L, F-DMY21.L, F-DMY19.N, F-DMY20.N) at the last cut in fall 2019 and 2020. A fall dry matter yield combined over all the years and locations (F-DMY) was estimated with a mixed model to remove the year and the location effects. Fall plant height measurements were made with an electronic ruler when heights were less than 35 cm and with a conventional ruler when heights were greater than 35 cm. Three heights per plot were measured, randomly in the high plant density plots and on the most developed plants in the degraded plots. All measurement and cutting dates are available in Table 1. In Lusignan, plant height was measured several times between the last two cuttings in the fall 2019 and 2021, and the stem elongation speed (SE19.L, SE21.L, in cm/degree.day) was obtained by using the slope of the regression between the height and the date of measurement expressed in degrees.days above 0°C. Finally, fall dormancy was visually scored on the 29th of October 2019, on a 1-11 scale based on regrowth height (D19.L). Due to a very dry fall in Lusignan in 2020, fall regrowth was not sufficient for a cutting and an estimation of dry matter yield.

TABLE 1 List of measurement dates of each trait and the prior cutting date in DD.MM.YYYY format.

Trait	Measurement date	Date of prior cut
D19.L	29.10.2019	17.09.2019
F-DMY19.L	19.11.2019	17.09.2019
F-DMY21.L	01.09.2021	28.07.2021
F-DMY19.N	16.10.2019	22.08.2019
F-DMY20.N	22.10.2020	11.08.2020
PH19.L	29.10.2019	17.09.2019
PH20.L	16.10.2020	28.07.2020
PH21.L	09.11.2021	28.07.2021
PH19.N	16.10.2019	22.08.2019
PH20.N	22.10.2020	11.08.2020
SE19.L	07.10.2019-14.10.2019-21.10.2019-29.10.2019-13.11.2019	17.09.2019
SE21.L	13.09.2021-22.09.2021-18.10.2021-09.11.2021	28.07.2021

Flowering date (FD), Dormancy (D), Fall Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE) for two years: 2019 (X19.X) and 2020 (X20.X) in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia.

2.6 Phenotypic adjustment and genetic parameter estimation

All traits were independently adjusted to field micro-environmental heterogeneity with the breedR package (Muñoz and Sanchez, 2020). Within trials, to capture the spatial heterogeneity at the plot level, a random effect was fitted thanks to the use of the tensor product of two B-splines bases with a covariance structure for the random knot effects (RKE) to account for spatial variability along the rows and the columns of the field design (Cantet et al., 2005; Cappa and Cantet, 2007; Robbins et al., 2012; Cappa et al., 2015). We used a genomic based mixed model for each year and each location. The genomic estimated breeding values (GEBV) for each trait were estimated with the best linear unbiased prediction based model (GBLUP) (Whittaker, 2000; Meuwissen et al., 2001):

$$y = \mu + Zu + Ws + \epsilon \quad (1)$$

where y was the raw phenotypes, μ the global mean, u the vector of random additive effects following $\mathbb{N}(0, G\sigma_a^2)$ with σ_a^2 the additive variance and G the genomic relationship matrix between accessions, s was the vector of random spatial effects containing the parameters of the B-splines tensor product following $\mathbb{N}(0, S\sigma_s^2)$ with σ_s^2 the variance of the RKE for rows and columns and S the covariance structure in two dimensions, ϵ the vector of residual effects following $\mathbb{N}(0, I\sigma_\epsilon^2)$with σ_ϵ^2 the residual variance. The design matrix Z and W are identity matrices connecting the plots to the random effects. The method used to obtain the genomic relationship matrix G is explained in the next section. B-splines were anchored at a given number of knots for rows and columns, a high number of knots smooths out the surfaces. breedR optimized the knot numbers by an automated grid search based on the Akaike information criterion (Akaike, 1974). The micro-environmental plot effect was subtracted from the observed phenotype to obtain a spatially adjusted phenotype. For the repeated accessions, we calculated an accession mean of the spatially adjusted phenotypes for each trait.

This model was used to estimate the narrow sense heritability of the trait. To avoid inflated heritability (Heckerman et al., 2016), the variance explained by the spatial effect is integrated in the heritability formula:

$$h^2 = \frac{VarG}{VarG + VarE + VarR} \quad (2)$$

With *VarG* the additive variance, *VarE* the micro-environmental plot variance and *VarR* the residual. We used a multi-trait model on adjusted phenotypes coupled with information from relatedness between individuals based on genomic information (Calus and Veerkamp, 2011) to extract the genetic correlation between traits and compare the genetic correlation with the phenotypic correlation calculated from the Pearson's correlation on adjusted phenotypes.

2.7 Relationship matrix estimation

The genomic relationship matrix (G) was based on (VanRaden, 2008), adapted to use allele frequencies (continuous values from 0

to 1) instead of allele dosage (Ashraf et al., 2014). The genotyping matrix (M) was normalized by the minimum allele frequency (P) to obtain the normalized genotyping matrix (Z) used to compute G, as follows:

$$G = \frac{ZZ'}{\frac{1}{n}\sum_{j=1}^{m} p_j(1-p_j)} \quad (3)$$

The denominator is a scaling parameter, corresponding to the sum of the expected SNP variance across genotypes (Ashraf et al., 2014), where m represents the number of markers, p_j equals the frequency of the jth marker, and n represents a scaling number to obtain a diagonal mean equal to 1. This has been recommended in previous studies on polyploid species (Ashraf et al., 2014; Cericola et al., 2018), with = 16 , the diagonal mean was close to 1.

2.8 GWAS

The GWAS analyses were performed with the MLMM method (Segura et al., 2012), while taking into account the genetic structure of the "cultivated material" with the genomic relationship matrix. The MLMM method uses a stepwise mixed-model regression approach with forward inclusion of the SNP as co-factors and a backward elimination. The variance components of the model are re-estimated at each step. This method is known to increase the detection power while decreasing the false detection rate. The maximum number of steps was limited to ten. The best step is selected with an adjusted (0.05/number of GBS loci) multiple Bonferroni criterion (mBonf). The percentage of phenotypic variation explained by each QTL was obtained by subtracting the R^2 of a linear model with all the QTL as fixed effects and the genomic relationship matrix (G) as random effect to the R^2 of the same model but without the focused QTL.

Genes located within 2500 bp flanking each QTL were determined using the Genome Browser (https://bbric-pipelines.toulouse.inra.fr/myGenomeBrowser?portalname=MSAT_XinJiangDaYe&owner=sebastien.carrere@inrae.fr&key=PyG9k9tK). Three sources of gene annotation were available, one from the reference genome used for the genotype calling (Chen et al., 2020), one from a partly assembled European genome (Carrère et al., 2020) and one from the genome of the model legume species *Medicago truncatula* (Pecrix et al., 2018).

2.9 Genomic prediction

2.9.1 Test of the size and the genetic composition of the training population

To assess the potential of genomic prediction, we used the predicting ability calculated by the correlation between the phenotype and the value predicted in the validation population. The validation population represents a portion of the complete dataset on which the phenotypes have been masked and only the genotype information is available. Here, two cases were considered. In a first case, 100 accessions were randomly taken to form the validation population, in a way that each group (see paragraph Genetic structure) is represented according to its size. We randomly sampled the remaining accessions to test the effect of the size of the training population. Nine sample sizes were tested: ranging from 10% of the remaining accessions (29 accessions) to 90% (270 accessions), ten iterations were performed for each sample size. In order to assess the effect of the training population composition on the quality of the predicting ability, a second case was studied. The validation population was composed of fifteen accessions from one group only and the training population of 210 accessions was randomly taken from the remaining groups. Ten repetitions were performed. Groups with less than fifteen accessions were excluded. For each repetition, the validation population was sampled from within a group and was predicted using two different training populations. To test the predicting ability across groups, the first training population excluded the other accessions belonging to the same group as the validation population. The second training population was a random sample of all groups. We ensured that the potential confusion between the effect of the composition of the training population and the effect of the sample size was avoided by using the same sampling but two different training populations.

2.9.2 Test of statistical models

First, the best linear unbiased prediction based on genomic information (GBLUP) (Whittaker, 2000; Meuwissen et al., 2001) was used to predict the genomic estimated breeding values with all the SNP used to compute the genomic relationship matrix matrix (G). The R package breedR was used.

Then, we used two Bayesian methods, the Bayesian Ridge Regression (BRR) (Pérez et al., 2010) and the Bayesian Lasso (lasso) (Tibshirani, 1996; Tibshirani et al., 2012), with the R package glmnet (Friedman et al., 2010; Simon et al., 2011). The BRR and the GBLUP method mimic an infinitesimal genetic architecture. The Bayesian Lasso method selects features (here SNPs) depending on their importance and uses them as a predictor.

We compared the different models in the situation of the first case scenario with a validation population of 100 accessions randomly selected and a training population of 270 randomly selected accessions (90%).

3 Results

3.1 Genotyping

The GBS pipeline optimization, as performed in this study, led to a genotyping dataset with little missing values among the accessions. Only five accessions out of 1017 were lost due to a poor sequencing depth (Figure 1A). We chose to keep SNPs with a maximum of 5% missing values per SNP (Figure 1B), the missing values were imputed with the minor allele frequency, and it represented less than 1% of all the genomic data. We chose this low threshold to avoid adding bias in the QTL detection and the genomic prediction. GBS-tagged loci were spread fairly evenly throughout the genome, with at least 1 polymorphic SNP marker

10.3389/fpls.2023.1196134

every 500 kb (Figure 1C). Linkage disequilibrium (LD), estimated with a squared partial correlation, dropped abruptly after 1000 bp, most SNP pairs presented a LD close to zero as shown by the yellow colour representing a high density (Figure 1D). Indeed, after a distance of 100 kb, on the 810 557 188 total pairs of SNP, only 872 SNP pairs showed a partial correlation >0.25; and 75 SNP pairs had a partial correlation >0.5. The longest distance between two SNPs were 89 Mb and 91 Mb, with a partial correlation of 0.5 and 0.25, respectively. Further examination revealed that this long-range LD is mainly due to a few pairs of SNP that have high LD with several flanking SNPs located in another region of the chromosome, suggesting a local problem in assembly or a wrong mapping.

3.2 Population structure

The number of principal components (PC) required to explain 90% of the genetic variation was 300 (Figure S1A). The optimal number of groups was the one that provided the lowest Bayesian Information Criterion (BIC), here seven (Figure S1B). On the PCA with the accessions coloured according to the seven DAPC groups (Figure 2), two groups were clearly separated from the others: group 6 with 15 Chinese accessions and group 1 with two accessions (an Italian variety that includes a *falcata* parent and a Hungarian variety). The five other groups showed a genetic continuum but each group can be related to the geographic origin of the accessions: group 3 with 139 accessions of European origin (France and Northern Europe), group 7 with 151 accessions mostly of European origin (Southern and Eastern Europe), group 4 with 61 accessions of Europe, North and South America, group 5 with 21 USA and 1 Chinese accessions and group 2 with 5 North American accessions. The European accessions as well as the American accessions were thus split into several groups. The group 4 probably illustrates the multiple origins of some varieties selected in the two continents.



FIGURE 2
Clustering of the cultivated material based on a Principal Component Analysis performed on a genotyping dataset of 89 216 SNPs without missing data. The analysis was done on 395 accessions, and a Discriminant Analysis of Principal Components (DAPC) analysis revealed seven groups (ellipses) linked to the geographic region of the origin or registration.

Finally, the European and the North American accessions displayed little overlap, but South-American accessions interestingly overlapped with these two groups. The three accessions from the Middle East were close to the American accessions and the Japanese accession was closer to the European-American groups than to the Chinese group. All Chinese varieties resided in the Chinese group except one variety that grouped into group 5, probably revealing a selection based on American material. Accessions from China seemed to be different from *falcata* material (group 1), even if the latter was represented by only two accessions in this study. The distance between the group 6 with Chinese accessions and the overlapping groups 3, 4, 5, 7 containing Western accessions suggested unconnected breeding programs. The F_{ST} value between the groups were low with an average value of 0.01 and a range between 0.001 and 0.026 (Table 2). Groups 6 and 2 were the most distinct, with the highest F_{ST} value (0.026).

The breeding material provided by five European breeders was compared to the groups obtained with the worldwide cultivated material (Figures 3, S2 for a detailed view by breeder). Three breeders provided materials with a narrow genetic diversity that were assigned to a single group: accessions of breeders B and E were assigned to group 3 (France and Northern Europe), those of breeder C were assigned to group 7 (Southern and Eastern Europe). The other two breeders provided more diversified genetic materials. The accessions of breeder A covered at least three groups: 3, 7, and 4 (Europe, North and South America). The accessions of breeder D covered groups 3 and 7, and more surprisingly, it seems that some of this material was crossed with *falcata* or possibly Chinese accessions. Nonetheless, the genetic material provided by the five European breeders in this study did not belong to the North American nor the Chinese groups.

A PCA based on phenotypic data did not reveal any structure and the genetic groups overlapped with each other (Figure 4; for a PCA based on molecular data and coloured per trait, see Figure S3). However, when the trait variation was displayed as boxplots per group (Figure 5), the group with the Chinese accessions showed the lowest values of the traits linked to fall dormancy (highly dormant) and the highest values for the flowering date. The other groups were

TABLE 2 Genetic distinction (F_{ST}) between the groups found by the DAPC method.

Groups	1	2	3	4	5	6
1						
2						
3		0.017				
4		0.013	0.003			
5		0.004	0.006	0.002		
6		0.026	0.019	0.012	0.015	
7		0.016	0.001	0.001	0.005	0.015

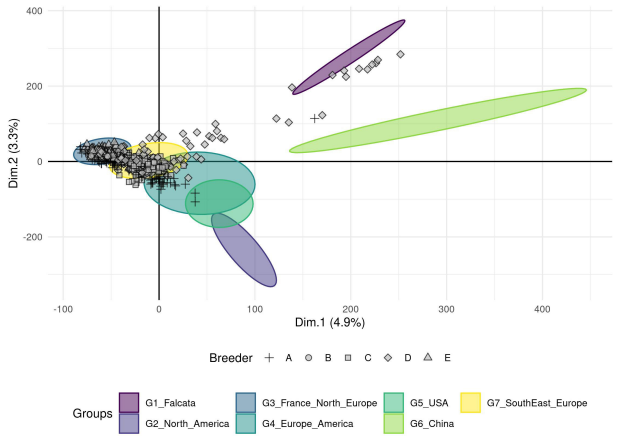


FIGURE 3
Projection of the 617 European breeders' accessions on the seven groups (ellipses) obtained from a DAPC analysis using a genotyping dataset of 89 216 SNPs without missing data. A particular point shape represents each breeder.

relatively similar, except group 1, which contained only two *falcata*-type accessions.

3.3 Heritability and genetic correlation

Model estimated variances, heritability and accession mean estimated after phenotypic adjustment, are presented in Table 3. Our study showed a wide variation in heritability, ranging from 0.01 for F-DMY20.N to 0.79 for PH20.L that evaluates fall dormancy. F-DMY20.N, PH20.N, PH21.L, and F-DMY21.L showed the lowest heritability (<0.04). These low heritabilities can be explained by a high micro-environmental variance (VarE) and not by the absence of genetic variability (VarG) for the trait. On average, the heritabilities were higher in Lusignan (0.34) than in Novi Sad (0.13). Similarly, the average heritabilities of the traits measured in 2019 (0.48) were higher than those of the traits measured in 2020 (0.09) and 2021 (0.06). Finally, the measurements of SE and PH had higher average heritabilities (0.27) than the measurements of F-DMY (0.18).



FIGURE 4
Principal Component Analysis (PCA) based on the phenotypic values, the accessions are colored per DAPC group. The traits are related to Flowering date (FD) and autumn dormancy depending on different measurements in autumn: Dormancy (D), Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE) for two years: 2019 (X19.X) and 2020 (X20.X) in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia. F-DMY without letter or number is the Dry Matter Yield adjusted for year and location effects. The biggest points represent the centroids of each cluster.

Table 4 shows the phenotypic and genetic correlation between traits. All the traits related to fall dormancy presented positive phenotypic and genetic correlations between each other. As expected, FD.L showed negative phenotypic and genetic correlation with all the traits related to fall dormancy. Among the traits related to fall dormancy, the traits measured in Lusignan presented stronger average genetic correlation (0.83) than at Novi Sad (0.78) and stronger than between locations (0.48). Interestingly, PH19.L and PH19.N that were the same trait measured in the same year but in two different locations, presented a genetic correlation of 0.997. These results suggest a low genotype by environment interaction for this trait linked to fall dormancy. Among the fall dormancy traits, F-DMY19.N showed the lowest but still positive genetic correlation (lower than 0.6) with all the other traits.

We found QTL for four out of the eleven traits of this study (Figure S4). For D19.L, a single QTL was found on chromosome 8 (chr 8) explaining 14.6% (Table 5) of the phenotypic variation. For F-DMY20.N, five QTL were detected: one QTL each on chr 2, 5, 6 and two on chr 7, explaining between 6% and 9% each; overall, they explained 32.1% of the phenotypic variation. For PH19.N, we found six QTL, one on chr 2 and chr 4, two on chr 3 and chr 7, each explaining between 7% and 11.9% of the variation; overall they explained 42.6%. Finally, for F-DMY, we found five QTL on chr 2, and four on chr 3, each explaining between 9.5% and 15.3% of the variation; overall they explained 43.2% of the phenotypic variation. QTL located on the same chromosome were spaced at least 9 Mb apart. No QTL was detected for flowering date. To understand the lack of common QTL between genetically correlated traits, we looked in detail at all the SNPs that were detected as potential QTL by the MLMM method. This iterative method added the potential QTL one by one as a co-factor in the model before estimating which model is the best and thus which are the "true" QTL. The additional data table (Table S2) tracked all SNPs selected by each iteration and for each trait. We observed a few cases where QTL for the different traits were located in close proximity. In four cases, the significantly associated SNPs are less than 1000 bp apart. Among these cases, for a pair of SNPs (chr2_12854184 - chr2_12854196), neither of them passed the threshold but were selected by the mlmm method at certain steps, these two SNPs would have an effect on the following traits: PH19.L and F-DMY20.N. In the remaining three cases, one of the two SNPs of the pair was retained as a QTL: chr3_61230828 - chr3_61230888 for PH20.N and F-DMY, respectively; chr3_89061410 - chr3_89061526 for F-DMY and PH19.L, respectively; chr8_51582915 - chr8_51582964 for PH20.N and D19.L. Three SNPs were selected for two different traits: chr2_14385543 for D19.L and F-DMY19.L; chr7_46045300 for PH21.L and SE21.L; chr7_90860527 for PH21.L

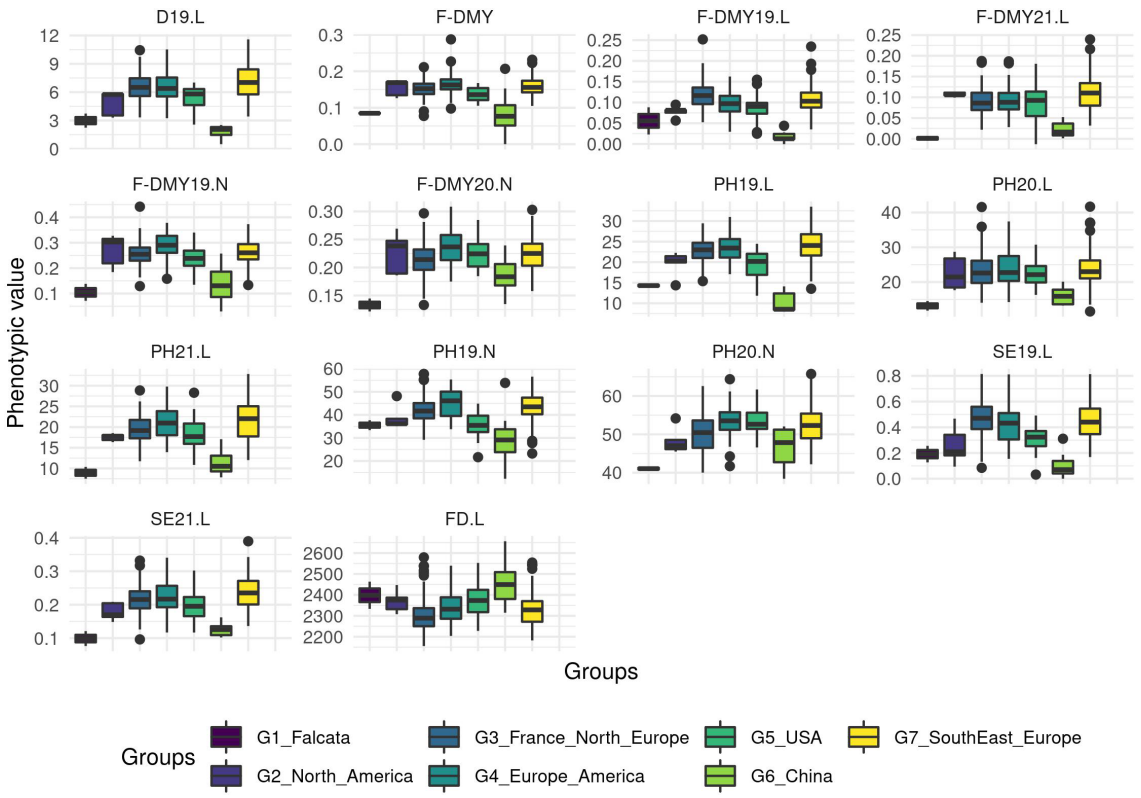


FIGURE 5
Boxplot of the phenotypic values per group for all the traits related to Flowering date (FD) and autumn dormancy scored from different measurements in autumn. Dormancy (D), Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE) for two years: 2019 (X19.X) and 2020 (X20.X) in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia. F-DMY without letter or number is the Dry Matter Yield measured in autumn adjusted for year and location effects.

TABLE 3 Results of models fit by trait with mean by phenotype, genetic (VarG), spatial by location and year (VarEX.X) and residual (VarR) variances used for heritability estimation (h^2).

Traits	mean	VarG	VarE18.L	VarE19.L	VarE20.L	VarE21.L	VarE19.N	VarE20.N	VarR	h^2
D19.L	6.62	2.129		0.281					0.5929	0.71
FD.L	2327.82	3285	4481						3809	0.28
F-DMY	0.15	0.00035		0.0000214	0.0000089	0.0036	0.00213	0.00056	0.00181	0.15
F-DMY19.L	0.10	0.0006		0.000105					0.00046	0.52
F-DMY19.N	0.26	0.00186		0.00495					0.00153	0.22
F-DMY20.N	0.22	0.00027			0.02653				0.00167	0.01
F-DMY21.L	0.10	0.00064				0.06298			0.0009	0.01
PH19.L	22.98	9.454		0.5518					1.928	0.79
PH19.N	42.16	22.95		41.2					20.59	0.27
PH20.L	23.27	14.78			32.8				14.37	0.24
PH20.N	51.66	4.441			87.93				52.77	0.03
PH21.L	20.31	11.38				1223			5.707	0.01
SE19.L	0.43	0.00816		0.00159					0.0133	0.35
SE21.L	0.22	0.00136				0.00539			0.00105	0.17

The traits are: flowering date (FD) and fall dormancy (Dormancy: D, plant height: PH, stem elongation rate: SE, Fall Dry Matter Yield: F-DMY) for different locations (Lusignan: L, and Novi Sad: N) over three years of trials (2019: X19.X, 2020: X20.X and 2021: X21.X) or overall (F-DMY).

TABLE 4 Correlation between traits measuring flowering date (FD) and fall dormancy (Dormancy: D, plant height: PH, stem elongation rate: SE, Fall Dry Matter Yield: F-DMY) for different locations (Lusignan: L, and Novi Sad: N) over three years of trials (2019: X19.X, 2020: X20.X and 2021: X21.X) or overall (F-DMY).

	D19.L	F-DMY	F-DMY19.L	F-DMY21.L	PH19.L	PH20.L	PH21.L	SE19.L	SE21.L	F-DMY19.N	F-DMY20.N	PH19.N	PH20.N	FD.L
D19.L		0.738	0.743	0.883	0.930	0.998	0.856	0.848	0.966	0.352	0.599	0.768	0.658	-0.313
F-DMY	0.639		0.768	0.710	0.843	0.806	0.746	0.804	0.776	0.886	0.816	0.919	0.723	-0.239
F-DMY19.L	0.631	0.603		0.578	0.726	0.605	0.553	0.926	0.649	0.554	0.484	0.659	0.319	-0.396
F-DMY21.L	0.553	0.561	0.409		0.828	0.999	0.934	0.700	0.964	0.365	0.553	0.647	0.735	-0.054
PH19.L	0.865	0.666	0.639	0.602		0.995	0.850	0.882	0.897	0.367	0.333	0.997	0.677	-0.292
PH20.L	0.400	0.358	0.309	0.535	0.469		0.746	0.590	0.998	0.107	0.119	0.990	0.844	0.030
PH21.L	0.674	0.600	0.441	0.796	0.707	0.538		0.735	0.993	0.291	0.314	0.703	0.852	-0.104
SE19.L	0.568	0.388	0.455	0.338	0.602	0.178	0.417		0.800	0.215	0.163	0.294	0.150	-0.492
SE21.L	0.698	0.584	0.462	0.748	0.707	0.542	0.926	0.456		0.257	0.297	0.383	0.248	-0.130
F-DMY19.N	0.364	0.707	0.266	0.188	0.429	0.488	0.398	0.538	0.398		0.796	0.706	0.681	-0.403
F-DMY20.N	0.351	0.495	0.202	0.234	0.642	0.730	0.683	0.534	0.654	0.352		0.726	0.995	-0.048
PH19.N	0.484	0.546	0.324	0.304	0.518	0.223	0.417	0.833	0.714	0.528	0.321		0.788	-0.373
PH20.N	0.308	0.347	0.094	0.195	0.276	0.138	0.289	0.505	0.762	0.251	0.636	0.284		-0.038
FD.L	-0.235	-0.248	-0.282	-0.123	-0.227	-0.083	-0.161	-0.221	-0.156	-0.219	-0.066	-0.172	-0.046	

The lower part in grey represents the phenotypic correlation between traits after phenotypic adjustment and the upper part represents the genetic correlation between traits based on the covariance matrix estimated with a multi-trait model.

and SE21.L, but not detected as QTL. Table S4 shows the distance between the QTL (highlighted) and other non-conserved SNPs that are less than 1000 bp apart. Between the detected QTL, the distance was large and the linkage disequilibrium values were very low, 0.006 on average. However, in cases where the SNPs are in close vicinity, the linkage disequilibrium locally increased strongly. These regions certainly contained QTL related to fall dormancy, but we cannot consider them as such due to our conservative threshold, which allowed us to limit both the number of false positives and a too small number of accessions leading to a low power of detection.

For 14 out of the 17 QTL, genes were found in the 5 kb flanking genomic region. Details are given in Table 5 and the corresponding annotations are listed in Table S3. Some genes encode similar functions, or may be involved in a common biological process. These functional annotations include: drought stress response (2-methylene-furan-3-one reductase (Singh et al., 2022), transcription factor C3H family (Kumar et al., 2019)), growth and development (ABC-type xenobiotic transporter (Verrier et al., 2008), F-box domain (Gupta et al., 2015), malate dehydrogenase (oxaloacetate-decarboxylating) (NADP(+)) (Kujur et al., 2016)) and diverse biological processes (ACP-like superfamily (Zhao et al., 2022), transcription factor WD40-like family (Xu and Min, 2011), and leucine-rich repeat domain superfamily (Liu et al., 2022)). The functions of certain genes linked to our QTL (peptidase C78, ubiquitin modifier-specific peptidase 1/2, thimet oligopeptidase, Type I protein exporter) were not well known in plants in general nor legumes in particular.

3.4 Genomic prediction

The predicting ability obtained in this study for flowering date and the traits linked to fall dormancy varied from 0.25 to 0.80, depending on the trait and the method (Figure 6). The GBLUP and the BRR method gave equivalent predicting ability (Figure 6), the Bayesian Lasso method gave slightly lower prediction quality and a higher standard deviation. The main differences in predicting ability depended on the trait. Fall dormancy in Lusignan showed a higher predicting ability (from 0.50 to 0.80) than in Novi Sad (from 0.25 to 0.68). The results showed less variation between years in Lusignan (2019 and 2021) than in Novi Sad (2019 and 2020). The predicting ability for all traits obtained with GBLUP were above 0.25. The lowest predicting ability on average (above 0.25) were obtained for PH20.L, PH20.N, F-DMY20.N and FD.L. The best predicting abilities (average >0.75) were obtained with D19.L, PH19.L and PH21.L. Other traits had a predicting ability higher than 0.5. Only the results with GBLUP were considered for the rest of the study.

The mean predicting ability (GBLUP method) increased markedly by increasing the size of the training population from 29 to 59 accessions, and reached a plateau with a training population of at least 89 accessions (Figure 7). The variance between replicates still decreased in training populations with more than 89 accessions, as shown for five example traits: D19.L, F-DMY19.L, FD.L, PH19.L and SE19.L (Figure 7). This observation was consistent across traits and locations (Figure S5). Unlike the training population size, which appeared to have a similar effect on

TABLE 5 QTL for autumn dormancy from a GWAS analysis with the MLMM method.

Trait	SNP	r2	r2 Global	Chromosome	Position	Annotation of Chen et al., 2020		Mercedes				Medicago tuncatula			
						Gene	Function	Gene 1	Function	Gene 2	Function	Gene 1	Function	Gene 2	Function
D19.L	chr8_51582964	0.146		chr8.2	51582964	MS.gene035451	ACP-like superfamily; Protein kinase-like domain superfamily	MsNRG001015g01756361	hypothetical protein	MsNRG000350g01027551	hypothetical protein	MtrunA17Chr7g0240251	hypothetical protein		
F-DMY20.N	chr7_53172294	0.090	0.321	chr7.2	53172294	MS.gene010897	Heat shock protein 70kD, peptide-binding domain superfamily	MsNRG000432g01160181	hypothetical protein						
F-DMY20.N	chr7_62988589	0.059		chr7.2	62988589	MS.gene67894	Alcohol dehydrogenase, N-terminal	MsNRG000037g00224451	Putative 2-methylene-furan-3-one reductase			MtrunA17Chr7g0228011	Putative 2-methylene-furan-3-one reductase		
F-DMY20.N	chr6_40691177	0.086		chr6.2	40691177										
F-DMY20.N	chr5_72247110	0.060		chr5.2	72247110	MS.gene98525	Lipid-binding serum glycoprotein, N-terminal ; Lipid-binding serum glycoprotein, C-terminal								
F-DMY20.N	chr2_12854196	0.086		chr2.2	12854196	MS.gene073096	Nucleotide-binding alpha-beta plait domain superfamily ; U2 auxiliary factor small subunit	MsNRG001447g01984221	Putative transcription factor C3H family	MsNRG001933g02106001	Putative transcription factor C3H family	MtrunA17Chr2g0320811	hypothetical protein		
PH19.N	chr3_90724627	0.119	0.426	chr3.2	90724627	MS.gene014383	Tetratricopeptide-like helical domain superfamily	MsNRG000840g01620971	Putative tetratricopeptide-like helical domain superfamily	MtrunA17Chr3g0143411	Putative tetratricopeptide-like helical domain superfamily	MtrunA17Chr3g0143411	Putative tetratricopeptide-like helical domain superfamily		
PH19.N	chr7_90834458	0.106		chr7.2	90834458										
PH19.N	chr7_10758065	0.105		chr7.2	10758065	MS.gene020657		MsNRG143424g02636541	Putative diacylglycerol kinase (ATP)	MsNRG000355g01035221	hypothetical protein				
PH19.N	chr4_5061600	0.080		chr4.2	5064100										
PH19.N	chr3_73810222	0.074		chr3.2	78458544	MS.gene06565	ABC transporter type 1, transmembrane domain	MsNRG000105g00468091	Putative Type I protein exporter	MsNRG000921g01687971	Putative ABC-type xenobiotic transporter	MtrunA17Chr3g0128361	Putative peptidase C78, ubiquitin modifier-specific peptidase 1/ 2	MtrunA17Chr3g0128381	Putative Type 1 protein exporter
PH19.N	chr2_13623107	0.103		chr2.2	13623107	MS.gene69859		MsNRG066413g02467211	Putative protein-serine/threonine kinase CMGC-GSK family						
F-DMY	chr3_44341487	0.130	0.432	chr3.2	44341487	MS.gene027073	WD40/YVTN repeat-like-containing domain superfamily	MsNRG000206g00737911	hypothetical protein	MsNRG000206g00737801	Putative transcription factor WD40-like family				
F-DMY	chr3_89061410	0.095		chr3.2	89061410	MS.gene40203	ACP-like superfamily	MsNRG000661g01450031	Putative thimet oligopeptidase	MsNRG000350g01027551	Putative thimet oligopeptidase	MtrunA17Chr7g0240251	hypothetical protein		
F-DMY	chr3_61230888	0.101		chr3.2	61230888	MS.gene048728	Leucine-rich repeat domain superfamily;F-box domain	MsNRG000832g01614171	Putative F-box domain, leucine-rich repeat domain superfamily, F-box-like domain superfamily	MsNRG000832g01614181	Putative leucine-rich repeat domain superfamily				
F-DMY	chr2_2373521	0.137		chr2.2	2373521	MS.gene36359	Malic enzyme, N-terminal domain superfamily; Malic oxidoreductase	MsNRG000410g01125591	Putative malate dehydrogenase (oxaloacetate-decarboxylating) (NADP(+))	MsNRG000006g00057221	Putative malate dehydrogenase (oxaloacetate-decarboxylating) (NADP (+))	MtrunA17Chr2g0331051	Putative malate dehydrogenase (oxaloacetate-decarboxylating) (NADP (+))	MtrunA17Chr2g0331061	hypothetical protein
F-DMY	chr3_78458544	0.153		chr3.2	78458544	MS.gene06565	ABC transporter type 1, transmembrane domain	MsNRG000105g00468091	Putative Type I protein exporter	MsNRG000921g01687971	Putative ABC-type xenobiotic transporter	MtrunA17Chr3g0128361	Putative peptidase C78, ubiquitin modifier-specific peptidase 1/ 2	MtrunA17Chr3g0128381	Putative Type 1 protein exporter

This table summarizes the exact position of the QTL on the reference genome for each phenotypic trait, the percentage of phenotypic variance explained by the QTL (r^2) and the available annotation on the cv. XinJiangDaYe reference genome (Chen et al., 2020), the cv. Mercedes genome sequence (Carrère et al., 2020) and the corresponding region on the model species Medicago truncatula (version 5.1.8; Pecrix et al., 2018).

10.3389/fpls.2023.1196134

all traits, the genetic composition of the training population affected the predicting ability across traits in different ways (Figure 8). In order to observe the impact of the training population composition on the predicting ability, Figure 8 illustrates the difference between a prediction of accessions of one group with a training population including accessions from the other groups and a prediction of the same accessions with a training population of the same size (210) but including accessions from all groups. In the majority of cases, the prediction quality decreased when no accession from the targeted group was included in the training population. Nevertheless, in some cases, the difference was in favour of the prediction across groups for some of the group/traits. The groups 3 (France, Northern Europe) and 7 (Southern and Eastern Europe) seemed to be the most impacted by the absence of their own accessions in the training population. In contrast, group 6 (China) was little impacted by the genetic composition of the training population.

4 Discussion

We designed this study firstly to describe the genetic structure among currently grown alfalfa plant materials ('cultivated material') and plant breeding materials from the major European breeders ('breeding material'). Secondly, we wanted to test our hypothesis that breeding programs conducted within a narrow range of fall dormancy could induce a structure within worldwide breeding pools depending on fall dormancy as previously studied by Munjal et al., 2018. The third objective was to study the genetic determinism of major traits related to phenology such as flowering date and fall dormancy, and the possibility to predict it with genomic data.

4.1 Genotyping

Our study used a large number of markers (227 092 SNP) with few missing data (<5% per locus), in comparison to other studies in alfalfa (Annicchiarico et al., 2015b; Li et al., 2015b; Nazzicari et al., 2016; Jia et al., 2018; Medina et al., 2020). Previous studies on genomic prediction and genetic diversity showed that a large number of markers allowed to compensate for the amount of missing values (Heslot et al., 2013; Li et al., 2015b). However, a large amount of missing data can lead to a wrong estimation of linkage disequilibrium (Li et al., 2015b) and as a consequence to a biased estimation of the genomic relationship between accessions (Schopp et al., 2017). Genotypic imputation has shown promising results to complete missing values in different species (Marchini et al., 2007; Howie et al., 2009; Marchini and Howie, 2010; Daetwyler et al., 2011; Faville et al., 2018; Pégard et al., 2019). So far, the imputation methods proposed for tetraploid species (Nazzicari et al., 2016; Bastien et al., 2018) have shown interesting results to impute individual genotyping (allele dosage) but not for pool sequencing (allele frequency). These methods are not effective enough to be reliable and routinely used. The development of a method to impute genotypes based on population allele frequencies of tetraploid species could increase the number of useful markers. This would be of particular interest for alfalfa, which showed a very rapid decay of linkage disequilibrium between 1 kb and 20 kb in our case and in several other studies (Herrmann et al., 2010; Li et al., 2014). Indeed, in case of a short LD, a large number of SNPs is required to capture QTL and to obtain a good genomic prediction quality (Wientjes et al., 2013; Liu et al., 2015). A short LD can also influence the prediction methods that use LD to connect to putative QTL by feature selection (e.g., BayesB, Bayes Lasso) and decrease their effectiveness because of inaccurate LD estimation (Habier et al., 2007; Shengqiang et al., 2009; Jannink et al., 2010).

4.2 Diversity

With the large number of SNPs that we obtained, a genetic structure was revealed among cultivated material. In previous studies conducted with a limited number of SSR, AFLP or RAPD markers, genetic structure was observed only when a wide genetic diversity was studied (Crochemore et al., 1996; Riday et al., 2003; Qiang et al., 2015) but not when the diversity was restricted to breeding material (Flajoulot et al., 2005; Annicchiarico et al., 2016; Herrmann et al., 2018). We identified seven genetic groups among which two groups were clearly separated from the others. One group consisted of two accessions related to ssp *falcata*. This separation is consistent with another study showing that the ssp *falcata* are clearly separated from cultivated alfalfa (Li et al., 2014). The other group consisted of 15 Chinese accessions. Their clear separation from American and European accessions is more marked in our study than the separation observed in previous studies with SSR markers (Qiang et al., 2015) or genomic markers (Chen et al., 2020; Long et al., 2022). Within accessions of Europe, North and South America, a clear geographic structure was also obtained with partial overlap of the groups, finally showing a continuum. As all American germplasm originates from the "Ancient World", the complete overlap of European and American genetic diversity was expected, as also found by (Shen et al., 2020). Two reasons could explain our results: (1) some diversity from the "Ancient World" was not represented in our study, such as that of North Africa and Near or Middle East, (2) the selective pressure exerted by American breeders or a genetic drift generated a shift in the diversity. The wide range of dormancy within American materials is not in favour of the first explanation.

When using phenotypic traits related to fall dormancy and flowering date, no structure was observed. Except some of the Chinese accessions that showed a high fall dormancy, the other groups showed similar range of fall dormancy. Fall dormancy was shown to be a good indicator for genetic structure among accessions when the widest range of diversity is studied (Li et al., 2014) but was less efficient within intermediate range of fall dormancy, as studied here. Our results seemed to refute our hypothesis that preferential crossing within dormancy groups and strong selection pressure on dormancy could have induced genetic structure.

The majority of 'breeding material' provided by European breeders was genetically close to the groups that contained European 'cultivated material'. This indicates that cultivated material from North and South America and China are either not introduced in European breeding

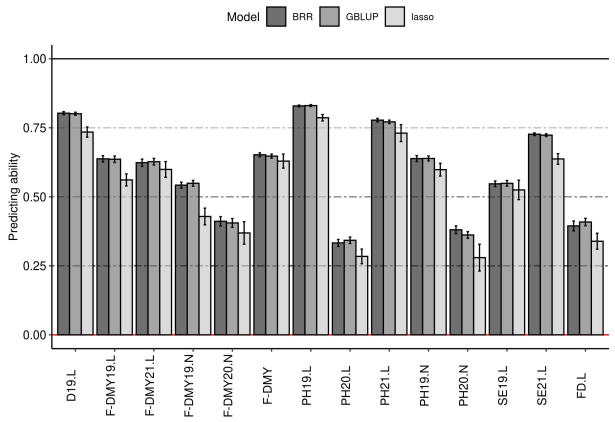


FIGURE 6
Impact of the model on the predicting ability (y-axis) for phenology traits. Three models were tested: best linear unbiased prediction model (GBLUP), Bayesian Ridge-regression (BRR) and Bayesian Lasso (lasso). The error bars are the standard deviation estimated on 10 repetitions. Traits were: Dormancy (D), Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE) for two years: 2019 (X19.X) and 2020 (X20.X) in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia and Flowering date (FD). F-DMY without letter or number is the Dry Matter Yield adjusted for year and location effects.

programs, or, if introduced, not retained after selection steps. This observation is an invitation to go more deeply into a phenotypic analysis of the cultivated material to exploit them in European, American and Chinese breeding programs.

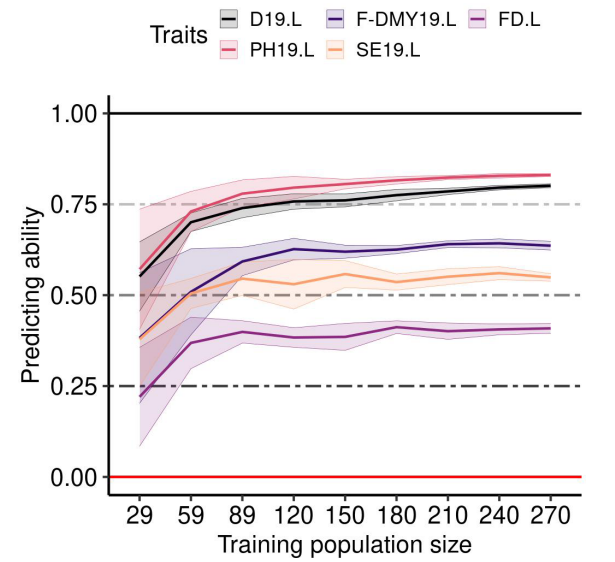


FIGURE 7
Impact of the training population size on the predicting ability for five traits (Flowering date (FD), Dormancy (D), Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE)) recorded in Lusignan in 2019. On the x-axis, the number of accessions used to train the model to predict a validation population composed of 100 accessions. The accessions were randomly taken among all the clusters. The average predicting ability (ten repetitions) estimated with the spearman correlation between the phenotype and its prediction is the middle solid line, the standard deviation is represented by the two solid lines above and below the middle line and colored by trait.

4.3 GWAS

We detected several QTL for traits related to fall dormancy on chromosomes 2, 3, 6, 7 and 8. Some of them explained a high percentage of phenotypic variation (>14%). Previous studies already evidenced QTL for fall dormancy. In a first attempt, QTL were found in a mapping population (Brouwer et al., 2000) but the assignment of linkage groups to current physical maps is not available. More recently, also in mapping populations, QTL were found for fall dormancy traits recorded in several environments on chr 1 and 7 (Li et al., 2015a), chr 1, 2, 3, 4, 5, 6, 7 (Adhikari et al., 2018), and chr 1, 7, 8 (Pecetti et al., 2021). In a GWAS, a QTL was found in a region of chr 7 that contained a Flowering locus T gene (MsFTa2), known to be part of the flowering pathway (Shen et al., 2020) and four QTL were found on chr 2, 3, 5, 6 (Long et al., 2022). A precise comparison of QTL positions was difficult because different genome references were used by the respective authors.

4.4 Genomic prediction

First of all, this study showed that it was possible to reach a high predicting ability (>0.75 for D19.L, PH19.L and PH21.L) for phenology related traits. However, this was not observed for all traits, the prediction capacity was around 0.30 for some traits (F-DMY20.N, PH20.L, PH20.N and FD.L). In our study, the difference between traits depended mainly on the year of measurement, with the year 2020 having a lower predicting ability on average, and on the location of measurement, with the prediction of traits measured in Novi Sad (Serbia) being less accurate than that measured in Lusignan (France). We do not have a clear explanation for these results, we suppose that it may be related to the difference between the continental climate (Novi Sad) and the oceanic climate (Lusignan) that is warmer in fall allowing a higher potential growth and therefore a better distinction of the fall dormancy of the accessions. Heritability was often presented (Luan et al., 2009; Lorenz et al., 2011; Clark et al., 2012; Kaler et al., 2022) as a factor that explained the difference for predicting ability between traits. In the present study, the quality of prediction could differ greatly for similar estimated heritability. This difference might be explained by the way heritability was expressed in our study, the variance explained by the spatial effect was integrated in the heritability formula, to avoid inflated heritabilities (Heckerman et al., 2016). Indeed, in Supplementary Figure S6, we have compared the relationship between predicting ability and heritability with and without taking into account the spatial variance in the heritability estimation. This showed that without taking into account the spatial variance in the estimation of heritability, the relationship between predicting ability and heritability was the same as that expressed in several studies (Luan et al., 2009; Lorenz et al., 2011; Clark et al., 2012; Kaler et al., 2022). However, (Heckerman et al., 2016) emphasized the importance of modelling environmental effects in the estimation of heritability and that the inflation was not stable for all traits, what was observed in our study as well. Even if the difference in heritability between traits did not allow to explain the difference in prediction quality, the traits with a high genetic

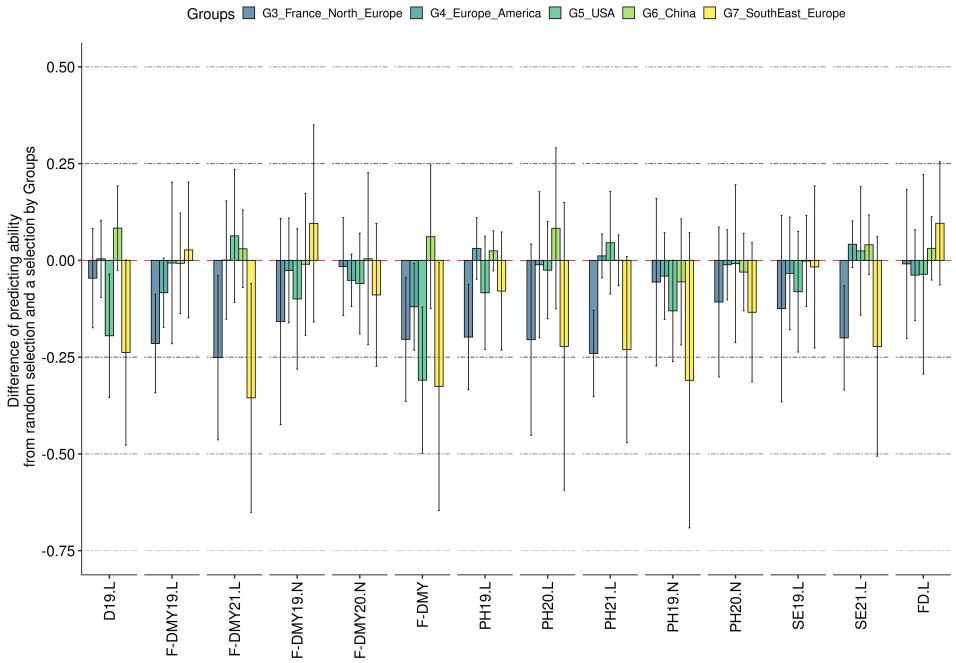


FIGURE 8
Impact of the training population composition on the predicting ability. The y-axis represents the difference of predicting ability between the predicting ability obtained when no other accessions from the same group were selected in the training population and the predicting ability when some accessions from the same group were in the training population. A negative predicting ability means that the predicting ability obtained with accessions of the same group in the training population was greater than the predicting ability when no accession of the same group was in the training population. The error bars are the standard deviation estimated on 10 repetitions. The traits are represented in the x-axis: Flowering date (FD), Dormancy (D), Dry Matter Yield (F-DMY), plant height (PH), Speed of elongation (SE) for two years: 2019 (X19.X) and 2020 (X20.X) in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia. F-DMY without letter or number is the Dry Matter Yield adjusted for year and location effects.

correlation (D19.L-PH19.L- F-DMY19.L-PH21.L-SE21.L and F-DMY20.N-PH20.N) showed a similar predicting ability.

In this study, the different GP models showed similar results, which was expected for GBLUP and BRR, but not necessarily for the Bayesian Lasso, as the latter selected the SNPs having an effect in the model and was supposed to perform better in the case of strong QTL effects (Meher et al., 2022). However, our results were consistent with those obtained in other studies in alfalfa for various trait (Annicchiarico et al., 2015b; Biazzi et al., 2017; Jia et al., 2018; Medina et al., 2020; Zhang et al., 2023).

Similar to other authors (Nakaya and Isobe, 2012; Tayeh et al., 2015; Cericola et al., 2017), increasing the number of accessions in the training population increased the prediction quality until stabilisation occurred with training populations of 89 accessions. The additional accessions added after this point were useful in reducing the variation due to sampling. By setting the size of the training and validation populations, we ensured that this did not affect the quality of the prediction between the groups. The same applied to the composition of the training population (Lorenz and Smith, 2015; Tayeh et al., 2015; Norman et al., 2018; Akdemir and Isidro-Sánchez, 2019; Pégard et al., 2020). The predicting ability observed in our study was higher and less variable when the training and validation populations were related and when all the groups were represented in the training population, than in the case of across groups prediction.

Our study has shown that for traits such as fall dormancy or flowering date, we can achieve good genomic prediction quality in a diverse panel. These results, combined with other studies on alfalfa (Annicchiarico et al., 2015b; Biazzi et al., 2017; Jia et al., 2018; Andrade et al., 2022; Zhang et al., 2023), show that genomic selection is an interesting and efficient lever in alfalfa breeding. It is possible to rethink the alfalfa improvement scheme by using genomic prediction. It allows to play with several genetic gain parameters such as selection intensity, cycle length and genetic diversity. The value of genomic prediction has been demonstrated, even with a low predicting ability, to increase genetic gain compared to the current selection method (Annicchiarico et al., 2017). Currently, breeders have several breeding populations, as they tend to select within a fall dormancy group. Our study shows that with genomic prediction, this selection process can be simplified by mixing different fall dormancy groups within a single breeding population. Varieties can be selected for fall dormancy during the breeding cycle using molecular markers. Our results also show that flowering date can be predicted and used to optimise seed crosses between parents to increase genetic gain and manage genetic diversity, as has been proposed in other species (Tiret et al., 2021). Advances in genomics-assisted breeding offer exceptional prospects for improving alfalfa's performance.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

MP contributed to the genotyping bioinformatics pipeline, conceived and performed the analyses and wrote the paper draft. SD performed DNA extraction, prepared libraries for sequencing. FS contributed to data collection. DK conceived the project with BJ. DM defined the trial experiments and plant material with BJ and coordinated data collection in Novi Sad. MZ set up the experimental augmented design. TR helped with bioinformatics analysis. PB covered the aspects of GWAS and genomic prediction on the project. BJ conceived the project and obtained financial support, contributed to data collection and analyses. All authors participated to the draft and revision of the manuscript and approved the final version. All authors contributed to the article and approved the submitted version.

Funding

This project has received funding from the European Union's Horizon 2020 Program for Research & Innovation under grant agreement n°727312, project EUCLEG.

Acknowledgments

We thank the seed companies and institutes that provided us with seeds of their accessions: AgriObtentions, Alforex, ART, Barenbrug, Cérience, CGS Sementi, CREA, DLF Seeds, Forage Genetics, Foratec Argentina, IFVCNS, IKBKS, INTA, Lantmännen, NordGen, RAGT-2n. The technical staffs of INRAE (URP3F, FERLUS and GeT-PlaGe) and IFVCNS are thanked for their deep investment to collect genotyping and phenotyping data. We would like to thank Rodrigue Veron in particular for his investment in the phenotyping of fall dormancy. We would like to thank Sébastien Blugeon for his investment in the design and implementation of the bioinformatics pipeline that enabled the genotyping of accessions. We thank the five breeders who provided us with part of their breeding pool.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1196134/full#supplementary-material

References

Adhikari, L., Lindstrom, O. M., Markham, J., and Missaoui, A. M. (2018)QTLs associated with fall dormancy & winter-hardiness in alfalfa shows potential for independent improvement of the two traits. In: *ASA-CSSA-SSSA*. Available at: https://scisoc.confex.com/scisoc/2018am/meetingapp.cgi/Paper/112970 (Accessed 26, 2022).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control.* 19 (6), 716–723.

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* 9, 1446. doi: 10.1038/s41598-018-38081-6

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel Evol.* 47, 38. doi: 10.1186/s12711-015-0116-6

Andrade, M., Acharya, J., Benevenuto, J., Oliveira, I., Lopez, Y., Munoz, P., et al. (2022). Genomic prediction for canopy height and dry matter yield in alfalfa using family bulks. *Plant Genome* 15, 1–16. doi: 10.1002/tpg2.20235

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., and Marshall, A. H. (2015a). Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* 34, 327–380. doi: 10.1080/07352689.2014.898462

Annicchiarico, P., Nazzicari, N., Ananta, A., Carelli, M., Wei, Y., and Brummer, E. (2016). Assessment of cultivar distinctness in alfalfa: a comparison of genotyping-by-Sequencing, simple-sequence repeat marker, and morphophysiological observations. *Plant Genome* 9, 1–12. doi: 10.3835/plantgenome2015.10.0105

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015b). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16, 1020. doi: 10.1186/s12864-015-2212-y

Annicchiarico, P., Nazzicari, N., Wei, Y., Pecetti, L., and Brummer, E. C. (2017). Genotyping-by-Sequencing and its exploitation for forage and cool-season grain legume breeding. *Front. Plant Sci.* 8, 679. doi: 10.3389/fpls.2017.00679

Ashraf, B. H., Jensen, J., Asp, T., and Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *Theor. Appl. Genet.* 127, 1331–1341. doi: 10.1007/s00122-014-2300-4

Basigalup, D., Irwin, J., Mi, F., and Abdelguerfi-Laouar, M. (2014). Perspectives of alfalfa in australia, china, africa and latin america. *Legume Perspect.* 4, 9–10.

Bastien, M., Boudhrioua, C., Fortin, G., and Belzile, F. (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. *Genome* 61, 449–456. doi: 10.1139/gen-2017-0236

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., et al. (2017). Genome-wide association mapping and genomic selection for alfalfa (Medicago sativa) forage quality traits. *PLoS One* 12, e0169234. doi: 10.1371/journal.pone.0169234

Blondon, F., MT, C., and Guy, P. (1967). Action de photopériode et de thermopériodes sur le comportement de deux types extrêmes de luzernes françaises. *Comptes Rendus Hebdomadaires Des. Séances l'Académie Des. Sciences série D* 264, 596.

Brouwer, D. J., Duke, S. H., and Osborn, T. C. (2000). Mapping genetic factors associated with winter hardiness, fall growth, and freezing injury in autotetraploid alfalfa. *Crop Sci.* 40, 1387–1396. doi: 10.2135/cropsci2000.4051387x

Brummer, E. C., Shah, M. M., and Luth, D. (2000). Reexamining the relationship between fall dormancy and winter hardiness in alfalfa. *Crop Sci.* 40, 971–977. doi: 10.2135/cropsci2000.404971x

Calus, M., and Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43, 26. doi: 10.1186/1297-9686-43-26

Cantet, R. J. C., Birchmeier, A. N., Cayo, A. C., and Fioretti, C. (2005). Semiparametric animal models *via* penalized splines as alternatives to models with contemporary groups. *J. Anim. Sci.* 83, 2482–2494. doi: 10.2527/2005.83112482x

Cappa, E. P., and Cantet, R. J. C. (2007). Bayesian Estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi: 10.1139/X07-116

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11, 1–15. doi: 10.1007/s11295-015-0917-3

Carrère, S., Gouzy, J., Debellé, F., Marande, W., Julier, B., and Barre, P. (2020). "Medicago sativa cv," in *Mercedes Genome sequence*. Available at: https://hal.inrae.fr/hal-02993163v2/document.

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PloS One* 12, e0169606. doi: 10.1371/journal.pone.0169606

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., et al. (2018). Optimized use of low-depth genotyping-by-Sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (Lolium perenne l.). *Front. Plant Sci.* 9, 369. doi: 10.3389/fpls.2018.00369

Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., et al. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* 11, 1–11. doi: 10.1038/s41467-020-16338-x

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44, 4. doi: 10.1186/1297-9686-44-4

Crochemore, M., Huyghe, C., Kerlan, M.-C., Durand, F., and Julier, B. (1996). Partitioning and distribution of RAPD variation in a set of populations of the medicago sativa complex. *Agronomie* 16, 421–432. doi: 10.1051/agro:19960702

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

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., and Goddard, M. E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics* 189, 317–327. doi: 10.1534/genetics.111.128082

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, 10. doi: 10.1371/journal.pone.0019379

Faville, M. J., Ganesh, S., Cao, M., Jahufer, M. Z. Z., Bilton, T. P., Easton, H. S., et al. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor. Appl. Genet.* 131, 703–720. doi: 10.1007/s00122-017-3030-1

Federer, W. T., and Raghavarao, D. (1975). On augmented designs. *Biometrics* 31, 29. doi: 10.2307/2529707

Flajoulot, S., Ronfort, J., Baudouin, P., Barre, P., Huguet, T., Huyghe, C., et al. (2005). Genetic diversity among alfalfa (Medicago sativa) cultivars coming from a breeding program, using SSR markers. *Theor. Appl. Genet.* 111, 1420–1429. doi: 10.1007/s00122-005-0074-4

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S.IV (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. Soft.* 33 (1), 1–22. doi: 10.18637/jss.v033.i01

Grattapaglia, D., and Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi: 10.1007/s11295-010-0328-4

Grünwald, N. J., and Goss, E. M. (2011). Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches. *Annu. Rev. Phytopathol.* 49, 249–267. doi: 10.1146/annurev-phyto-072910-095246

Gupta, S., Garg, V., Kant, C., and Bhatia, S. (2015). Genome-wide survey and expression analysis of f-box genes in chickpea. *BMC Genomics* 16, 67. doi: 10.1186/s12864-015-1293-y

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci.* 113, 7377–7382. doi: 10.1073/pnas.1510497113

Herrmann, D., Barre, P., Santoni, S., and Julier, B. (2010). Association of a CONSTANS-LIKE gene to flowering and height in autotetraploid alfalfa. *Theor. Appl. Genet.* 121, 865–876. doi: 10.1007/s00122-010-1356-z

Herrmann, D., Flajoulot, S., Barre, P., Huyghe, C., Ronfort, J., and Julier, B. (2018). Comparison of morphological traits and molecular markers to analyse diversity and structure of alfalfa (Medicago sativa l.) cultivars. *Genet. Resour. AND Crop Evol.* 65, 527–540. doi: 10.1007/s10722-017-0551-z

Heslot, N., Rutkoski, J., Poland, J., Jannink, J. L., and Sorrells, M. E. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PloS One* 8.9 (2013), e74612. doi: 10.1371/journal.pone.0074612

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PloS Genet.* 5, e1000529. doi: 10.1371/journal.pgen.1000529

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128, 145–158. doi: 10.1007/s00122-014-2418-4

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

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., et al. (2018). Genomic prediction for 25 agronomic and quality traits in alfalfa (Medicago sativa). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01220

Jombart, T. (2008). Adegenet: a r package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

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

Julier, B., Barre, P., Lambroni, P., Delaunay, S., Thomasset, M., Lafaillette, F., et al. (2018). Use of GBS markers to distinguish among lucerne varieties, with comparison to morphological traits. *Mol. Breed.* 38, 133. doi: 10.1007/s11032-018-0891-1

Julier, B., Blugeon, S., Delaunay, S., Mappa, G., Ruttink, T., Pégard, M., et al. (2021). Optimisation of GBS protocols for efficient genotyping of forage species. *In Eucarpia-Section Fodder Crops and Amenity Grasses Meeting*. doi: 10.5507/vup.21.24459677.17

Julier, B., Gastal, F., Louarn, G., Badenhausser, I., Annicchiarico, P., Crocq, G., et al. (2017). "Lucerne (alfalfa) in European cropping systems," in *Legumes in cropping systems*. Eds. D. Murphy-Bokern, F. L. Stoddard and C. A. Watson (Wallingford: CABI), 168–192. doi: 10.1079/9781780644981.0168

Julier, B., Huyghe, C., and Ecalle, C. (2000). Within- and among-cultivar genetic variation in alfalfa: forage quality, morphology, and yield. *Crop Sci.* 40, 365–369. doi: 10.2135/cropsci2000.402365x

Kaler, A. S., Purcell, L. C., Beissinger, T., and Gillman, J. D. (2022). Genomic prediction models for traits differing in heritability for soybean, rice, and maize. *BMC Plant Biol.* 22, 87. doi: 10.1186/s12870-022-03479-y

Khanna, A., Larson, D., Srivatsan, S., Mosior, M., Abbott, T., Kiwala, S., et al. (2022). Bam-readcount - rapid generation of basepair-resolution sequence metrics. *JOSS* 7, 3722. doi: 10.21105/joss.03722

Kujur, A., Upadhyaya, H. D., Bajaj, D., Gowda, C., Sharma, S., Tyagi, A. K., et al. (2016). Identification of candidate genes and natural allelic variants for QTLs governing plant height in chickpea. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep27968

Kumar, M., Chauhan, A. S., Kumar, M., Yusuf, M. A., Sanyal, I., and Chauhan, P. S. (2019). Transcriptome sequencing of chickpea (Cicer arietinum l.) genotypes for identification of drought-responsive genes under drought stress condition. *Plant Mol. Biol. Rep.* 37, 186–203. doi: 10.1007/s11105-019-01147-4

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : an r package for multivariate analysis. *J. Stat. Soft.* 25 (1), 1–18. doi: 10.18637/jss.v025.i01

Lesins, K. A., and Lesins, I. (2012). *Genus medicago (Leguminosae): a taxogenetic study* (Springer Science & Business Media).

Li, X. (2013)Development and use of a high density SNP array to evaluate genetic diversity in alfalfa. In: *Plant and animal genome*. Available at: https://pag.confex.com/pag/xxi/webprogram/Paper7784.html (Accessed 9, 2021).

Li, X., Alarcon-Zuniga, B., Kang, J., Tahir, M., Jiang, Q., Wei, Y., et al. (2015a). Mapping fall dormancy and winter injury in tetraploid alfalfa. *Crop Sci.* 55, 1995–2011. doi: 10.2135/cropsci2014.12.0834

Li, X., Wei, Y., Acharya, A., Hansen, J., Crawford, J., Viands, D., et al. (2015b). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8. doi: 10.3835/plantgenome2014.12.0090

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014). A saturated genetic linkage map of autotetraploid alfalfa (Medicago sativa l.) developed using genotyping-by-sequencing is highly syntenous with the medicago truncatula genome. *G3: genes genomes Genet.* 4, 1971–1979. doi: 10.1534/g3.114.012245

Lin, C.-S., and Poushinsky, G. (1985). A modified augmented design (type 2) for rectangular plots. *Can. J. Plant Sci.* 65, 743–749. doi: 10.4141/cjps85-094

Lin, C.-Y., Xing, G., and Xing, C. (2012). Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity (Edinb).* 109, 401–402. doi: 10.1038/hdy.2012.54

Liu, Z., Ren, Z., Yan, L., and Li, F. (2022). DeepLRR: an online webserver for leucine-Rich-Repeat containing protein characterization based on deep learning. *Plants* 11, 136. doi: 10.3390/plants11010136

Liu, H., Zhou, H., Wu, Y., Li, X., Zhao, J., Zuo, T., et al. (2015). The impact of genetic relationship and linkage disequilibrium on genomic selection. *PLoS One* 10, e0132379. doi: 10.1371/journal.pone.0132379

Long, R., Zhang, F., Zhang, Z., Li, M., Chen, L., Wang, X., et al. (2022). Genome assembly of alfalfa cultivar zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics Proteomics Bioinf.* 20 (1), 14–28. doi: 10.1016/j.gpb.2022.01.002

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). Genomic selection in plant breeding. knowledge and prospects. *Adv. Agron.* 110, 77–123. doi: 10.1016/B978-0-12-385531-2.00002-5

Lorenz, A. J., and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci.* 55, 2657–2667. doi: 10.2135/cropsci2014.12.0827

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183, 1119–1126. doi: 10.1534/genetics.109.107391

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb).* 108, 285–291. doi: 10.1038/hdy.2011.73

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi: 10.1038/nrg2796

Marchini, J., Howie, B. N., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

Martin, G., Durand, J.-L., Duru, M., Gastal, F., Julier, B., Litrico, I., et al. (2020). Role of ley pastures in tomorrow's cropping systems. *A review. Agron. Sustain. Dev.* 40, 17. doi: 10.1007/s13593-020-00620-9

Medina, C., Hawkins, C., Liu, X., Peel, M., and Yu, L. (2020). Genome-wide association and prediction of traits related to salt tolerance in autotetraploid alfalfa (Medicago sativa l.). *Int. J. Mol. Sci.* 21. doi: 10.3390/ijms21093361

Medina, C., Kaur, H., Ray, I., and Yu, L. (2021). Strategies to increase prediction accuracy in genomic selection of complex traits in alfalfa (Medicago sativa l.). *CELLS* 10. doi: 10.3390/cells10123372

Meher, P. K., Rustgi, S., and Kumar, A. (2022). Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. *Heredity* 128, 519–530. doi: 10.1038/s41437-022-00539-9

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. doi: 10.1093/genetics/157.4.1819

Michaud, R., Lehman, W. F., and Rumbaugh, M. D. (1988). World distribution and historical development. *Alfalfa alfalfa improvement* 29, 25–91. doi: 10.2134/agronmonogr29.c2

Muller, M.-H., Poncet, C., Prosperi, J.-M., Santoni, S., and Ronfort, J. (2006). Domestication history in the medicago sativa species complex: inferences from nuclear sequence polymorphism. *Mol. Ecol.* 15, 1589–1602. doi: 10.1111/j.1365-294X.2006.02851.x

Munjal, G., Hao, J., Teuber, L. R., and Brummer, E. C. (2018). Selection mapping identifies loci underpinning autumn dormancy in alfalfa (Medicago sativa). *G3: genes genomes Genet.* 8, 461–468. doi: 10.1534/g3.117.300099

Muñoz, F., and Sanchez, L. (2020) *breedR: statistical methods for forest genetic resources analysts*. Available at: https://github.com/famuvie/breedR.

Nakaya, A., and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110, 1303–1316. doi: 10.1093/aob/mcs109

Nazzicari, N., Biscarini, F., Cozzi, P., Brummer, E., and Annicchiarico, P. (2016). Marker imputation efficiency for genotyping-by-sequencing data in rice (Oryza sativa) and alfalfa (Medicago sativa). *Mol. Breed.* 36, 69. doi: 10.1007/s11032-016-0490-y

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 G3* 200311, 2018. doi: 10.1534/g3.118.200311

Pecetti, L., Barre, P., Delaunay, S., Lambroni, P., Annicchiarico, P., and Julier, B. (2021). QTL analysis for grazing tolerance, autumn dormancy and growth habit offers prospects for marker-assisted selection in lucerne. *Euphytica* 217, 171. doi: 10.1007/s10681-021-02897-7

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., et al. (2018). Whole-genome landscape of medicago truncatula symbiotic genes. *Nat. Plants* 4, 1017–1025. doi: 10.1038/s41477-018-0286-7

Pégard, M., Leuenberger, J., Julier, B., and Barre, P. (2021). Genomic prediction of lucerne forage yield and quality. *In Eucarpia-Section Fodder Crops and Amenity Grasses Meeting* doi: 10.5507/vup.21.24459677.25

Pégard, M., Rogier, O., Bérard, A., Faivre-Rampant, P., Paslier, M.-C. L., Bastien, C., et al. (2019). Sequence imputation from low density single nucleotide polymorphism panel in a black poplar breeding population. *BMC Genomics* 20, 1–16. doi: 10.1186/s12864-019-5660-y

Pégard, M., Segura, V., Muñoz, F., Bastien, C., Jorge, V., and Sanchez, L. (2020). Favorable conditions for genomic evaluation to outperform classical pedigree evaluation highlighted by a proof-of-concept study in poplar. *Front. Plant Sci.* 1552. doi: 10.3389/fpls.2020.581954

Pembleton, L. W., Cogan, N. O. I., and Forster, J. W. (2013). St AMPP : an r package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol. Ecol. Resour.* 13, 946–952. doi: 10.1111/1755-0998.12129

Pérez, P., Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in r. *Plant Genome 3 plantgenome* 2010, 4.0005. doi: 10.3835/plantgenome2010.04.0005

Poux, X., and Aubert, P.-M. (2018). An agroecological Europe in 2050: multifunctional agriculture for healthy eating. *Findings Ten Years For Agroecology (TYFA) Model. exercise Iddri-AScA Study* 9, 18.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Qiang, H., Chen, Z., Zhang, Z., Wang, X., Gao, H., and Wang, Z. (2015). Molecular diversity and population structure of a worldwide collection of cultivated tetraploid alfalfa (Medicago sativa subsp sativa l.) germplasm as revealed by microsatellite markers. *PloS One* 10. doi: 10.1371/journal.pone.0124592

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

Riday, H., Brummer, E. C., Campbell, T. A., Luth, D., and Cazcarro, P. M. (2003). Comparisons of genetic and morphological distance with heterosis between medicago sativa subsp. sativa and subsp. falcata. *Euphytica* 131, 37–45. doi: 10.1023/A:1023050126901

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics* 192, 715–728. doi: 10.1534/genetics.112.141473

Robbins, K. R., Backlund, J. E., and Schnelle, K. D. (2012). Spatial corrections of unreplicated trials using a two-dimensional spline. *Crop Sci.* 52, 1138–1144. doi: 10.2135/cropsci2011.08.0417

Rubiales, D., Annicchiarico, P., Vaz Patto, M. C., and Julier, B. (2021). Legume breeding for the agroecological transition of global agri-food systems: a European perspective. *Front. Plant Sci.* 2548. doi: 10.3389/fpls.2021.782574

Sakiroglu, M., and Brummer, E. C. (2017). Identification of loci controlling forage yield and nutritive value in diploid alfalfa using GBS-GWAS. *Theor. Appl. Genet.* 130, 261–268. doi: 10.1007/s00122-016-2782-3

Schaumont, D. (2020) *GBprocesS: genotyping-by-Sequencing data processing toolkit*. Available at: https://gitlab.com/dschaumont/GBprocesS.

Schaumont, D., Veeckman, E., van der Jeugt, F., Haegeman, A., van Glabeke, S., Bawin, Y., et al. (2022). Stack mapping anchor points (SMAP): a versatile suite of tools for read-backed haplotyping. *bioRxiv*. doi: 10.1101/2022.03.10.483555

Schopp, P., Müller, D., Technow, F., and Melchinger, A. E. (2017). Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. *Genetics* 205, 441–454. doi: 10.1534/genetics.116.193243

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. doi: 10.1038/ng.2314

Shen, C., Du, H., Chen, Z., Lu, H., Zhu, F., Chen, H., et al. (2020). The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Mol. Plant* 13, 1250–1261. doi: 10.1016/j.molp.2020.07.003

Shengqiang, Z., Dekkers, J. C. M., 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: 10.1534/genetics.108.098277

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model *via* coordinate descent. *J. Stat. Soft.* 39 (5), 1–13. doi: 10.18637/jss.v039.i05

Singh, P. K., Indoliya, Y., Agrawal, L., Awasthi, S., Deeba, F., Dwivedi, S., et al. (2022). Genomic and proteomic responses to drought stress and biotechnological interventions for enhanced drought tolerance in plants. *Curr. Plant Biol.* 29, 100239. doi: 10.1016/j.cpb.2022.100239

Small, E. (2011). *Alfalfa and relatives: evolution and classification of medicago* (NRC Research Press).

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Front. Plant Sci.* 6, 1–11. doi: 10.3389/fpls.2015.00941

Teuber, L. R., and Brick, M. A. (1988). Morphology and anatomy. In *Alfalfa and Alfalfa Improvement* (eds A. A. Hanson, D. K. Barnes and R. R. Hill). doi: 10.2134/agronmonogr29.c4

Teuber, L., Taggard, K., Gibbs, L., McCaslin, M., Peterson, M., and Barnes, D. (1998). Fall dormancy. standard tests to characterize alfalfa cultivars, a-1.

Tibshirani, R. (1996). Regression shrinkage and selection *Via* the lasso. *J. R. Stat. Society: Ser. B (Methodological)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., et al. (2012). Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Society: Ser. B (Statistical Methodology)* 74, 245–266. doi: 10.1111/j.1467-9868.2011.01004.x

Tiret, M., Pégard, M., and Sánchez, L. (2021). How to achieve a higher selection plateau in forest tree breeding? fostering heterozygote× homozygote relationships in

optimal contribution selection in the case study of populus nigra. *Evolutionary Appl.* 14, 2635–2646. doi: 10.1111/eva.13300

UPOV (2005). *Guidelines for the conduct of tests for distinctness, uniformity and stability*. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwig4KHIuYX6AhXPRvEDHVU3A4sQFnoECBAQAQ&url=https%3A%2F%2Fwww.upov.int%2Fedocs%2Ftgdocs%2Fen%2Ftg006.doc&usg=AOvVaw2PqWNLKlNiWrlv0EhoJDxl (Accessed 3, 2022).

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

Verrier, P. J., Bird, D., Burla, B., Dassa, E., Forestier, C., Geisler, M., et al. (2008). Plant ABC proteins – a unified nomenclature and updated inventory. *Trends Plant Sci.* 13, 151–159. doi: 10.1016/j.tplants.2008.02.001

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research* 75(2), 249–252.

Wientjes, Y. C. J., Veerkamp, R. F., and Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi: 10.1534/genetics.112.146290

Willame, J., Gensollen, V., Montegano, B., and Baril, C. (2002). The use of fall dormancy for distinguishing between varieties of lucerne (Medicago sativa l.). *Plant Varieties Seeds* 14, 133–141.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., and Schön, C.-C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587. doi: 10.1534/genetics.113.150078

Xu, C., and Min, J. (2011). Structure and function of WD40 domain proteins. *Protein Cell* 2, 202–214. doi: 10.1007/s13238-011-1018-1

Zhang, F., Kang, J., Long, R., Li, M., Sun, Y., He, F., et al. (2023). Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Horticulture Res.* 10, uhac225. doi: 10.1093/hr/uhac225

Zhao, Z., Fan, J., Yang, P., Wang, Z., Opiyo, S. O., Mackey, D., et al. (2022). Involvement of arabidopsis acyl carrier protein 1 in PAMP-triggered immunity. *MPMI* 35, 681–693. doi: 10.1094/MPMI-02-22-0049-R

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 August 2023
DOI 10.3389/fpls.2023.1236147

Check for updates

OPEN ACCESS

EDITED BY
Leif Skot,
Aberystwyth University, United Kingdom

REVIEWED BY
Elsayed Mansour,
Zagazig University, Egypt
Odd Arne Rognli,
Norwegian University of Life Sciences, Norway

*CORRESPONDENCE
Christiane Balko
christiane.balko@julius-kuehn.de

RECEIVED 07 June 2023
ACCEPTED 14 August 2023
PUBLISHED 30 August 2023

CITATION
Balko C, Torres AM and Gutierrez N (2023) Variability in drought stress response in a panel of 100 faba bean genotypes. *Front. Plant Sci.* 14:1236147. doi: 10.3389/fpls.2023.1236147

COPYRIGHT
© 2023 Balko, Torres and Gutierrez. 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.

Variability in drought stress response in a panel of 100 faba bean genotypes

Christiane Balko[1]*, Ana M. Torres[2] and Natalia Gutierrez[2]

[1]Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Sanitz, Germany, [2]Área de Mejora Vegetal y Biotecnología, Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA), Centro Alameda del Obispo, Córdoba, Spain

Faba bean is an important protein crop for food and feed worldwide and provides a range of advantages in crop rotations. Its limited use in modern agriculture is mainly due to the high fluctuations in yield. A well known limiting factor in most legumes, and particularly in faba bean, is the high sensitivity to water shortage, which is further aggravated by climate change. The present study was undertaken to exploit the genetic variation in drought stress response in a faba bean collection of 100 accessions with diverse origins and to assess selection criteria for identifying drought tolerant genotypes. Physiological, phenological and yield related traits evaluated under drought or water-sufficient conditions responded significantly to the end-terminated drought stress. Comparison of yield relations showed the advantage of using a stress tolerance index (STI) to identify genotypes combining high yield potential with high stress yield. With regard to physiological traits, SPAD (chlorophyll content) values were significantly related to yield as well as to STI, while the other traits also contributed to different extents to variation in yield formation. Among the yield related traits, seeds per plant proved to be the most important trait followed by pods per plant. Interestingly, the eight genotypes with the best STI performance use different strategies to cope with drought stress.

KEYWORDS

faba bean, drought stress, phenotyping, physiological traits, senescence, genetic diversity, heritability, yield

Introduction

Grain legumes are a major source of plant protein for food and feed worldwide and provide multiple advantages in crop rotations regarding soil fertility, plant health and sustainability. Faba bean (*Vicia faba* L.) is one of the most widely distributed crops, being grown both as a grain (pulse) and green-manure legume (Jensen et al., 2010). With a production varying between 4.5 and 5.5 Mt in the last decade, faba bean ranks 6th in terms of world production of pulse crops (FAOSTAT, 2022). China is the largest producer followed by Ethiopia, the United Kingdom, Australia, France and Germany. In the last 70

years, there has been more than 50% decline in the global cropping area of the faba bean, which is mainly caused by poor yield stability owing to climate variability, diseases, weeds and other pests, which represent the major constraints of faba bean production.

Faba bean yield fluctuations are likely to increase with the predicted climate change in this century. Increasing atmospheric CO2 concentrations will change surface temperatures and precipitation patterns, causing more periods of extreme precipitation and drought (Farooq et al., 2014). Therefore, drought stress has become one of the most uncontrollable and unpredictable agricultural challenges and is today one of the major yield limiting factors in grain legume crops (Swann et al., 2016; Nadeem et al., 2019). Improving drought tolerance is a key strategy to enhance performance and stability of yield in these crops. Nevertheless, the complexity of the trait, the lack of efficient selection protocols and the mixed faba bean breeding system pose major challenges for effective implementation in plant breeding programs (Muktadir et al., 2020).

Legumes respond to drought with morphological, physiological and biochemical changes in roots, stems and leaves (Shakeel et al., 2011). Drought stress results in many interactive modifications including changes in the expression of drought-resistance genes, the synthesis of hormones, the overproduction of reactive oxygen species and the osmotic adjustments through active ions or organic compound such as proline and carbohydrates (Farooq et al., 2012; Kaur and Asthir, 2015; Wahab et al., 2022). Plants minimize water loss by closing stomata and reducing light absorbance, canopy leaf area and peaked water absorption.

The extent and type of responses will depend on the intensity and length of the stress, but in all cases grain yield is substantially reduced. A meta-analysis of legume yield responses to drought under field conditions from 1980 to 2014 revealed a yield reduction of 40% following a 65% reduction in water availability, with cultivar and environmental factors being important cofactors (Daryanto et al., 2015). The study confirmed, that legumes have a high yield potential, but respond in a sensitive way to water shortage (Müller et al., 1985; Grashoff, 1990; Khan et al., 2010; Torres et al., 2010; Redden et al., 2014; Muktadir et al., 2020). Faba bean is more sensitive to drought than other field crops and largely exceeds other grain legumes such as common bean, pea and chickpea (McDonald and Paulsen, 1997; Amede and Schubert, 2003; Manning et al., 2020).

Drought impairs grain legume yield at all growth stages. Germination, leaf area and photosynthetic activity are significantly reduced (Nadeem et al., 2019). However, the most sensitive period is the reproductive stage where drought stress leads to earlier flowering, and ultimately, reduced pod and grain set (Plies-Balzer et al., 1995; Khan et al., 2007; Alghamdi et al., 2015). Against this background, improvement of drought tolerance is an important faba bean breeding goal. Understanding of drought response patterns and associated traits are key factors for achieving higher yield stability under stress conditions. However, compared to other crops, improvement of productivity under drought stress has rarely been included in faba bean breeding programmes. Breeding progress has been relatively slow so far due to the limited number of genotypes included in the studies (Abdelmula et al., 1999), the low heritability of advantageous traits and the lack of efficient and reproducible screening methods (Stoddard et al., 2006; Muktadir et al., 2020). However, a wide genotypic variation in faba bean water stress response has been reported (Link et al., 1999; Amede et al., 2001; Ricciardi et al., 2001; Khan et al., 2007), indicating a high potential for breeding in drought-prone environments. To exploit this genetic variation, the application of an accurate and relevant phenotyping method plays a key role for the selection of drought-resilient genotypes and for the dissection and genetic analysis of a complex trait such as the adaptive response of crops to drought (Tuberosa, 2012).

Phenotypic attributes are the most frequently used criteria used for identifying drought-tolerant genotypes. In legumes, selection for drought tolerance based on highly heritable morphological and phenological traits, together with physiological attributes such as the accumulation of proline or soluble sugars, has proven highly successful for screening genotypes under limited water supply (Lafitte et al., 2003; Richards, 2006; Stoddard et al., 2006; Annicchiarico and Iannucci, 2008; Alderfasi and Alghamdi, 2010; Ammar et al., 2015; Ali et al., 2016). However, phenotyping under field conditions is time- and labour-consuming, and the reproducibility of water stress conditions is often poor, because of variations in the timing of onset, duration and severity of the drought (Khan et al., 2007). The recent use of a controlled water supply combined with rainout shelters reduces these uncertainties and generates results that are more accurate. These tools allow phenotyping of large field populations under conditions of adequate light intensity and quality, avoiding the effects of unpredictable rainfall patterns.

Direct selection for grain yield under drought conditions often proves inefficient to identify stress-tolerant genotypes. For this reason, several selection indices have been developed to evaluate yield stability based on grain yield under normal and stress conditions (reviewed in Sofi et al., 2018). Among these, the stress tolerance index (STI) defined by Fernandez (1992) has been consistently correlated with other indexes (Farshadfar et al., 2013; Mansour et al., 2021; Sharifi et al., 2021; Memari et al., 2022), indicating that it can be used as an alternative to select drought tolerant genotypes with high yield performance in stress and non-stress conditions.

Considering the scarce number of studies so far reported and the limited number of faba bean genotypes used, breeding progress for drought stress tolerance in faba bean has been relatively slow. The goal of the present study was to exploit the genetic variation of a wider collection of 100 faba bean accessions with diverse origins to assess the selection criteria for identifying drought tolerant genotypes and to dissect a complex trait such as the adaptive response to drought. Thus, our main aims were: (i) to evaluate the performance of this faba bean collection, one of the largest screenings reported so far, under water-sufficient and drought stress conditions and (ii) to investigate the relevance of morphological and phenological traits with respect to yield and yield stability in water stress conditions in an effort to identify and to improve selection efficiency under drought stress environments.

Materials and Methods

Plant material and experimental design

The test set included 100 faba bean inbred lines of diverse origins, which were grown in 2019 and 2020 in Groß Lüsewitz (north-eastern Germany). The location is characterized by an annual mean temperature of 8.3°C. Soil type is a slightly loamy sand with a pH of 5.7.

The worldwide diversity panel includes genetic stocks aiming at gathering a wide range of genetic diversity from diverse geographical origins: Africa (8 accessions), North and South America (2), Asia (27) and Europe (39). The remaining 24 accessions were provided by the International Center for Agricultural Research in the Dry Areas (ICARDA), but their exact origin is unknown. Europe, with 9 countries, is the most represented geographical area, followed by Asia, Africa and America (7, 4 and 2 countries, respectively). Spain accounts for the highest number of accessions (23). The diversity panel was built in collaboration with the public institutes ICARDA, the Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA) and the French National Institute for Agricultural Research (INRA). Prior to the genotyping analysis, all the Spanish lines had been self-pollinated for at least four generations. The remaining accessions were purified for two generations by single seed descent (SSD) in insect-proof cages. A detailed description of this collection is provided in Supplementary Table S1.

The experimental set up was a randomized complete-block design with 4 replications in both treatments. Drought stress treatment was performed in two mobile rainout shelters, with two blocks in each. The same design was applied in the open field control treatment. Plants were grown in single row plots of 1.2 m length with 14 plants each (one seed per hill) and the distance between plots was 0.5 m. Additional water was applied by drip irrigation (Supplementary Figure S1). Irrigation has been scheduled in the range of 60 - 70% of field capacity of the soil, determined over winter after excessive rainfall. Water content in the soil was assessed by time-domain reflectance (TDR) probes in about 0.40 m depth (Figure 1). Irrigation in the stress treatment was stopped when about 30% of the genotypes had started flowering. During rainfall the shelter covered the plots. An end-terminated drought stress was applied. The key data of the experiments are explained in Supplementary Table S2.

Weather conditions

The mean temperature, the Crop Heat Unit (CHU) based heat sums and the precipitation conditions during the vegetative period in the two years of the experiment are summarized in Table 1. The experimental year 2019 was moderate with respect to rainfall but warm with an early heat period in June. The year 2020 was cooler with slightly more rainfall and a late heat period in August, when faba beans were partly harvested already. This resulted in a more intensive drought stress in 2019 compared to 2020.

Measurements of traits

Four physiological traits were assessed in the youngest fully-developed leaves from the main stem of the inner 10 plants of each plot. Two leaf samples mixed from 5 plants per plot each were taken about 4 weeks after onset of stress, according to the stress development, for determining the content of free proline (PRO) and total soluble sugars (TSS). Samples were immediately frozen in liquid nitrogen. PRO was analyzed as described by Bates et al. (1973) and TSS was estimated by the anthrone reagent method according to Yemm and Willis (1954). SPAD measurements as an indirect parameter for chlorophyll content were carried out with a Chlorophyll Meter SPAD 502 Plus (Konica Minolta) at the beginning of stress treatment (SPAD1) and in the same time slot of TSS and PRO sampling (SPAD2). The mean value of 30 clippings per plot was used. The difference between these SPAD values, DiffSPAD (DiffSPAD = SPAD1 – SPAD2), was calculated to standardize the possible differences in chlorophyll content among genotypes and to obtain a more accurate measure of chlorophyll degradation (senescence). Two phenological traits were analyzed: end of flowering (EF), recorded as the days after sowing when no open flowers were visible any more in the plot (BBCH 69); and maturity date (MAT), defined as the date when more than 90% of the pods have ripened (BBCH 89) following the extended BBCH Code (Hack et al., 1992). Moreover, five morphological traits were recorded: plant height in cm (PH); number of pods per plant (PP); number of seeds per plant (SP); hundred seed weight (HSW) in grams; and plot yield (PY) in kg. Plant height was measured close to maturity (BBCH 80-85) from the ground to the tip of 10 inner plants per plot. At the end of the vegetative period, 10 single plants per plot were harvested by hand and pods and seeds were counted. Dry matter content was determined at 105°C. PY and HSW are given on a basis of 86% dry matter.

Data analyses

Statistical analyses were performed by SAS 9.4 (SAS Institute, USA). Analysis of variance (ANOVA) used a linear mixed model with year as a random effect. Means are shown as LSmeans. Minimum variance quadratic unbiased estimators (MIVQUEs) of variance components were used to calculate broad sense heritability h^2 according to Becker (2011). The response of the accessions to drought stress was calculated as relative yield of the drought stress treatment:

$$Yrel = PY_DS \times 100/PY_C$$

with PY_DS as plot yield of the drought stress and PY_C as plot yield of control treatment.

The Stress Tolerance Index (STI) was calculated for the yield measured in both treatments (control and drought) according to Fernandez (1992) as follows:

$$STI = PY_C \times PY_DS/(\bar{Y}_C)^2$$

with PY _ C as plot yield of control, PY_DS as plot yield of the drought stress treatment and $\bar{Y}_C$as mean yield of the control environment.

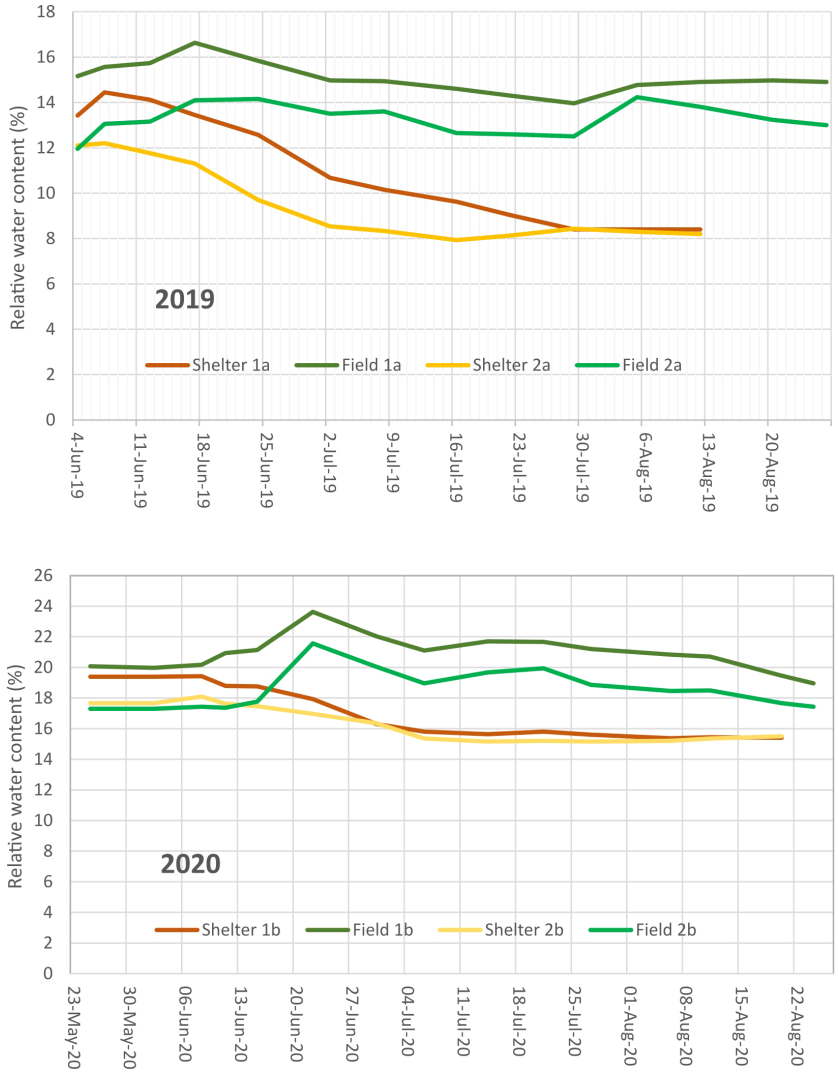


FIGURE 1
Soil water content of the experimental areas measured by TDR probes in about 0.40 m depth for the first and second experimental year.

Pearson's correlations between STI, PY_C and PY_DS data were calculated in order to check for possible dependencies between the performance of the accessions under control conditions and the strength of response to drought stress. A principal component analysis (PCA) was calculated based on correlations using JMP Genomics 9 (SAS Institute, USA). The results were represented using a biplot that combines both the principal component scores and the loading vectors in a single plot. In addition, a cluster analysis

TABLE 1 Weather conditions during the vegetative period in 2019 and 2020 at Gross Lüsewitz.

	Mean Temperature (°C)		CHU* heat sums (°C)		Precipitation (mm)	
	2019	2020	2019	2020	2019	2020
April	9.2	8.4	259	188	23.9	10.9
May	10.9	10.9	311	279	33.1	20.5
June	19.1	16.6	665	579	87.4	131.2
July	17.7	16.0	644	571	79.7	62.1
August	18.5	19.4	685	708	55.7	82.2
Sum			2564	2325	279.8	306.9

*Crop Heat Units.

of observations was performed to find subgroups within phenotypic data in control and drought conditions by means of the K-means clustering method (Hartigan and Wong, 1979). The data were scaled to make variables comparable. The average silhouette approach was used to determine the optimal number of clusters. K-means clustering was performed using the built-in k-means libraries of R suite and similarities were measured by squared Euclidean distance to classify all the accessions into groups (R Core Team, 2022). The results were visualised by means of the 'fviz_cluster' function in the 'factoextra' R-package (Kassambara and Mundt, 2020).

Results

Statistics of the data

Descriptive statistics and analysis of variance (ANOVA) of the traits recorded over two years is shown in Table 2. The table includes the mean, range and Least Significant Difference (LSD) values together with the heritability (h^2) of the traits in both conditions. LSDs of all the traits studied in the 100 faba bean genotypes in control and drought stress conditions are provided in Supplementary Table S3. Most of the parameters displayed a wide range of differences among the treatments. A significant effect of genotype and drought stress treatment as well as an interaction between them could be observed for all traits.

The results obtained for the physiological traits showed clear differences between treatments. TSS and PRO were measured when drought stress was fully developed and senescence started in the stress treatment. As expected, both parameters increased significantly (24% and 269% respectively) under drought stress (Figure 2). The significant interaction between genotype x treatment indicates a different response of genotypes to drought stress (Table 2). The leaf chlorophyll concentration, assessed as SPAD values, showed a marked decrease (44%) after 4 weeks of drought stress compared to the control (SPAD2 and DiffSPAD, Figure 3), which was clearly visible (Figure 4). This was confirmed by the difference between the two SPAD values (DiffSPAD = SPAD1 – SPAD2) which minimizes the effects of variability in chlorophyll content among genotypes and stands for the degree of senescence. Whereas in the control treatment values were increasing until full flowering (negative difference), chlorophyll content was already decreasing in the stress treatment (positive difference).

For all the phenological and morphological traits, the mean values also decreased under drought stress conditions with values ranging from 9.7% for MAT to 11.3% for EF (Table 2). PH decreased by 12.5% and the taller forms were more affected than the shorter types (data not shown). In contrast, strict declines were observed for the remaining yield related traits PP, SP and PY. Mean PY value under drought stress was significantly reduced by 46.4%, while PP and SP were reduced by 40.8% and 42.8%, respectively. Similarly, HSW values were lower, compared to the control, but the mean decrease of 9.7% was not as high as that of the other yield components (Figure 5). A significant interaction between genotypes and drought stress treatment was detected.

Heritability computed across the two water regimes (control vs. drought) ranged from 0.3 (PRO) to 0.91 (HSW), both in stress and non-stress conditions (Table 2). Heritability was similarly high (> 0.82) under both conditions for the yield related traits PH, PP, SP and HSW. Moderate to high heritability was recorded for the physiological traits TSS (0.56-0.66), MAT (0.75-0.68), SPAD2 (0.73-0.89), DiffSPAD (0.38-0.86) and PY (0.83-0.65) in both conditions. The most contrasting heritability values between water regimes were found in EF (0.73-0.44) and PRO (0.63-0.3).

Yield relationships and genotypes

For a better understanding of the relationship between yield under control and stress conditions, a regression analysis was conducted. This relationship provides a benchmark for identifying the most drought tolerant genotypes (Figure 6). A positive correlation was detected between the two parameters (r = 0.75), whereby the high yielding genotypes were more affected by drought stress than the low yielding ones.The non-linear regression line shows that stress yield may not exceed a distinct level predefined by the stress environment. A correlation coefficient of -0.72 (Figure 7) revealed the negative and relatively close correlation between the yield under controlled conditions (standing for yield potential) and the relative yield (standing for yield stability). The figure clearly depicts that selection of genotypes with a high relative yield in most cases leads to rather low yielding genotypes. For this reason, a stress tolerance index (STI) based on seed yield (PY) was calculated for each genotype in response to severe drought and well-watered conditions, in order to identify genotypes producing high yields under both conditions. The variability of this index is shown in Figure 8. By using STI as target parameter, it was possible to identify genotypes that combine a high yield potential with a reasonable stress yield in this faba bean assortment. Accession ID and origin of the eight most drought-tolerant genotypes (5, 26, 35, 49, 56, 64, 75 and 85), according to STI are given in Table 3.

The K-means cluster analysis using the morphological, phenological and physiological faba bean data and the STI index in control and drought conditions classified the accessions into groups based on their dissimilarities (Figure 9). For control conditions, two clusters were identified: cluster 1 (15 genotypes) showed highest STI value and highest average values for all traits (except for HSW), containing six of the eight best performing STI genotypes (5, 35, 56, 64, 75 and 85), while cluster 2 (84 genotypes) had the maximum average value for HSW and contained the remaining genotypes 26 and 49 (Table 4). Under drought conditions, the genotypes were classified in three clusters: cluster 1 (56 genotypes), cluster 2 (13 genotypes) and cluster 3 (30 genotypes). Cluster 2 showed the best average value for PRO, SPAD2 (and correspondingly low DiffSPAD), MAT, PH, PP, SP, PY and STI, while cluster 1 had a similar yield but an intermediate value for STI and cluster 3 had the lower yield and STI mean values but the highest ones for TSS, EF and HWS (Table 4). Again, six of the best STI genotypes (accessions 5, 35, 49, 56, 64 and 85) were included in cluster 2 while cluster 1 contained the other two accessions (26 and 75). According to cluster analysis, genotypes from cluster 2 were superior regarding the STI index and yield

TABLE 2 Descriptive statistics and analysis of variance (ANOVA) of the traits over both years.

Trait		Control	Stress	Unit	F values				Heritability (h^2)	
					Genotype	Treatment	G x T	Year	Control	Stress
Soluble sugars (TSS)	Range LSD	1117.6 769.5 - 1709.4 587.1	1387.9 998.2 - 2242.4 659.9	µmol/ g DM	7.44 ***	339.79 ***	1.86 ***	196.73 *** 2019>2020	0.56	0.66
Free proline (PRO)	Range LSD	2.29 1.50 - 6.55 1.20	8.45 1.79 - 57.02 26.75	µmol/ g DM	2.93 ***	171.09 ***	2.57 ***	128.78 *** 2019>2020 stress only	0.63	0.3
SPAD 2	Range LSD	42.8 29.2 - 57.4 8.5	24.0 11.5 - 51.3 9.8	SPAD Units	22.57 ***	6780.40 ***	9.30 ***	208.32 *** 2020>2019	0.73	0.89
Difference SPAD (DiffSPAD)	=Senescence Range LSD	-5.0 -15.1 - 5.2 10.1	15.5 -6.8 - 25.2 11.0	SPAD Units	11..56 ***	6100.08 ***	8.30 ***	54.74 *** 2019>2020	0.38	0.86
End of flowering (EF)	Range LSD	97,0 85.5-118.6 15.7	86.3 82.0-100.3 11.2	DAS	8.13 ***	1169.23***	2.33***	230.27*** 2019>2020	0.73	0.44
Maturity (MAT)	Range LSD	127.4 118.0-143.0 10.1	115.1 105.9 - 126.5 9.2	DAS	9.28 ***	2404.57 ***	1.51 **	54.41 *** 2019>2020 C 2020>2019 S	0.75	0.68
Plant height (PH)	Range LSD	64.6 42.1 - 102.3 18.5	56.5 37.1 - 71.6 12.2	cm	17.19 ***	359.77 ***	3.19 ***	170.93 *** 2020>2019 control only	0.87	0.82
Pods per plant (PP)	Range LSD	10.8 4.5-27.1 6.9	6.4 3.5 - 12.0 3.2	number	19.11 ***	1124.49***	5.22***	1.59*** 2020>2019 control only	0.89	0.88
Seeds per plant (SP)	Range LSD	24.9 8.9 - 77.9 17.3	14.2 5.8 - 27.0 7.5	number	21.3 ***	1039.57 ***	6.86 ***	1.60 *** 2020>2019	0.9	0.85
Hundred seed weight (HSW)	Range LSD	66.93 30.27 - 115.38 23.67	60.44 26.07 - 103.24 16.28	g/100 seeds	28.6 ***	133.13 ***	1.39 **	420.52 *** 2019>2020	0.83	0.91
Plot yield (PY)	Range LSD	0.155 0.047 - 0.321 0.119	0.083 0.031 - 0.117 0.054	kg/plot	9.19 ***	1003.33 ***	3.26 ***	5.0 * 2019>2020	0.83	0.65
Stress tolerance index (STI)	Range LSD	0.564 0.074 - 1.378 0.048			4.94 ***			1.38 n.s.	0.80	

***, ** and * are significant for $\alpha \leq 0.1$, 1 and 5%, respectively, n.s., not significant; LSD, Least Significant Difference; DAS, Days after sowing.

related traits, so cultivation of these genotypes would be recommended under drought stress conditions.

Correlations and principal component analysis (PCA)

A correlation analysis was performed to evaluate the relationships among morphological, phenological and physiological traits as well as their association with PY and STI. As mentioned above, the correlations of plot yield under control (PY_C) and drought stress conditions (PY_DS) with STI are similar and highly significant, with PY_C having the greater impact on STI (Table 5). The most yield influencing factors were SP and PP, although physiological traits such as chlorophyll content (SPAD values), PRO and TSS together with phenological traits such as EF and MAT, also contributed significantly to the respective yield and to the STI values.

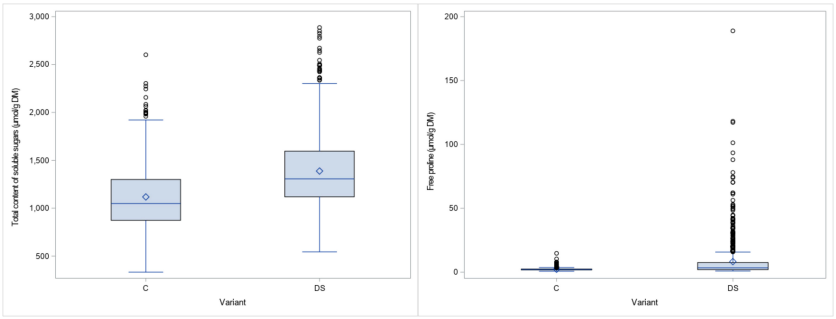


FIGURE 2
Boxplots for total content of soluble sugars (TSS) and free proline (PRO) for control (C) and drought stress (DS) treatment.

Under control conditions, PH was significantly correlated with most traits, particularly PP and SP, highlighting their contribution to yield. Strong positive correlations were also observed for PP and SP with STI, PY and PRO. Interestingly, HSW showed a strong negative correlation with PP and SP. Under drought stress conditions, DiffSPAD showed a significant negative correlation with all traits except HSW. SPAD2 showed strong positive correlation with MAT and PRO and lower but still significant correlation with PH, PP and SP, which showed positive inter-correlations. In control conditions, all measured yield-related traits (PY, PP, and SP) were associated positively except HSW, indicating that they could simultaneously improve drought tolerance at the expense of seed weight. Correlations were stronger in the control treatment in nearly all cases with the exception of EF and TSS (Table 5).

In order to determine the most contributing traits to yield under both conditions, a principal component analysis (PCA) was performed (Figure 10). Results showed that the first and second component (PC1 and PC2) represented 63.9% of the variance observed in control conditions and 59.3% of the variance in the drought stress treatment. Biplots for control and drought conditions were constructed from PC1 (42.6% and 36% variation explained, respectively) and PC2 (21.3% and 23.3% variation explained, respectively) with the distribution of the 100 genotypes, the morphological, phenological and physiological traits and STI. The PCA matrix revealed that the genotypes assayed followed different strategies with respect to seed set, seed size or physiological parameters (Figure 10). In the control condition, the PC1 axis was primarily associated with SP and PP, followed by PH and PRO, while TSS and differences in chlorophyll content (DiffSPAD) were the most contributing factors for the dispersion of the genotypes along PC2. In the drought stress treatment, chlorophyll content (SPAD2), followed by SP and PP had the largest contribution to PC1, while EF and MAT mainly contributed to PC2. The nearly 180° angle between SPAD2 and DiffSPAD indicated a strong negative association, as expected. PCA biplot analysis further demonstrates that the eight superior genotypes according to STI follow different strategies with respect to phenological, physiological and yield components. Thus, in drought conditions the genotypes 26, 35, 64 and 75 showed a higher number of pods (PP_DS), seeds per plant (SP_DS) and yield (PY_DS) as well as a higher STI value, but low total soluble sugar content (TSS_DS). On the other hand, genotypes 5, 49, 56 and 85 showed higher values for plant height (PH_DS) and the physiological traits SPAD2 and PRO, together with a low hundred seed weight (HSW_DS).

Discussion

In this study, a faba bean collection of 100 accessions was used to assess the relevance of morphological, phenological and physiological traits with respect to yield under different water regimes. One determining factor is the chlorophyll content, measured indirectly by SPAD metre. A decrease in chlorophyll

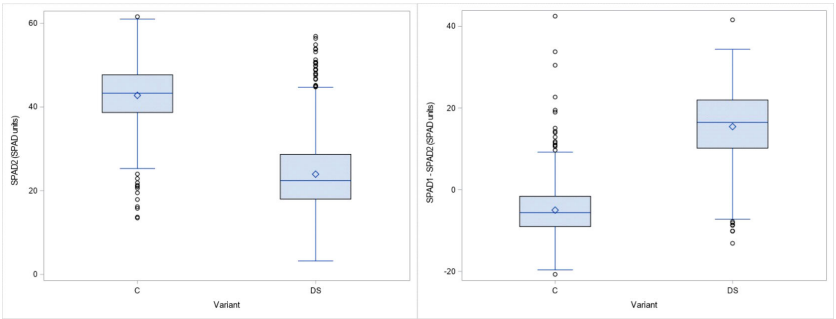


FIGURE 3
Chlorophyll content by SPAD Meter about 4 weeks after onset of drought stress (SPAD 2) and difference to SPAD 1 values (DiffSPAD) for control (C) and drought stress (DS) treatment.



FIGURE 4
Visible signs of early senescence under drought stress in rain out-shelter (left side) compared to the control (right side) at the beginning of July 2019.

under drought stress has been widely reported (Ammar et al., 2015; Siddiqui et al., 2015; Abid et al., 2017). Drought-induced early senescence leading to early chlorophyll degradation was also observed in the present study with genotypes responding differently to the stress factor. Late chlorophyll degradation in connection with prolonged photosynthetic activity is a desirable trait, especially under drought stress,

Accumulation of free proline and of carbohydrates as free sugars contributes to osmotic adjustment in plants (Yoshiba et al., 1997; Hekneby et al., 2006). An increase in free proline between 42% and 202%, under 30% of field capacity was reported in a faba bean greenhouse trial (Abid et al., 2017), These results evidenced that accumulation of proline has a role in the faba bean tolerance to drought. Our study further supports this idea, as the amount of proline was four times higher under drought stress than in controlled conditions. Siddiqui et al. (2015) reported that drought-tolerant faba bean genotypes accumulated more proline than sensitive genotypes. In contrast, Migdadi et al. (2016) found a negative correlation between proline and yield over all treatments, which was mainly due to the high accumulation of proline under drought stress conditions rather than to differences between genotypes within the treatment. This is in accordance with our study, in which there was no close correlation between proline accumulation and other traits scored in the drought stress treatment. The significant correlation in the control treatment (PY_C) could be due to limitations in the measurements, as the content of free proline and therefore the genotypic differences are very low in the absence of osmotic stress (Balko, 2005; Stoddard et al., 2006; Kabbadj et al., 2017)

Another adaptation to water stress is the accumulation of carbohydrates, one of the main components of osmoregulation in many plant species (Morgan, 1984). In the forage legume *Lotus japonicus* L., accumulation of sugars such as fructose, galactose, glucose and maltose occurred when affected by drought (Sanchez et al., 2012). In soybean leaves, about 30% of the metabolites identified were soluble sugars and sugar alcohols, which are important for plant adaptation to stresses, especially during water deficit (Benkeblia et al., 2007). These findings are supported by the present study where a 24% increase of TSS under drought conditions was observed.

Plant height in faba bean is strongly determined by the growth type. Indeterminate types reach a greater height under favorable conditions while determinate (topless) types are lower. In this study, the indeterminate types showed a broad variability in plant height. Drought stress reduced plant height and this effect was more pronounced in the tall forms. Importantly, the yield-deciding parameter is not plant height itself but the number of reproductive nodes. Drought during the reproductive phase may reduce plant height and therefore the number of reproductive nodes (Gnanasambandam et al., 2012). In our field trials, plant height in both treatments was most significantly correlated to the number of pods per plant as well as to seeds per plant.

Pods per plant, seeds per pod and seed weight are among the main yield components (Ayaz et al., 2004; Tadesse et al., 2011; Ouji et al., 2017). Under drought stress, the most susceptible developmental stage is the reproductive phase including flowering, early podding and pod setting (Muktadir et al., 2020). The early podding was the most sensitive stage reported by Mwanamwenge et al. (1999), causing a reduction in the number of pods and seeds due to high rates of abscission (Sekara et al., 2001). In our study, seeds per plant showed the closest relation to

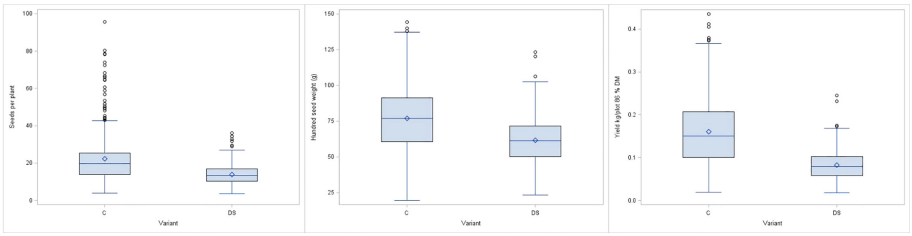


FIGURE 5
Boxplots for seeds per plant (SP), hundred seed weight (HSW) and plot yield (PY) for control (C) and drought stress (DS) treatment.

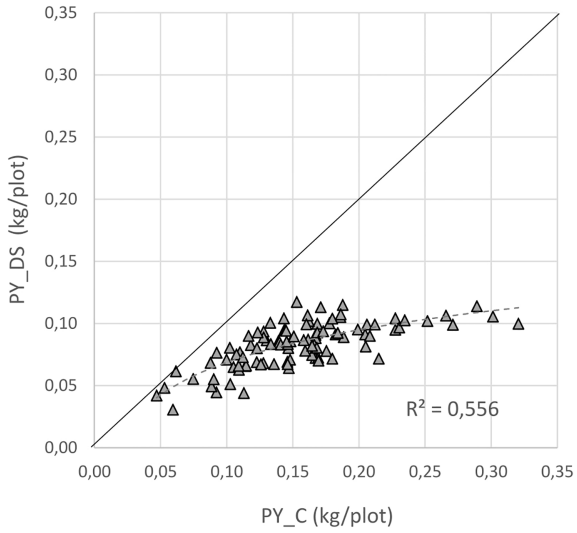


FIGURE 6
Relation between yield under drought stress (PY_DS) and yield in the control (PY_C); the solid black line stands for 100% yield stability.

seed yield in both the control and the drought stress treatment, followed by pods per plant. Similar outcomes were reported by Pilbeam et al. (1992) and Tofiq et al. (2016). Lopez et al. (1996) reported that the number of pods per plant was most affected by drought stress during flowering, with yield reductions up to 70%.

Yield relations and genotypes

Previous studies pointed out that high yield potential under optimal conditions does not necessarily correlate with yield under limited water irrigation (Ramirez-Vallejo and Kelly, 1998; Abdelmula et al., 1999; El-Hendawy et al., 2017). This is in agreement with the results of our study where the performance of the faba bean lines was not consistent across the two treatments. The negative correlation between yield potential (yield under non-stress) and yield stability (yield under severe stress) suggests that the use of yield stability as a target parameter for drought tolerance could lead to the selection of low yielding genotypes. Nevertheless, varieties with high yield potential will often have an advantage over varieties with lower yield potential under moderate drought stress. Whether genotypes with medium yield potential and high stress yield might have an advantage for regions with regularly occurring drought stress remains a matter of debate. For regions with occasional drought stress, the relatively close correlation between stress yield and control yield (Figure 6) implies that both yield under well watered and drought stress should be considered, as shown by the STI value (Fernandez, 1992). Several studies point towards STI as the most suitable index for selecting the best yielding genotypes under contrasting conditions (Fernandez, 1992; El-Hendawy et al., 2017; Memari et al., 2022). According to our results, STI exhibits a strong correlation with yield (PY) and yield

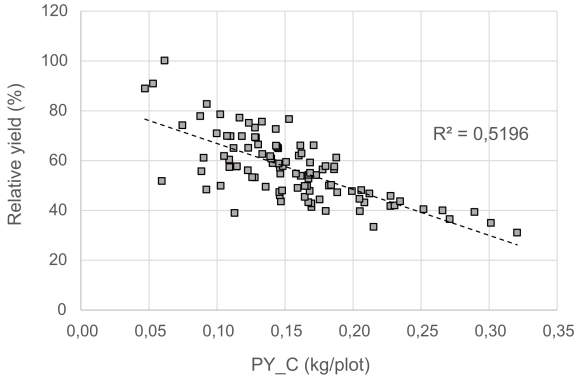


FIGURE 7
Relationship between relative yield and yield in the control (PY_C).

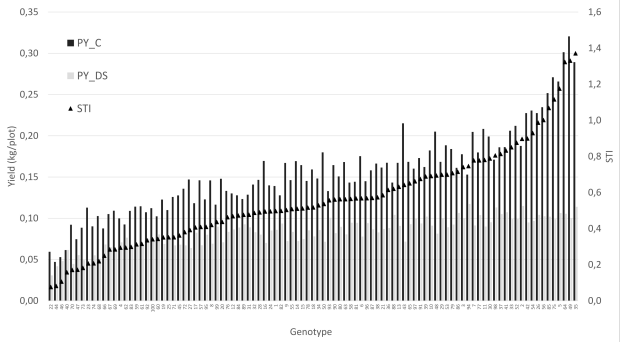


FIGURE 8
Genotypes of the test set ordered according to their STI value. PY_C: control yield. PY_DS: drought stress yield.

components (PP and SP) both under control conditions and drought stress, and appears to be an effective index for the selection of genotypes with high yield potential under both environmental conditions.

Correlations, cluster analysis and principal component analysis

Genotypes with higher yields in favorable as well as stress environments are the ones preferably selected in breeding programs. Nevertheless, the presence of genotype by environment (GxE) interactions is a major concern, since it may reduce correlation between genotypic and phenotypic values and slow down the selection progress (Romagosa et al., 2009).

Our results revealed a high GxE interaction for all the traits studied. Nevertheless, a clear association between the yield related traits, PP and SP and the physiological trait SPAD2 was observed through their significant correlation in both control and drought stress conditions. A moderate correlation was also found between PRO and SP in both environments. The high heritability estimates and the positive correlations of SPAD2 with PP and SP in stress conditions indicates that SPAD2 may be a suitable selection criterion for drought tolerance in faba bean.

A positive correlation between maturity and yield was found in the control condition, but not in the stress treatment. Katsoulieri et al. (2020) described that later ripening genotypes showed greater susceptibility to drought stress during early pod set compared to genotypes with early maturity. This was supported by Khan et al. (2010) as well as Manning et al. (2020), where early flowering is considered an important adaptation trait for cool-season legumes growing under semi-arid conditions.

PCA is a powerful strategy to uncover genetic factors that contribute to complex traits. It has been considered a useful complementary tool in drought tolerance screenings and in the selection of the most tolerant genotypes (Mohamamdi et al., 2017; Santos et al., 2020; Reyes et al., 2022; Weng et al., 2022). This approach has been successfully adopted in drought stress tolerance evaluations in different crop legumes such as soybean (Bouslama and Schapaugh, 1984; Toum et al., 2022), lentil (Siahsar et al., 2010), chickpea (Shah et al., 2020; Arif et al., 2021) and faba bean (Mansour et al., 2021; Afzal et al., 2022). Here we applied the same strategy to integrate multiple parameters for grouping the faba bean accessions according to their drought tolerance. Our results show that in controlled conditions, the morphological traits SP, PP and PH were the most contributing factors for the dispersion of the genotypes along the PC1 axis, while PC2 was positively loaded with the TSS in the leaf, which is often increased during grain filling and may point towards late senescence.

Under drought stress conditions, there was a shift towards traits related to a longer photosynthesis period in PC1 (as late chlorophyll content or SPAD2), resulting in a higher number of seeds per plant. On the other hand, the phenological traits EF and MAT, which are strongly related to the length of the vegetation period, were the main factors contributing to PC2. These findings were supported by investigations in *Miscanthus*, where a PCA performed with relative values and the physiological traits chlorophyll content and chlorophyll fluorescence revealed their important role in drought stress response (Weng et al., 2022). The relevance of both physiological traits for final yield under drought stress was also reported in durum wheat (Mohamamdi et al., 2017; Reyes et al., 2022), further evidencing that a delay of chlorophyll degradation during senescence can improve performance under drought conditions. These observations reinforce the idea that screening for chlorophyll content under drought conditions could significantly

TABLE 3 STI value, ID and origin of the eight best genotypes.

STI value	Genotype	Accession ID	Accession	Origin
1.371	35	EUC_VF_077	1272-1	RUS
1.330	49	EUC_VF_130	INRA 2394	DEU
1.323	64	EUC_VF_194	INRA 612	EGY
1.176	5	EUC_VF_ 009	359-1	TUR
1.116	75	EUC_VF_ 303	IG 72242	CHN
1.068	85	EUC_VF_336	Misr 3	EGY
1.001	56	EUC_VF_174	1248-3	IRN
0.987	26	EUC_VF_064	1157-1	ESP

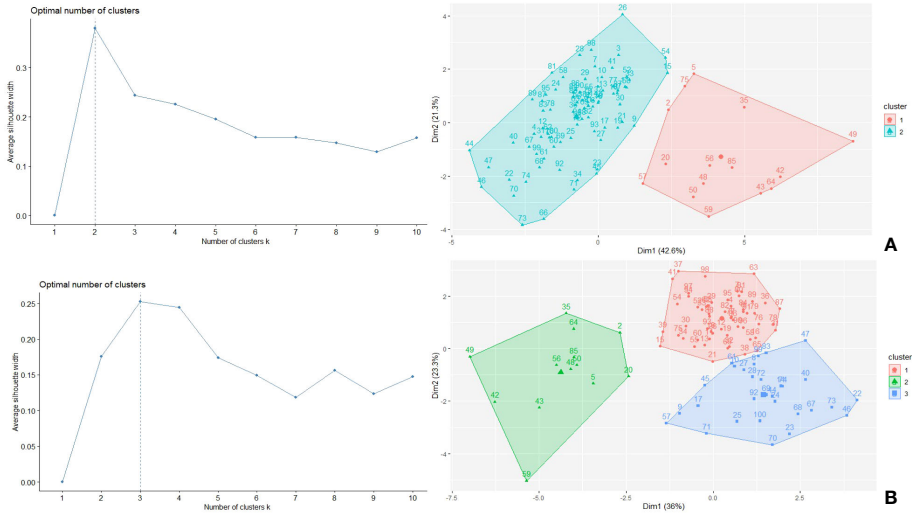


FIGURE 9
Optimum number of clusters and clustering of the 100 faba bean accessions under control **(A)** and drought **(B)** conditions.

TABLE 4 Average value based on physiological, phenological and morphological and the STI index under control (C) and drought stress conditions (DS) after K-means cluster analysis.

Cluster	TSS_C	PRO_C	SPAD2_C	DiffSPAD_C	EF_C	MAT_C	PH_C	PP_C	SP_C	HSW_C	PY_C	STI
1 (15)*	1212.53	2.95	44.66	-5.51	102.27	132.62	84.89	17.85	44.33	51.04	0.21	0.82
2 (84)	1099.55	2.13	42.45	-4.90	96.01	126.39	60.48	9.46	21.28	69.97	0.14	0.51
Cluster	TSS_DS	PRO_DS	SPAD2_DS	DiffSPAD_DS	EF_DS	MAT_DS	PH_DS	PP_DS	SP_DS	HSW_DS	PY_DS	STI
1 (56)	1276.416	6.671	21.529	18.183	84.221	112.875	54.782	6.405	14.869	62.408	0.091	0.612
2 (13)	1310.345	17.398	37.205	2.385	88.485	119.037	65.862	9.148	20.596	45.726	0.092	0.899
3 (30)	1634.117	6.995	22.449	16.267	88.961	117.319	55.341	5.157	10.324	63.628	0.065	0.321

* Number of genotypes in each cluster.

TABLE 5 Correlation matrix with Pearson's correlation coefficient r for all relevant traits under control (above) and drought stress conditions (down).

Trait	TSS _C	PRO_C	SPAD2_C	DiffSPAD_C	EF_C	MAT_C	PH_C	PP_C	SP_C	HSW_C	PY_C	STI
TSS_C	1	-0.097	**-0.290**	**0.482**	**0.385**	**0.276**	**0.284**	0.118	0.026	-0.219	-0.210	**-0.256**
PRO_C		1	**0.260**	-0.164	0.185	**0.357**	**0.560**	**0.562**	**0.616**	**-0.358**	**0.421**	**0.352**
SPAD2_C			1	**-0.812**	0.035	**0.372**	**0.263**	**0.310**	**0.345**	0.129	**0.509**	**0.474**
DiffSPAD_C				1	-0.004	**-0.259**	-0.163	**-0.261**	**-0.267**	-0.200	**-0.474**	**-0.448**
EF_C					1	**0.694**	**0.408**	**0.255**	0.132	-0.106	-0.017	-0.105
MAT_C						1	**0.619**	**0.444**	**0.377**	-0.007	**0.374**	**0.275**
PH_C							1	**0.783**	**0.767**	**-0.340**	**0.590**	**0.484**
PP_C								1	**0.925**	**-0.515**	**0.663**	**0.566**
SP_C									1	**-0.499**	**0.737**	**0.650**
HSW_C										1	0.153	0.201
PY_C											1	**0.961**
Trait	**TSS _DS**	**PRO_DS**	**SPAD2_DS**	**DiffSPAD_DS**	**EF_DS**	**MAT_DS**	**PH_DS**	**PP_DS**	**SP_DS**	**HSW_DS**	**PY_DS**	**STI**
TSS_DS	1	0.217	0.197	0.187	**0.267**	**0.333**	0.035	**-0.293**	**-0.414**	0.123	**-0.441**	**-0.410**

(Continued)

TABLE 5 Continued

Trait	TSS _C	PRO_C	SPAD2_C	DiffSPAD_C	EF_C	MAT_C	PH_C	PP_C	SP_C	HSW_C	PY_C	STI
PRO_DS		1	**0.510**	**-0.459**	0.100	**0.343**	**0.279**	0.171	**0.254**	-0.052	0.187	**0.369**
SPAD2_DS			1	**-0.905**	**0.282**	**0.558**	**0.442**	**0.423**	**0.478**	**-0.325**	0.137	**0.373**
DiffSPAD_DS				1	**-0.269**	**-0.504**	**-0.392**	**-0.420**	**-0.432**	**0.327**	-0.068	**-0.317**
EF_DS					1	**0.680**	**0.253**	-0.014	-0.172	-0.029	**-0.405**	**-0.250**
MAT_DS						1	**0.393**	0.076	-0.090	0.032	-0.182	0.023
PH_DS							1	**0.500**	**0.398**	-0.195	0.206	**0.381**
PP_DS								1	**0.819**	**-0.626**	**0.332**	**0.429**
SP_DS									1	**-0.651**	**0.499**	**0.523**
HSW_DS										1	**0.268**	0.129
PY_DS											1	**0.836**

Numbers in bold are significant for $\alpha \leq 5\%$. PRO, content of free proline; TSS, total content of soluble sugars; EF, end of flowering; MAT, maturity; PH, plant height; PP, pods per plant; SP, seeds per plant; HSW, hundred seeds weight; PY, plot yield; STI, stress tolerance index; C, control; DS, drought stress.

contribute to efficient selection in breeding programs. The PCA was also helpful to group the faba bean genotypes according to their phenotypic response (Abou-Zaitoun et al., 2018; Or-Rashid et al., 2021). It further evidences that the lines tested follow different strategies to reach high yields under drought stress.

Cluster analysis is a multivariate analysis that categorizes the genotypes into several subgroups according to their similarities, based on the capability and performance of the traits examined. This approach has been used to predict the best-performing genotypes under drought stress in other crops (Hannok et al., 2021; Mutlu et al., 2022; Aslam et al., 2023). Grouping breeding materials using multivariate data under a given condition allows plant breeders to select breeding lines more efficiently. Here we applied the same strategy to integrate phenological, physiological as well as yield

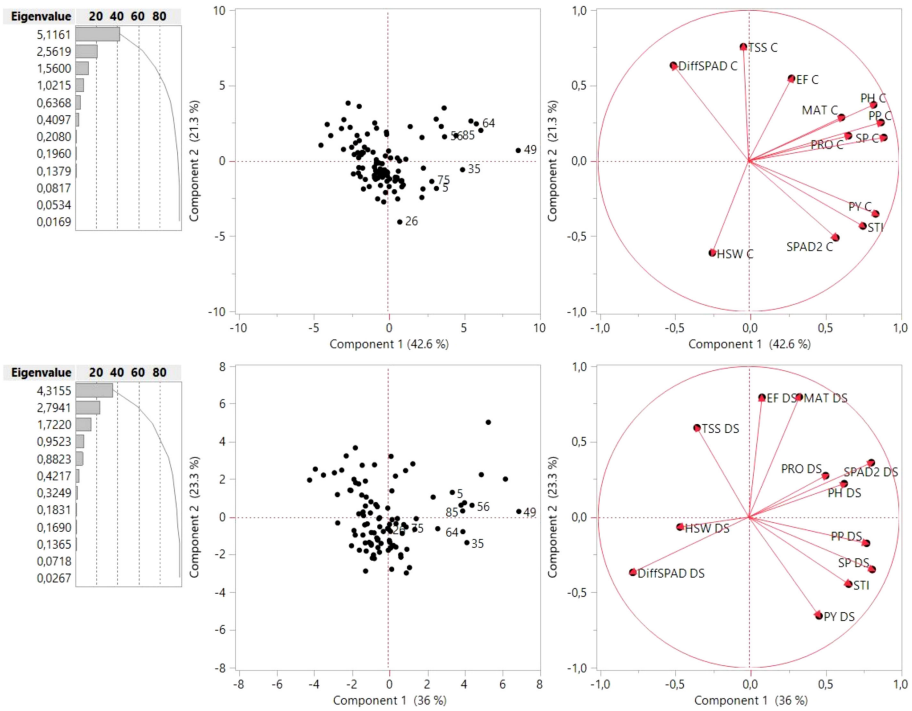


FIGURE 10
PCA based on all Least Squares (LS) means of the 100 genotypes for the respective traits assessed under control (above) and drought stress conditions (down). Points with numbers are the eight best genotypes according to STI (see). The lines indicate eigenvectors representing the strength (length of the vector) and the direction of the parameter correlation relative to the first two principal componente (PC1 and PC2). TSS, total content of soluble sugars; PRO, free proline content; SPAD, measures of chlorophyll content 4 weeks after onset of stress (SPAD 2) and differences in chlorophyll content (DiffSPAD); EF, end of flowering; MAT, maturity; PH, plant height; PP, pods per plant; SP, seeds per plant; HSW, hundred seed weight; PY, plot yield; STI, stress tolerance index; C, control, DS, drought stress.

components for grouping the faba bean accessions according to their drought tolerance and the clustering method was effective to differentiate the best performing genotypes under drought stress.

PCA biplot and cluster analysis demonstrate that the eight superior genotypes according to STI follow different strategies with respect to phenological, physiological and yield components. Thus, in drought conditions the genotypes of cluster 2, that included six of the best performing STI genotypes (5, 35, 64, 56, 64 and 85), were the best group in terms of performance. These genotypes are characterized not only by the highest yield (PY_DS) and STI values but also by the highest mean values for proline (PRO_DS), chlorophyll content (SPAD2_DS), plant height (PH_DS), number of pods (PP_DS), seeds per plant (SP_DS) and later maturity days (MAT_DS), thus being the most promising traits for selection. On the other hand, these genotypes displayed lower total soluble sugar content (TSS_DS) and hundred seed weight values (HSW).

In summary, our results revealed that the Stress Tolerance Index (STI) is a useful criterion for selecting drought tolerant and high yielding faba bean genotypes. Using STI and biplot analysis, eight accessions with relatively high yield under both normal and drought stress conditions were identified and six of them further validated by K-means cluster analyses. Thus, genotypes 5, 35, 64, 56, 64 and 85 could potentially be used as genetic resources in faba bean breeding programs to develop varieties with enhanced drought resistance traits. Our study also suggests that SPAD2, followed by SP and PP can be reliable indicators for selection. To bridge the gap between traditional and molecular breeding, this faba bean collection has recently been genotyped with a high-density SNP genotyping array to conduct a genome-wide association (GWAS) study for drought resistance (Gutierrez et al., 2023). After validation, the candidate genes identified can be used for marker-assisted selection to accelerate and improve faba bean yield in agricultural areas where long water deficit periods are expected.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AMT and NG selected and multiplied the seeds for the experiments. CB planned and carried out the drought evaluations and processed the experimental data. CB and AT wrote the manuscript and NG contributed in the review and editing. CB and AMT acquired the financial support for the project. All authors contributed to the article and approved the submitted version.

Funding

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n° 727312 (EUCLEG Project) and PID2020-114952RR-I00, co-financed by ERDF.

Acknowledgments

Many thanks to the seed companies and institutes which provided us with faba bean seeds of their accessions: INRA, Dijon (Jean-Bernard Magnin-Robert), ICARDA (Fouad Maalouf), and IFAPA.

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.

The handling editor LS declared a past collaboration with the author AT.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1236147/full#supplementary-material

References

Abdelmula, A. A., Link, W., von Kittlitz, E., and Stelling, D. (1999). Heterosis and inheritance of drought tolerance in faba bean, *Vicia faba* L. *Plant Breed.* 118 (6), 485–490. doi: 10.1046/j.1439-0523.1999.00411.x

Abid, G., Hessini, K., Aouida, M., Baudoin, J. P., Aroua, I., Baudoin, J.-P., et al. (2017). Agro-physiological and biochemical responses of faba bean (*Vicia faba* L. var 'minor') genotypes to water deficit stress. *Biotechnol. Agron. Soc Environ.* 21 (2), 146–159. doi: 10.25518/1780-4507.13579

Abou-Zaitoun, S., Chandresakar, K., Assili, S., Shtaya, M., Jamous, R., Mallah, O., et al. (2018). Unlocking the genetic diversity within a middle-east panel of Durum wheat landraces for adaptation to semi-arid climate. *Agronomy* 8, 233. doi: 10.3390/agronomy8100233

Afzal, M., Alghamdi, S. S., Migadi, H. H., El-Harty, E., and Al-Faifi, S. A. (2022). Agronomical and physiological responses of faba bean genotypes to salt stress. *Agriculture* 12, 235. doi: 10.3390/agriculture12020235

Alderfasi, A. A., and Alghamdi, S. S. (2010). Integrated water supply with nutrient requirements on growth, photosynthesis productivity, chemical status and seed yield of faba bean. *Am.-Eurasian J. Agron.* 3, 8–17.

Alghamdi, S. S., Al-Schameri, A. M., Migdadi, H. M., Ammar, M. H., El-Harty, E. H., Khan, M. A., et al. (2015). Physiological and molecular characterization of faba bean (*Vicia faba* L.) genotypes for adaptation to drought stress. *J. Agron. Crop Sci.* 201, 401–409. doi: 10.1111/jac.12110

Ali, M. B. M., Welna, G. C., Sallam, A., Martsch, R., Balko, C., Gebser, B., et al. (2016). Association analyses to genetically improve drought and freezing tolerance of faba bean (*Vicia faba* L.). *Crop Sci.* 56, 1036–1048. doi: 10.2135/cropsci2015.08.0503

Amede, T., and Schubert, S. (2003). Mechanisms of drought resistance in grain legumes. I. Osmotic adjustment. *Ethiop J. Sci.* 26, 37–46.

Amede, T., von Kittlitz, E., and Schubert, S. (2001). Differential drought responses of faba bean (*Vicia faba* L.) inbred lines. *J. Agron. Crop Sci.* 183, 35–45. doi: 10.1046/j.1439-037x.1999.00310.x

Ammar, M. H., Anwar, F., El-Harty, E. H., Migdadi, H. M., Abdel-Khalik, S. M., Al-Faifi, S. A., et al. (2015). Physiological and Yield Responses of Faba bean (*Vicia faba* L.) to Drought Stress in Managed and Open Field Environments. *J. Agron. Crop Sci.* 201 (4), 280–287. doi: 10.1111/jac.12112

Annicchiarico, P., and Iannucci, A. (2008). Adaptation strategy, germplasm type and adaptive traits for field pea improvement in Italy based on variety responses across climatically contrasting environments. *Field Crops Res.* 108 (2), 133–142. doi: 10.1016/j.fcr.2008.04.004

Arif, A., Parveen, N., Waheed, M. Q., Atif, R. M., Waqar, I., and Shah, T. M. (2021). A comparative study for assessing the drought-tolerance of chickpea under varying natural growth environments. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.607869

Aslam, S., Hussain, S. B., Baber, M., Shaheen, S., Aslam, S., Waheed, R., et al. (2023). Estimation of drought tolerance indices in upland cotton under water deficit conditions. *Agronomy* 13, 984. doi: 10.3390/agronomy13040984

Ayaz, S., Mckenzie, B., Hill, G., and Mcneil, D. (2004). Variability in yield of four grain legume species in a subhumid temperate environment. II. Yield components. *J. Agric. Sci.* 142, 21–28. doi: 10.1017/S0021859604004113

Balko, C. (2005). Drought tolerance in potatoes and faba beans. Through breeding research to suitable crop plants.Forschungs-Report. *Ernahrung Landwirtschaft Forsten* 1), 10–13.

Bates, C. J., Waldren, R. P., and Teare, I. D. (1973). Rapid determination of free proline for water - stress studies. *Plant Soil* 39, 205–207. doi: 10.1007/BF00018060

Becker, H. (2011). *Pflanzenzüchtung* (Stuttgart: UTB GmbH).

Benkeblia, N., Shinano, T., and Osaki, M. (2007). Metabolite profiling and assessment of metabolome compartmentation of soybean leaves using non-aqueous fractionation and GC-MS analysis. *Metabolomics* 3, 297–305. doi: 10.1007/s11306-007-0078-y

Bouslama, M., and Schapaugh, W. T. (1984). Stress tolerance in soybeans. I. Evaluation of three screening techniques for heat and drought tolerance. *Crop Sci.* 24, 933–937. doi: 10.2135/cropsci1984.0011183X002400050026x

Daryanto, S., Wang, L., and Jacinthe, P. A. (2015). Global synthesis of drought effects on food legume production. *PloS One* 10, e0127401. doi: 10.1371/journal.pone.0127401

El-Hendawy, S. E., Hassan, W. M., Al-Suhaibani, N. A., and Schmidhalter, U. (2017). Spectral assessment of drought tolerance indices and grain yield in advanced spring wheat lines grown under full and limited water irrigation. *Agric. Water Manage.* 182 (7), 1–12. doi: 10.1016/j.agwat.2016.12.003

FAOSTAT (2022). FAOSTAT (FAOSTAT). Available at: http://www.fao.org/faostat/en/#data (Accessed October, 2021).

Farooq, M., Hussain, M., and Siddique, K. H. M. (2014). Drought stress in wheat during flowering and grain-filling periods. *Crit. Rev. Plant Sci.* 33, 331–349. doi: 10.1080/07352689.2014.875291

Farooq, M., Hussain, M., Wahid, A., and Siddique, K. H. M. (2012). "Drought stress in plants: An overview," in *Plant responses to drought stress: From morphological to molecular features, 1st edn*. Ed. R. Aroca (Berlin Heidelberg: Springer), 1–33.

Farshadfar, E., Mohammadi, R., Farshadfar, M., and Dabiri, S. (2013). Relationships and repeatability of drought tolerance indices in wheat-rye disomic addition lines. *Aust. J. Crop Sci.* 7 (1), 130–138.

Fernandez, G. C. (1992). "Effective selection criteria for assessing plant stress tolerance," in *Proceedings of the International Symposium on Adaptation of Vegetables and Other Food Crops in Temperature and Water Stress*, Adaptation of food crops to temperature and water stress. (Shanhua, Taiwan: AVRDC) Kuo, C. George Book editor. Vol. 18. 257–270.

Gnanasambandam, A., Paull, J., Torres, A., Kaur, S., Leonforte, T., Li, H., et al. (2012). Impact of molecular technologies on faba bean (*Vicia faba* L.) breeding strategies. *Agronomy* 2, 132–166. doi: 10.3390/agronomy2030132

Grashoff, C. (1990). Effect of pattern of water supply on *Vicia faba* L. 1. Dry matter partitioning and yield variability. *Netherlands J. Agric. Sci.* 38 (1), 21–44. doi: 10.18174/njas.v38i1.16608

Gutierrez, N., Pégard, M., Balko, C., and Torres, A. M. (2023). Genome-wide association analysis for drought tolerance and associated traits in faba bean (*Vicia faba* L.). *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1091875

Hack, H., Bleiholder, H., Buhr, L., Meier, U., Schnock-Fricke, U., Weber, E., et al. (1992). Einheitliche Codierung der phänologischen Entwicklungsstadien mono- und dikotyler Pflanzen – Erweiterte BBCH-Skala, Allgemein. *Nachr.bl. Dtsch. Pflanzenschutzd* 44 (12), 265–270.

Hannok, P., Kaewunta, P., Khunyota, W., Kaewnut, A., Pachimkanthong, P., and Tangthavonkarn, C. (2021). Exploration of phenotypic dissimilarity for drought tolerance in maize inbred line collection. *Curr. Appl. Sci. Technol.* 23, (1). doi: 10.55003/cast.2022.01.23.001

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Appl. Stat.* 28, 100–108. doi: 10.2307/2346830

Hekneby, M., Antolin, M. C., and Sanchez-Diaz, M. (2006). Frost resistance and biochemical changes during cold acclimation in different annual legumes. *Environ. Exp. Bot.* 55 (3), 305–314. doi: 10.1016/j.envexpbot.2004.11.010

Jensen, E. S., Peoples, M., and Hauggaard-Nielsen, H. (2010). Review: Faba beans in cropping systems. *Field Crop Res.* 115, 203–216. doi: 10.1016/j.fcr.2009.10.008

Kabbadj, A., Makoudi, B., Mouradi, M., Pauly, N., Frendo, P., and Ghoulam, C. (2017). Physiological and biochemical responses involved in water deficit tolerance of nitrogen-fixing *Vicia faba*. *PloS One* 12 (12), e0190284. doi: 10.1371/journal

Katsoulieri, A., Papastylianou, P., Travlos, I., Vlachostergios, D., Tigka, E., Kargiotidou, A., et al. (2020). Yield performance of faba bean cultivars under different environmental conditions in Greece. *Bull. UASVM Horticult.* 77 (2), 30–36. doi: 10.15835/buasvmcn-hort:2020.0016

Kaur, G., and Asthir, B. (2015). Proline: a key player in plant abiotic stress tolerance. *Biol. Plant* 59, 609–619. doi: 10.1007/s10535-015-0549-3

Khan, H. R., Link, W., Hocking, T. J., and Stoddard, F. L. (2007). Evaluation of physiological traits for improving drought tolerance in faba bean (*Vicia faba* L.). *Plant Soil* 292, 205–217. doi: 10.1007/s11104-007-9217-5

Khan, H. R., Paull, J. G., Siddique, K. H. M., and Stoddard, F. L. (2010). Faba bean breeding for drought-affected environments: A physiological and agronomic perspective. *Field Crops Res.* 115, 279–286. doi: 10.1016/j.fcr.2009.09.003

Kassambara, A., and Mundt, F. (2020). Extract and visualize the results of multivariate data analyses [R package factoextra] (CRAN).

Lafitte, H. R., Blum, A., and Atlin, G. (2003). "Using Secondary Traits to Help Identify Drought-Tolerant Genotypes," in *Breeding Rice for Drought-Prone Environments*. Eds. K. S. Fischer, R. H. Lafitte, S. Fukai, G. Atlin and B. Hardy (Manila: International Rice Research Institute), 37–48.

Link, W., Abdelmula, A. A., Kittlitz, E. V., Bruns, S., Riemer, H., and Stelling, D. (1999). Genotypic variation for drought tolerance in *Vicia faba*. *Plant Breed* 118, 477–483. doi: 10.1046/j.1439-0523.1999.00412.x

Lopez, F. B., Johansen, C., and Chauhan, Y. S. (1996). Effects of timing of drought stress on phenology, yield and yield components of short-duration pigeonpea. *J. Agron. Crop Sci.* 177 (5), 311–320. doi: 10.1111/j.1439-037X.1996.tb00251.x

Manning, B. K., Adhikari, K. N., and Trethowan, R. (2020). Impact of sowing time, genotype, environment and maturity on biomass and yield components in faba bean (*Vicia faba*). *Crop Pasture Sci.* 71, 147–154. doi: 10.1071/CP19214

Mansour, E., Mahgoub, H. A. M., Mahgoub, S. A., El-Sobky, E.-S.E.A., Abdul-Hamid, M. I., Kamara, M. M., et al. (2021). Enhancement of drought tolerance in diverse Vicia faba cultivars by inoculation with plant growth-promoting rhizobacteria under newly reclaimed soil conditions. *Sci. Rep.* 11 (1), 24142. doi: 10.1038/s41598-021-02847-2

McDonald, G. K., and Paulsen, G. M. (1997). High temperature effects on photosynthesis and water relations of grain legumes. *Plant Soil* 196, 47–58. doi: 10.1023/A:1004249200050

Memari, M., Dadashi, M. R., Sheikh, F., and Ajamnoruzi, H. (2022). Evaluation of drought stress tolerance in some faba bean genotypes using drought tolerance indices. *Romanian Agric. Res.* n.39. 11, 139–151 Print ISSN 1222-4227; online ISSN 2067-5720. doi: 10.59665/rar3904

Migdadi, H. M., El-Harty, E. H., Salamh, A., and Khan, M. A. (2016). Yield and proline content of faba bean genotypes under water stress treatments. *J. Anim. Plant Sci.* 26 (6), 1772–1779.

Mohamamdi, R., Etminan, A., and Shooshtari, L. (2017). Statistical methods for identifying traits associated with high yield potential in durum wheat under drought conditions. *Indian J. Genet. Plant Breed.* 77 (4), 513–523. doi: 10.5958/0975-6906.2017.00068.2

Morgan, J. M. (1984). Osmoregulation and water stress in higher plants. *Ann. Rev. Plant Physiol.* 35, 299–319. doi: 10.1146/annurev.pp.35.060184.001503

Muktadir, A., Adhikari, K. M., Merchant, A., Belachew, K. Y., Vandenberg, A., Stoddard, F. L., et al. (2020). Physiological and biochemical basis of faba bean breeding for drought adaptation – a review. *Agronomy* 10, 1345. doi: 10.3390/agronomy10091345

Müller, U., Meyer, C., Ehlers, W., and Bohm, W. (1985). Water-uptake and water-use of field beans and oats grown on a loess-derived grey-brown podzolic soil (EUTROBORALF). *Z. Pflanzenernähr. Bodenkd.* 4, 389–400. doi: 10.1002/jpln.19851480404

Mutlu, S. S., Mutlu, N., Karaguzel, O., Sari, H., and Shearman, R. C. (2022). Drought resistance of Bermudagrass accessions collected from Eastern Mediterranean. *Eur.J.Hortic.Sci* 87 (3), 1–12. doi: 10.17660/eJHS.2022/034

Mwanamwenge, J., Loss, S. P., Siddique, K. H. M., and Cocks, P. S. (1999). Effect of water stress during floral initiation, flowering and podding on the growth and yield of faba bean (Vicia faba L.). *Eur. J. Agron.* 11 (1), 1–11. doi: 10.1016/S1161-0301(99)00003-9

Nadeem, M., Li, J. J., Yahya, M., Sher, A., Ma, C. X., Wang, X. B., et al. (2019). Research progress and perspective on drought stress in legumes: A review. *Int. J. Mol. Sci.* 20 (10), 2541. doi: 10.3390/ijms20102541

Or-Rashid, M. H., Miah, M. A. B., and Islam, S. M. S. (2021). *In vitro* morphophysiological screening of drought tolerant potato genotypes. *In Vitro Cell. Dev. Biol.* 57, 519–528. doi: 10.1007/s11627-020-10155-z

Ouji, A., Naouari, M., Mouelhi, M., and Ben Younes, M. (2017). Yield and yield components of faba bean (Vicia faba L.) as influenced by supplemental irrigation under semi-arid region of Tunisia. *World J. Agric. Res. World J. Agric. Res.* 5 (1), 52–57. doi: 10.12691/wjar-5-1-7

Pilbeam, C. J., Akatse, J. K., Hebblethwaite, P. D., and Wright, C. D. (1992). Yield production in two contrasting forms of spring-sown faba beans in relation to water supply. *Field Crops Res.* 29, 273–287. doi: 10.1016/0378-4290(92)90030-D

Plies-Balzer, E., Kong, T., Schubert, S., and Mengel, K. (1995). Effect of water stress on plant growth, nitrogenase activity and nitrogen economy of 4 different cultivars of *Vicia faba L. Eur. J. Agron.* 4 (2), 167–173. doi: 10.1016/S1161-0301(14)80043-9

Ramirez-Vallejo, P., and Kelly, J. D. (1998). Traits related to drought resistance in common bean. *Euphytica* 99 (2), 127–136. doi: 10.1023/A:1018353200015

R Core Team (2022). *A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing). URL https://www.R-project.org/.

Redden, R., Paull, J., Zong, X., Sass, O., Yang, T., and Lintg, L. (2014). "Faba bean," in *Broadening the genetic base of grain legumes*. Eds. M. Singh, I. S. Bisht and M. Dutta (New Delhi: Springer), 75–93.

Reyes, J. A. O., Delfin, E. F., Reyes, M. E. C., and Ocampo, E. T. M. (2022). Physiological and yield responses of selected mungbean genotypes to terminal drought. *Philipp J. Sci.* 151 (2), 739–750. doi: 10.56899/151.02.16

Ricciardi, L., Polignano, G. B., and De Giovanni, C. (2001). Genotypic response of faba bean to water stress. *Euphytica* 118, 39–46. doi: 10.1023/A:1004078017159

Richards, R. A. (2006). Physiological traits used in the breeding of new cultivars for water-scarce environments. *Agric. Water Manag* 80 (1-3), 197–211. doi: 10.1016/j.agwat.2005.07.013

Romagosa, I., van Eeuwijk, F. A., and Thomas, W. T. (2009). "Statistical analyses of genotype by environment data," in *Cereals* (Dordrecht, The Netherlands: Springer), 291–331.

Sanchez, D. H., Schwabe, F., Erban, A., Udvardi, M. K., and Kopka, J. (2012). Comparative metabolomics of drought acclimation in model and forage legumes. *Plant Cell Environ.* 35, 136–149. doi: 10.1111/j.1365-3040.2011.02423.x

Santos, R., Carvalho, M., Rosa, E., Carnicle, V., and Castro, I. (2020). Root and agro-morphological traits performance in cowpea under drought stress. *Agronomy* 10 (10), 1604. doi: 10.3390/agronomy10101604

Sekara, A., Poniedzialek, M., Ciura, J., and Jedrszczyk, E. (2001). The effect of meteorological factors upon flowering and pod setting of faba bean (Vicia faba L.) at different sowing times. *Veg. Crop Res. Bull.* 54 (1), 65–68.

Shah, T. M., Imran, M., Atta, B. M., Ashraf, M. Y., Hameed, A., Waqar, I., et al. (2020). Selection and screening of drought tolerant high yielding chickpea genotypes based on physio-biochemical indices and multi-environmental yield trials. *BMC Plant Biol.* 20, 1–16. doi: 10.1186/s12870-020-02381-9

Shakeel, A. A., Xiao-yu, X., Long-chang, W., Muhammad, F. S., Chen, M., and Wang, L. (2011). Morphological, physiological and biochemical responses of plants to drought stress. *Afr. J. Agric. Res.* 6 (9), 2026–2032. doi: 10.5897/AJAR10.027

Sharifi, P., Astereki, H., Sheikh, F., and Khorasanizadeh, M. (2021). Evaluation of faba bean genotypes in normal and drought stress conditions by tolerance and susceptibility indices. *CAJPSI* 1 (4), 176–191. doi: 10.22034/CAJPSI.2021.04.0

Siahsar, B. A., Ganjali, S., and Allahdoo, M. (2010). Evaluation of drought tolerance indices and their relationship with grain yield of lentil lines in drought-stressed and irrigated environments. *Aust. J. basic Appl. Sci.* 4 (9), 4336–4346.

Siddiqui, M. H., Al-Khaishany, M. Y., Al-Qutami, M. A., Al-Whaibi, M. H., Grover, A., Ali, H. M., et al. (2015). Response of different genotypes of faba bean plant to drought stress. *Int. J. Mol. Sci.* 16, 10214–10227. doi: 10.3390/ijms160510214

Sofi, P. A., Djanaguiraman, M., Siddique, K. H. M., and Prasad, P. V. V. (2018). Reproductive fitness in common bean (*Phaseolus vulgaris* L.) under drought stress is associated with root length and volume. *Ind. J. Plant Physiol.* 23, 796–809. doi: 10.1007/s40502-018-0429-x

Stoddard, F. L., Balko, C., Erskine, W., Khan, H. R., Link, W., and Sarker, A. (2006). Screening techniques and sources of resistance to abiotic stresses in cool-season food legumes. *Euphytica* 147 (1-2), 167–186. doi: 10.1007/s10681-006-4723-8

Swann, A. L., Hoffman, F. M., Koven, C. D., and Randerson, J. T. (2016). Plant responses to increasing CO2 reduce estimates of climate impacts on drought severity. *Proc. Natl. Acad. Sci. U S A.* 113 (36), 10019–10024. doi: 10.1073/pnas.1604581113

Tadesse, T., Fikere, M., Legesse, T., and Parven, A. (2011). Correlation and path coefficient analysis of yield and its components in faba beans. *Int. J. Biodivers. Conserv.* 3, 376–382.

Tofiq, S. E., Aziz, O. K., and Salih, S. H. (2016). Correlation and path coefficient analysis of seed yield and yield components in some faba bean genotypes in Sulaimani region. *ARO-The Sci. J. Koya Univ.* 4 (2), 1–6. doi: 10.14500/aro.10081

Torres, A. M., Avila, C. M., Gutierrez, N., Palomino, C., Moreno, M. T., and Cubero, J. I. (2010). Marker-assisted selection in faba bean (Vicia faba L.). *Field Crops Res.* 115 (3), 243–252. doi: 10.1016/j.fcr.2008.12.002

Toum, L., Perez-Borroto, L. S., Peña-Malavera, A. N., Luque, C., Welin, B., Berenstein, A., et al. (2022). Selecting putative drought-tolerance markers in two contrasting soybeans. *Sci. Rep.* 12, 10872. doi: 10.1038/s41598-022-14334-3

Tuberosa, R. (2012). Phenotyping for drought tolerance of crops in the genomics era. *Front. Physiol.* 3. doi: 10.3389/fphys.2012.00347

Wahab, A., Abdi, G., Saleem, M. H., Ali, B., Ullah, S., Shah, W., et al. (2022). Plants physio-biochemical and phyto-hormonal responses to alleviate the adverse effects of drought stress: A comprehensive review. *Plants-Basel* 11, 1620. doi: 10.3390/plants11131620

Weng, T., Nakashima, T., Villanueva-Morales, A., Stewart, J. R., Sacks, E., and Yamada, T. (2022). Assessment of drought tolerance of Miscanthus genotypes through dry-down treatment and fixed-soil-moisture-content techniques. *Agriculture* 12, 6. doi: 10.3390/agriculture12010006

Yemm, E. W., and Willis, A. J. (1954). The estimation of carbohydrates in plant extracts by anthrone. *Biochem. J.* 57, 508–514. doi: 10.1042/bj0570508

Yoshiba, Y., Kiyosue, T., Nakashima, K., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1997). Regulation of levels of proline as an osmolyte in plants under water stress. *Plant Cell Physiol.* 38 (10), 1095–1102. doi: 10.1093/oxfordjournals.pcp.a029093

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 January 2024
DOI 10.3389/fpls.2023.1304411

Check for updates

OPEN ACCESS

EDITED BY
Åshild Ergon,
Norwegian University of Life Sciences, Norway

REVIEWED BY
Lina Ahmed,
INRAE Nouvelle Aquitaine Poitiers, France
Mariola Staniak,
Institute of Soil Science and Plant Cultivation, Poland
Annick Bertrand,
Agriculture and Agri-Food Canada (AAFC), Canada

*CORRESPONDENCE
Tim Vleugels
tim.vleugels@ilvo.vlaanderen.be

†PRESENT ADDRESSES
Aamir Saleem,
Plant Breeding, Wageningen University and Research, Wageningen, Netherlands
Reena Dubey,
Faculty of Bioscience Engineering, Ghent University, Gent, Belgium
Irene Borra-Serrano,
ICA (Institute of Agricultural Sciences), Spanish National Research Council (CSIC), Tec4agro Group, Madrid, Spain

RECEIVED 29 September 2023
ACCEPTED 18 December 2023
PUBLISHED 12 January 2024

CITATION
Vleugels T, Saleem A, Dubey R, Muylle H, Borra-Serrano I, Lootens P, De Swaef T and Roldán-Ruiz I (2024) Phenotypic characterization of drought responses in red clover (*Trifolium pratense* L.). *Front. Plant Sci.* 14:1304411. doi: 10.3389/fpls.2023.1304411

COPYRIGHT
© 2024 Vleugels, Saleem, Dubey, Muylle, Borra-Serrano, Lootens, De Swaef and Roldán-Ruiz. 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.

Phenotypic characterization of drought responses in red clover (*Trifolium pratense* L.)

Tim Vleugels*, Aamir Saleem†, Reena Dubey†, Hilde Muylle, Irene Borra-Serrano†, Peter Lootens, Tom De Swaef and Isabel Roldán-Ruiz

ILVO (Flanders Research Institute for Agriculture, Fisheries and Food), Plant Sciences Unit, Melle, Belgium

Introduction: Red clover (*Trifolium pratense*) is a protein-rich, short-lived perennial forage crop that can achieve high yields, but suffers increasingly from drought in different cultivation areas. Breeding for increased adaptation to drought is becoming essential, but at this stage it is unclear which traits breeders should target to phenotype responses to drought that allow them to identify the most promising red clover genotypes. In this study, we assessed how prolonged periods of drought affected plant growth in field conditions, and which traits could be used to distinguish better adapted plant material.

Methods: A diverse panel of 395 red clover accessions was evaluated during two growing seasons. We simulated 6-to-8-week drought periods during two consecutive summers, using mobile rain-out shelters, while an irrigated control field was established in an adjacent parcel. Plant growth was monitored throughout both growing seasons using multiple flights with a drone equipped with RGB and thermal sensors. At various observation moments throughout both growing seasons, we measured canopy cover (CC) and canopy height (CH). The crop water stress index (CWSI) was determined at two moments, during or shortly after the drought event.

Results: Manual and UAV-derived measurements for CH were well correlated, indicating that UAV-derived measurements can be reliably used in red clover. In both years, CC, CH and CWSI were affected by drought, with measurable growth reductions by the end of the drought periods, and during the recovery phase. We found that the end of the drought treatment and the recovery phase of approximately 20 days after drought were suitable periods to phenotype drought responses and to distinguish among genotypes.

Discussion: Multifactorial analysis of accession responses revealed interactions of the maturity type with drought responses, which suggests the presence of two independent strategies in red clover: 'drought tolerance' and 'drought recovery'. We further found that a large proportion of the accessions able to perform well under well-watered conditions were also the ones that were less affected by drought. The results of this investigation are interpreted in view of the development of breeding for adaptation to drought in red clover.

KEYWORDS

Trifolium pratense, drought responses, phenotyping, UAV, canopy temperature, HTFP

Introduction

Red clover (*Trifolium pratense* L.) is an important forage legume crop in temperate regions around the world (Annicchiarico and Pagnotta, 2012). It is either grown in mixtures with grasses or as a monoculture (Taylor and Quesenberry, 1996). Farmers appreciate red clover for its capacity to fix atmospheric nitrogen in symbiosis with soil bacteria, its high forage yield and crude protein content, and its high palatability (Kjærgaard, 2003). Red clover is a short-lived perennial that remains productive for about three years, after which plants start to degenerate. Plants develop a deep, well-branched taproot during the first year after sowing, that is later complemented with adventitious roots (Taylor and Quesenberry, 1996). Commercial varieties are usually classified into two maturity types (Boucher, 2016), as this has important implications for their agronomic use. 'Double-cut' varieties develop early in spring, have a strong regrowth capacity, a higher shoot/root ratio, and keep growing until late in autumn if environmental conditions are favorable (Hejduk and Knot, 2010). 'Single-cut' varieties initiate growth later in spring, display slower re-growth capacity after mowing, and exhibit earlier autumn dormancy - traits that render them more persistent and winter-hardy (Hejduk and Knot, 2010). Single-cut varieties require longer photoperiods than double-cut varieties to initiate flowering (day lengths of 14 h or more), and rarely flower in-between mowing events (Taylor and Quesenberry, 1996; Boucher, 2016). Single-cut types are commonly grown in Nordic climates, while the double-cut types are better adapted to temperate and warm climates (Taylor and Quesenberry, 1996).

In the light of climate change, it is crucial that breeders anticipate and create new red clover varieties that can cope with periods of drought (Boucher, 2016). While red clover has been traditionally regarded as a fairly drought-tolerant crop when compared to other forages (Annicchiarico et al., 2015), changing climatic conditions in Europe are expected to bring about warmer summers with longer periods of drought and less rainfall (IPCC, 2023). Consequently, drought is expected to decrease productivity in European forage crops (Dellar et al., 2018), possibly also in red clover. Indeed, some red clover varieties, such as those selected for cultivation in temperate and Nordic climates, might display a high level of drought susceptibility.

Investigating the response of crops to drought is however not simple. Plants can express different mechanisms to cope with drought (Fang and Xiong, 2015), and according to Aslam et al. (2015) at least four responses can be distinguished: (1) drought escape, which prevents exposure to terminal drought stress through earliness; (2) drought avoidance, which implies increasing water uptake or reducing water losses during periods of drought; (3) drought tolerance, which maintains physiological processes during drought stress and allows to conserve a certain degree of productivity during drought; and (4) quick drought recovery, which represents the speed at which plant productivity resumes to a normal level after drought. In forage crops, mainly 'drought tolerance' and 'drought recovery' have been described (Pembleton and Sathish, 2014) but few studies have investigated specifically the mechanisms active in red clover. For example, in the related legume crop alfalfa (*Medicago sativa* L.), drought tolerance has been associated with the capacity to limit water losses through the reduction of transpiration rates or the reduction of leaf size and aboveground biomass, and with the capacity to absorb more water, through the formation of more lateral roots (Hanson et al., 2015; Quan et al., 2016). Drought-tolerant plants protect cell membranes from oxidative damage, and display higher antioxidant enzyme activity during stress (Quan et al., 2016). In both alfalfa and red clover, drought recovery has been associated with the ability to maintain viable tissues in the crown and with the availability of metabolites to support re-growth (Pembleton et al., 2010; Loucks et al., 2018). In these perennial crops, drought stress reduces the number and size of viable crown buds, and lowers starch concentrations in the taproot, both of which are essential for regrowth after a period of drought (Pembleton et al., 2010).

An important prerequisite for red cover breeders to identify genotypes adapted to drought conditions is an evaluation of the variation available in breeding germplasm and the definition of effective selection criteria. However, the response of red clover to drought has not been thoroughly characterized yet. Available publications either describe work performed in growth chambers using only a few accessions (Loucks et al., 2018), report the evaluation of accessions of narrow geographical ranges (Annicchiarico and Pagnotta, 2012), or focus on the molecular mechanisms associated with some responses to drought, such as the abundance of stress-related proteins (Vaseva et al., 2011) or changes in the transcriptome (Yates et al., 2014). To our knowledge, no prior studies have described the response to drought of a large collection of red clover genotypes in terms of growth and productivity, nor have they proposed specific easy-measurable traits that can be used by breeders. In breeding context, phenotyping drought responses is preferably done in the field, and rain-out shelters that block the precipitation, while having minimal effects on temperature and light conditions are ideal tools (Parra et al., 2012; De Swaef et al., 2021). In temperate climates, the most relevant period to evaluate the drought response of red clover is early summer, right after the first cut. In this period, soil moisture reserves built up during the winter can already be depleted, and the plants rely mostly on precipitation for growth. Furthermore, an optimal re-growth in this period is essential to obtain a good yearly productivity. All these aspects should be considered when designing relevant screening experiments.

The present work was performed in the framework of the EUCLEG project (Horizon 2020 Programme for Research and Innovation). A strategic goal of EUCLEG was to develop efficient breeding strategies in five legume crops, among which red clover, by analyzing key agronomic and quality traits. Here we present the work performed to characterize for the first time the diversity of responses to drought available in a collection of 395 red clover accessions from 14 countries. Field-grown plants were subjected to two periods of drought in early summer during two production years. Forage crop breeding relies mostly on visual observations and manual measurements, but high-throughput field

Abbreviations: BLUP, best linear unbiased predictor; CC, UAV-derived canopy cover; CH, UAV-derived canopy height; CH_rpm, manual measurement for canopy height; CWD, cumulative water deficit; CWSI, crop water stress index; DOY, day of the year; PCA, principal component analysis; UAV, unmanned aerial vehicle; Yr, relative performance index; HTFP, High-Throughput Field Phenotyping.

phenotyping (HTFP) using unmanned aerial vehicles (UAV) equipped with different imaging sensors are a promising alternative, especially when dealing with large numbers of plants or field plots (Araus and Cairns, 2014; Borra-Serrano et al., 2019). An additional advantage of UAV measurements is that objective, repeated measurements can be made over a course of time to follow-up plant growth in detail over an entire growing season. While, to the best of our knowledge, the use of indices derived from RGB or thermal sensors mounted on an UAV has not yet been explored in red clover, they have proven very useful to determine the drought responses of many other crops such as forage grasses (De Swaef et al., 2021), alfalfa (Surault et al., 2021), and soybean (Saleem et al., 2022).

The objectives of our study were: (i) to characterize for the first time the responses to drought of a diverse panel of red clover accessions; (ii) to appraise the use of UAV-derived measurements in characterizing growth and drought responses in red clover; (iii) to assess the "broad-sense" heritability of traits related to drought responses in this diverse collection, (iv) to assess whether the response of red clover to drought differs between the first and the second year of production, and (v) to identify traits that can be targeted by breeders to improve adaptation to drought in European red clover.

Materials and methods

Experimental design

The plant materials used in this study have been described by Frey et al. (2022). Here we used a subset of 395 red clover accessions of diverse geographical origin and including different population types: ecotypes, landraces, cultivars and breeding material from 14 countries (Supplementary Table S1).

Our trial comprised two adjacent fields of sandy loam soil, which were sown in August 2018 in Melle, Belgium (51.00° N, 3.80° E) (Figure 1A). One of these fields, further named the 'control field', was irrigated and served as control. The other field, further named the 'drought field', was subjected to drought using mobile rain-out shelters as described below. Both fields were established following an identical, incomplete block design with two randomized repetitions of 395 accessions, distributed over three blocks, with each rain-out shelter representing one block (Figure 1). Each block measured 10 m x 30 m and contained five strips (further named columns) with 53 individual plots (further named rows), spaced 45 cm apart. Each plot represented a single accession sown in a 1 m line. All accessions were replicated twice over the three blocks, except the control cultivar 'Lemmon' which was replicated seven times to prevent empty entries in the design. The control and drought fields each contained a total of 795 plots (3 blocks x 5 columns x 53 rows). Around each block, border strips containing 'Lemmon' were sown to reduce border effects. The control and drought fields were treated identically in terms of sowing, mowing regime, fertilization, weed control and most observations.

In the drought field, the perennial crop was subjected to early summer drought in two consecutive years, to evaluate the response to drought of plants in different stages of development, i.e. plants in the first production year, which do not yet have a fully developed root system, and the same plants in the second production year, which have a more developed root system but are also older. The

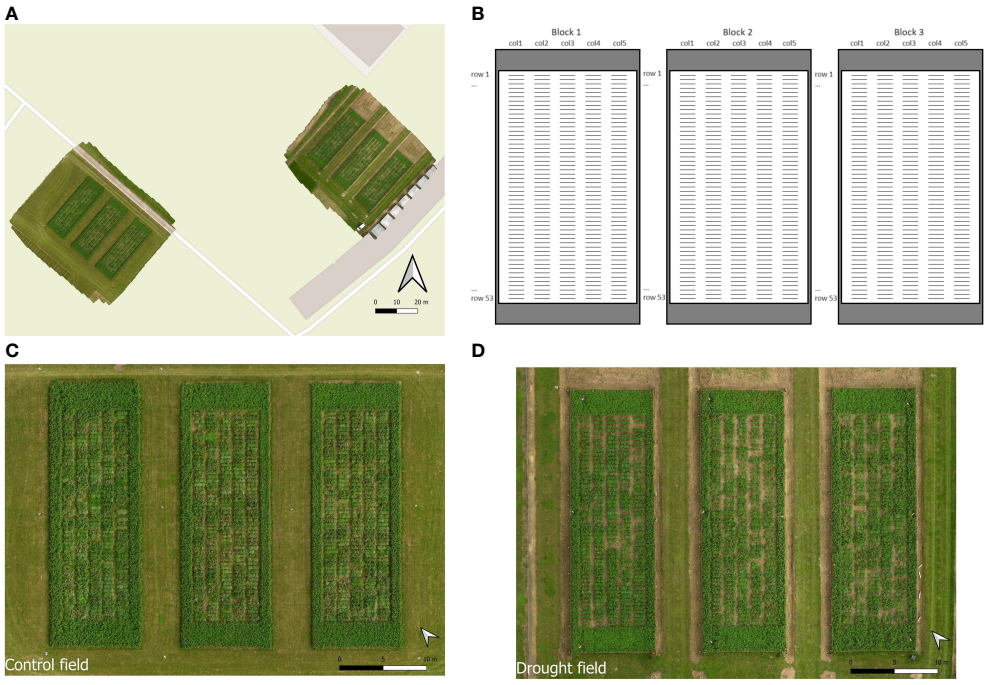


FIGURE 1
Overview of the trial with **(A)** a UAV-derived picture of the control (left) and drought (right) fields, **(B)** a schematic overview of the organization of the fields with blocks, columns and rows indicated, and UAV-derived photographs of the control **(C)** and drought **(D)** fields.

two trial years cannot be considered as replicates, as the developmental status of the perennial crop was different.

The maturity type of each accession was not known in advance, but it was possible to identify the accessions' maturity using flowering data recorded during the first production year. For each accession, we had data available from four plots in our trial: two control plots and two drought plots. We classified an accession as 'early flowering' when it flowered in at least one plot – representing mostly double-cut types, and 'late flowering' when it failed to flower in any of the four plots – representing mostly single-cut types.

Agronomic management

The trial was sown manually at 1 cm depth on a finely prepared seed bed. Each plot was sown with 90 germinable seeds. The number of actually sown seeds per plot depended on the germination of each accession seed lot, which was provided by the EUCLEG consortium: 90 seeds were sown in case of 100% germination, and more seeds in case of lower germination. After sowing, a synthetic fertilizer was applied, providing 12:24:60 units NPK. During the establishment phase, weeds were controlled manually by hoeing and grass weeds were controlled with a single application of fluazifop-P-butyl (Fusilade Max®, Nufarm B.V., Belgium) at 1.5 lha^{-1}. Due to dry weather in autumn 2018, germination was poor for some red clover accessions. Therefore, seedlings within poorly germinated plots were transplanted so that plants were equally distributed in the plot.

Agronomic activities, drought periods and observations are presented schematically in Figure 2. The crop was mown four times per growing season, as is usual practice in temperate European red clover. In the first production year (2019), further referred to as year 1, the crop was mown on DOY 134, 185, 232 and 277. In the second production year (2020), further named year 2, the crop was mown on DOY 134, 198, 246 and 290. Mowing was done with a Haldrup mower at 7 cm height. In year 1, weeds were controlled by carbetamide (Legurame®, Belchim Crop Protection, Belgium) at 5.0 lha^{-1} on DOY 51 and isoxaben (AZ500®, Dow Agrosciences, Belgium) at 0.1 lha^{-1} on DOY 78. Weed control was similar in year 2, with herbicide applications on DOY 69 and 106. Regular hoeing during the rest of the growing season kept the weeds under control. Both fields were fertilized in three doses with 0:45:328 units NPK on a yearly basis.

The drought treatment was initiated each year right after the first cut and maintained until soil moisture content reached 6% (v/v), close to the permanent wilting point (Loucks et al., 2018). At the end of the drought period, the second cut was taken and a modest irrigation was applied to the drought-treated field. In year 1, drought was maintained for 50 days from DOY 134 to 184. Due to unusual dry weather conditions, the control field was irrigated through sprinklers at a rate of 23 lm^{-2} on DOY 144, 152 and 167. After removal of the rain-out shelters, the drought field was irrigated through hosing at a rate of 10 lm^{-2} on DOY 185 and 191. In year 2, drought was initiated on DOY 134 and maintained for 69 days until DOY 203. The control field was irrigated with 23 lm^{-2} through sprinklers on DOY 143 and 152, except the first block that was accidentally provided 794 lm^{-2}. This accidental irrigation did not seem to have greatly influenced our results: the effect was compensated by the block effects in our mixed model analysis. After removal of the rain-out shelters, the drought field was irrigated by sprinklers on DOY 204 (23 lm^{-2}), 219 (29 lm^{-2}) and 224 (32 lm^{-2}).

Temperature, precipitation and daily solar shortwave radiation for the trial site were available from the Royal Meteorological Institute (KMI), in the form of interpolated data from nearby weather stations, the nearest station about 4 km away at 50° 58'49"N, 3°48'57"E. The weather conditions in years 1 and 2 are summarized in Supplementary Figure S1. Precipitation and reference evapotranspiration data were used to calculate the cumulative water deficit (CWD) as described in Saleem et al. (2022) (Figure 3). To assess possible differences in soil water content between blocks, soil moisture content was monitored at

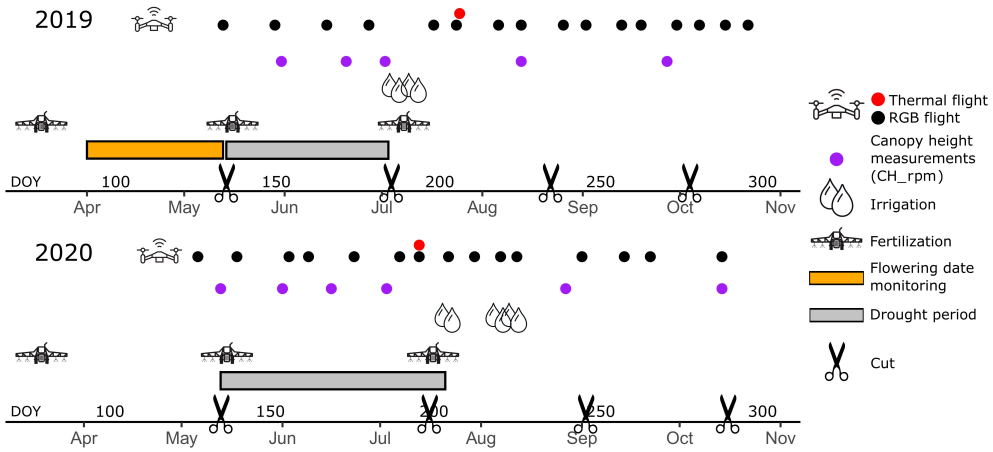


FIGURE 2
Timeline representing the two trial years for the drought and the control fields with schematic indication of the drought periods, mowing events, management, irrigation, and observations that were used in this study.

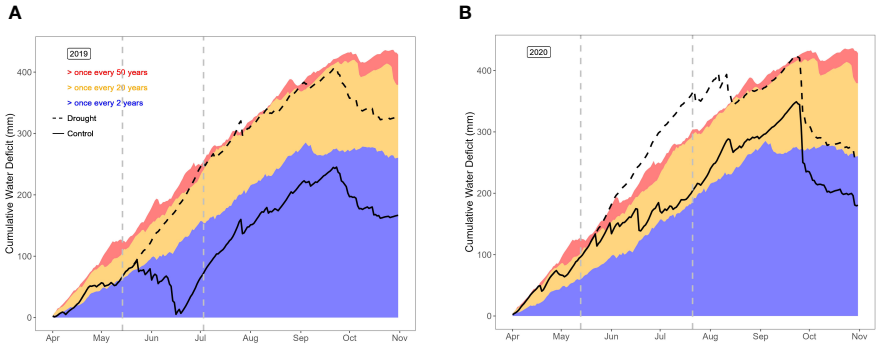


FIGURE 3
Cumulative water deficit in the control and drought fields in year 1 **(A)** and year 2 **(B)**, compared to historical data for the region of the trial.

regular occasions in the control and drought fields. At each event, 4 soil samples (0 – 30 cm) were taken per block, samples were oven-dried (70°C for 48h), and soil moisture was expressed as % (v/v).

Data acquisition and processing

Ground and UAV-derived measurements were made in both fields throughout the entire duration of the trial. In year 1, visual observations were made for flowering date until the first cut (Table 1, Figure 2). Throughout the growing seasons of years 1 and 2, canopy height (CH_rpm) was measured several times in a representative part of each plot using a rising plate meter (RPM, HerboMETRE, ARVALIS-Institut du Végétal, France) consisting of a square rising plate of 0.09 m^2 (30 × 30 cm). Additionally, regular UAV flights were performed to estimate the Canopy Cover (CC), Canopy Height (CH) and canopy temperature of each plot in the control and drought fields (Table 1). A total of 45 UAV flights was performed using a drone (model DJI Matrice 600 Pro; DJI, China) with an RGB sensor (α6000, 35mm lens, Sony Corporation, Japan) and a multispectral sensor (Micasense RedEdge MX RED in year 1 and Micasense RedEdge MX RED + MX BLUE in year 2). The flight altitude was set to 25 m, at a speed of 2.3 ms^{-1} and the side and forward overlap was 80-80%. Images were corrected and adjusted for white balance and exposure in Lightroom v6.5 (Adobe Systems Incorporated, USA) using a grey reference card (18% reference grey, Novoflex, Präzisionstechnik GmbH, Germany). The software Agisoft Metashape Professional v1.5.5 (Agisoft LLC, Russia) was used to stitch the images adding nine ground control points (coordinates measured with an RTK GPS Stonex S10 GNSS, Stonex SRL, Italy) to obtain a georeferenced orthomosaic and digital elevation model for each date. To extract data from each plot, rectangular polygons were created in QGIS 3.18.2 (QGIS Geographic Information System, QGIS Development Team, Open-Source Geospatial Foundation) and only pixels which corresponded to vegetation were considered. CC and CH were derived as in Borra-Serrano et al. (2020). For CH, Q90 values were used. Multispectral data were processed, but were not used further.

In order to assess the value of thermal data (CWSI values) for phenotyping drought tolerance in red clover, canopy temperature data were obtained from two successful thermal flights using a Wiris second generation (Workswell, Czech Republic) sensor. The two sets of thermal images (DOY 206 in year 1 - weather conditions 36°C and 38% RH; DOY 195 in year 2 – weather conditions 21°C and 54% RH) were pre-processed in ThermoFormat (Workswell, Czech Republic) and stitched in the software Pix4D Mapper 4.5.6 (Pix4D S.A., Switzerland). The CWSI (Maes and Steppe, 2012) was calculated as described in De Swaef et al. (2021).

TABLE 1 General information on the variables recorded through ground observations and UAV-derived measurements.

Name	Abbreviation	Acquired through	Description	Unit
Flowering date	FLD	Ground observation	Day of year (DOY) until the start of flowering in year 1. Non-flowering accessions are encoded 222.	DOY
Canopy cover	CC	UAV-derived	Canopy cover. Percentage of the polygon covered by the canopy.	%
Canopy height	CH	UAV-derived	Canopy height. Height of the canopy relative to the naked soil (Q90).	Cm
Canopy Height	CH_rpm	Ground measurement	Canopy height. Height of the canopy relative to the naked soil measured with a rising plate meter.	Cm
Crop water stress index	CWSI	UAV-derived	Relative transpiration rate at the time of measurement	%

10.3389/fpls.2023.1304411

Data analysis

Variables were named using the variable abbreviation (Table 1) followed by the DOY (e.g. CC_178 for canopy cover measured on DOY 178). Both, ground measured and UAV-derived measurements were filtered for faulty plots for each variable, i.e. plots in which the crop was not completely established by the start of the growing season in year 1, plots that experienced damage during an exceptional herbivore attack, and plots with outliers. Data from a limited number of UAV flights was not reliable: these data were not used. In total, we decided to use data from 31 flights. The filtered phenotypic data was further harnessed for statistical analysis. The number of retained plots per variable is represented in Supplementary Table S2.

All statistical analyses were carried out in R statistical software version 4.0.3 (R Core Team, 2022), implemented in RStudio (RStudio Team, 2022). A mixed model approach with the lme4 package (Bates et al., 2015) was performed to correct for environmental effects caused by block and position within block (row and column), and to obtain best linear unbiased predictor (BLUP) values for each accession.

The following base model was considered: Y = Intercept + Accession + Block + Column + Row + Residual.

Where, 'Y' = response variable, 'Intercept' = overall mean value of the response variable, 'Accession' = random effect representing the accession, 'Block', 'Column', 'Row' = random effects representing spatial components in the experimental design and 'Residual' = noise term. The random effects for Block, Column and Row were assumed to be independent, and originate from an identical, normal distribution. The residuals are also assumed to be independent and identically distributed. In the experimental design, column was nested in block. The base model was not applied as such because it would be 'overfitted' (incorporates the 'Block' and 'Column' as unique components while in our design columns were actually nested). For each response variable, six versions of the base model were tested:

1. Y = Intercept + Accession + Residual
2. Y = Intercept + Accession + Block + Residual
3. Y = Intercept + Accession + Column + Residual
4. Y = Intercept + Accession + Row + Residual
5. Y = Intercept + Accession + Block + Row + Residual
6. Y = Intercept + Accession + Column + Row + Residual

For each variable, the output was evaluated using the Akaike Information Criterion (Akaike, 1974), and the fit with the lowest value was chosen. From the best fit, the BLUP values were calculated for each accession as the sum of the 'Intercept' value and the value of the random effect of 'Accession'. The variance components of the fit were obtained using the VarCorr function from lme4 package (Bates et al., 2015). The broad sense heritability (H^2) was calculated from the variance components of the best fit as follows: $H^2 = V_G/(V_G + V_R)$, where V_G and V_R represent the variance due to 'Accession' and the residual variance, respectively. For each variable, the model used and the heritability is given in Supplementary Table S2.

Relative performance index Yr

Drought responses were measured in terms of growth reduction (using the proxies CC and CH) compared to control conditions (Bellague et al., 2016). To assess plant responses to drought, relative performance indices (Yr) were calculated for the UAV-derived observations CC, CH and CWSI, using the following formula: Yr = (Control – Drought)/Control. Yr variable names are composed of 'Yr' followed by the variable (CC, CH or CWSI) and the DOY (e.g. Yr_CC_178 for relative performance indices for canopy cover measured on DOY 178). A positive Yr value for a certain accession indicates a lower observation in terms of CC, CH or CWSI for that accession in the drought field compared to the control field, and vice versa. Two-sided Z-tests were used to assess for each variable and observation date whether average Yr values over all accessions differed statistically from zero, i.e. indicating no growth reduction due to drought.

Principal component analyses

Principal component analysis (PCA) was used to understand the relationships among variables in this complex dataset. First, PCA was performed on the data of the control field for each year separately to understand patterns of clustering and relations among CC, CH, maturity, variety type and geographic origin. At first, we included all observations for CC and CH between the 1st cut (start of drought) and the 4th cut (end of the growing season). As CC and CH observations taken during the same growing period were generally well correlated, we reduced the number of variables to 1 observation in the middle and 1 observation at the end of each of the three growing periods. By doing so, we obtained a set of informative CC and CH variables that explained an equal proportion of variation in the PCAs as the complete set of variables. In year 1, we retained observations for CC and CH on DOY 165, 178, 205 225, 256, and 273. In year 2, we retained observations for CC and CH on DOY 161, 195, 220, 245, 266, and 288.

Second, PCA was performed on the dataset of Yr indices to reveal key variables and time-points related to drought responses, using our informative set of CC and CH variables describes above, and the CWSI variables. FactoMineR (Lê et al., 2008) and factoextra (Kassambara and Mundt, 2020) packages in R were used to depict patterns related to maturity (early – late), variety type (breeding material, cultivars, landraces or ecotypes), and geographic origin (N Europe, W Europe, Central Europe, SE Europe, or Non-European material).

Relation between growth patterns under control and drought conditions

An important question is whether breeders can create varieties with improved performance under drought conditions by breeding in control conditions. To investigate this, we ranked all accessions according to the number of CC and CH variables for which they attained a value in the top 25%. In other words, as we had determined CC and CH at different moments in the two growing seasons, the number of times each accession ranked among the top 25% best-performing was counted. This allowed to identify the accessions with best overall performance for CC or CH in the control and drought fields in both years: accessions which a breeder would identify as 'promising'. For both years and both fields, we identified the 50 accessions (or slightly more in case of accessions with equal number of hits in the top 25%) with best overall performance in terms of CC and CH. This allowed to assess how many accessions in this subset performed best in both fields.

Results

Weather conditions and drought intensity in both trial years

Daily average temperatures and daily solar shortwave radiation for year 1 and year 2 are presented in Supplementary Figure S1. Year 1 (2019) and year 2 (2020) were warm, sunny and relatively dry with a warm, early spring and a warm summer. Daily average temperatures exceeded the long-term average of 10.6°C by 0.9°C and 1.6°C on a yearly basis, and total yearly precipitation was 54 mm and 121 mm below the long-term average of 852 mm, for year 1 and 2, respectively.

The weather conditions before, during, and after the drought periods were similar in both years. Before the drought treatment (DOY 1 to 134), both years displayed similar daily average temperatures (7.6 vs. 8.4°C), daily solar radiation (9.2 vs. 10.1 MJm^{-2}) and precipitation (248 vs. 280 mm), for year 1 and 2, respectively. During the drought periods (DOY 134-184 and 134-203 in year 1 and 2, respectively), daily average temperatures (16.5 vs. 16.6°C), daily solar radiation (20.4 vs. 20.7 MJm^{-2}) and cumulative precipitation (112 vs. 114 mm) were also similar. Both drought periods contained 12 warm days with maximal temperatures exceeding 25°C, and respectively 4 and 1 hot days with maximal temperatures exceeding 30°C. Major rainfall events during the drought periods occurred on DOY 156 (15.1 mm) and 165 (27.6 mm) in year 1, and DOY 169 (36.8 mm) and 182 (16.0 mm) in year 2. The recovery phases after the drought treatments and the remainder of the growing seasons (drought relief to DOY 300) remained warm in both years, with similar daily average temperatures (16.7°C in both years), daily solar radiation (13.4 vs. 12.5 MJm^{-2}) and precipitation (222 vs. 257 mm), for year 1 and 2, respectively. Multiple hot days with daily maximum temperatures exceeding 30°C occurred during the recovery phases: 8 days in year 1 and 9 days in year 2.

The cumulative water deficit (CWD) represents the soil moisture stress, by taking into account evapotranspiration, precipitation, irrigations applied in the control field, and the effect of the rain-out shelters in the drought field (Figure 3). In year 1, the control field represented nearly normal soil moisture conditions, with a short and modest CWD in early summer. The drought field, on the other hand, suffered a once-in-20-year-drought starting soon after the installation of the rain-put shelters until the end of the growing season, long after removal of the rain-out shelters. In year 2, drought was more harsh in both fields. In the control field, the CWD entered the once-in-20-year-drought zone early in the growing season, and remained at the margin of that zone until late in the growing season. The drought field suffered even more soil moisture stress, representing a once-in-50-year-drought from the installation of the rain-out shelters until mid-August. In other words, year 1 represents a comparison between normal (control field) vs. dry conditions (drought field), whereas year 2 compares dry (control field) vs. severely dry soil moisture conditions (drought field).

The actual soil moisture content was monitored during the drought treatments. In year 1, soil moisture in the drought field dropped from 11% to 6% (v/v) during the drought period. In year 2, soil moisture levels were already lower at the start of the drought period, and dropped from 8% to 7% (v/v). In year 2, around three weeks after drought relief (DOY 224), soil water contents were still low in both fields: 16% in the control field vs. 11% in the drought field (v/v).

Correlations between ground-measured and UAV-derived canopy height data

The rising plate meter is considered a reference method to measure plant height in phenotyping studies. UAV-derived measurements could produce the same data in considerably less time, but this method has not been thoroughly studied or validated in red clover. Therefore, we evaluated whether it was possible measure canopy height through UAV-derived measurements in red clover. Pearson correlation coefficients between ground-measured (CH_rpm) and UAV-derived (CH) measurements for canopy height were high in both years and in both fields (average over the different time points $r = 0.89$, lowest value $r = 0.80$), even for observations taken up to 7 days apart (Supplementary Table S3). Therefore, we decided to continue only with the UAV-derived CH measurements, as they included many more observations.

Performance under control conditions

Before analyzing the response of this red clover panel to drought, we investigated the general performance of the accessions in the control field. CC and CH were low during the winter and at the start of the growing season. During the growing season, CC increased to nearly 100% before each cut (Figure 4). Similarly, CH increased during each growing period, and reached a

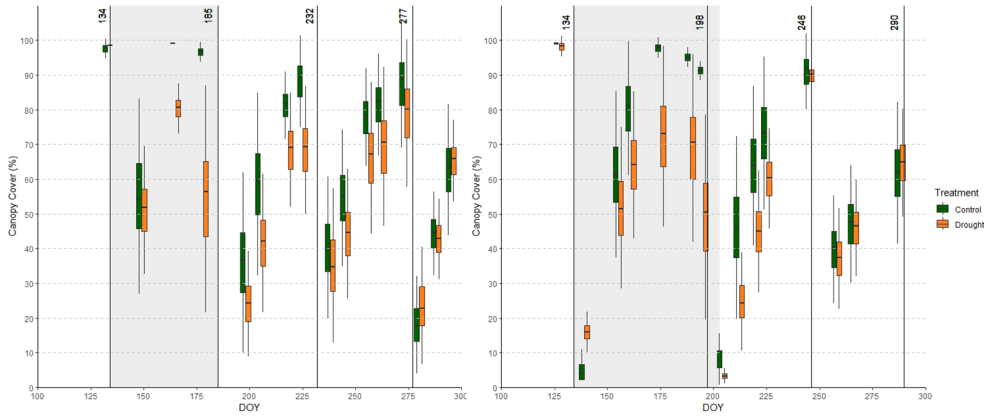


FIGURE 4
Average canopy cover (CC) over all accessions in the relevant part of the growing season of year 1 (left) and year 2 (right) in the control (green) and drought (orange) fields. Adjacent green and orange boxplots represent measurements in the control and drought fields obtained on the same day. Cuts are indicated with vertical lines and drought periods are shaded.

plateau towards the end of each growing period. In some cases a decline in CH was observed shortly before mowing (e.g. cut 1 in year 1), which can be attributed to the bending (or lodging) of the sward (Figure 5). Maximal canopy height reached 75 cm in some accessions. For CC, and CH to a lesser extent, a high degree of variability among accessions was present in the beginning of each growing period, which shrunk towards the end of the growing period. This implies that some accessions quickly re-grew after mowing, and quickly recovered to maximum CC (and CH), whereas other accessions needed more time to restore their maximum CC and CH. Broad-sense heritability values were always low just after mowing, increased during the growing period, and were maximal right before the next cut (Supplementary Table S2). This indicates that measurements at the end of each growing period represent best the genetic variation present in this collection. Growth patterns differed little between year 1 and 2. CC before the 4th cut was lower in year 2 than in year 1 (62% vs. 88% for CC_288 in year 2 and CC_273 in year 1, respectively). On average, CH before each cut was lower by approximately 10 cm in year 2.

Drought stress responses

Accession averages for all observations under control and drought conditions, together with the Yr indices are given in the Supplementary Materials. In both years, drought substantially inhibited plant growth, which became observable already two weeks after the initiation of the treatment. Drought-treated plots produced a less dense and shorter canopy than control plots. Leaf wilting was only observed by the end of the drought treatment, and only to a limited extent (personal observation). Drought effects were especially evident in year 1.

Drought caused reductions in CC: 19% and 44% at CC_165 and CC_178 in year 1, and 26% and 46% at CC_175 and CC_195 in year 2, respectively (Figure 4). In addition, we observed a

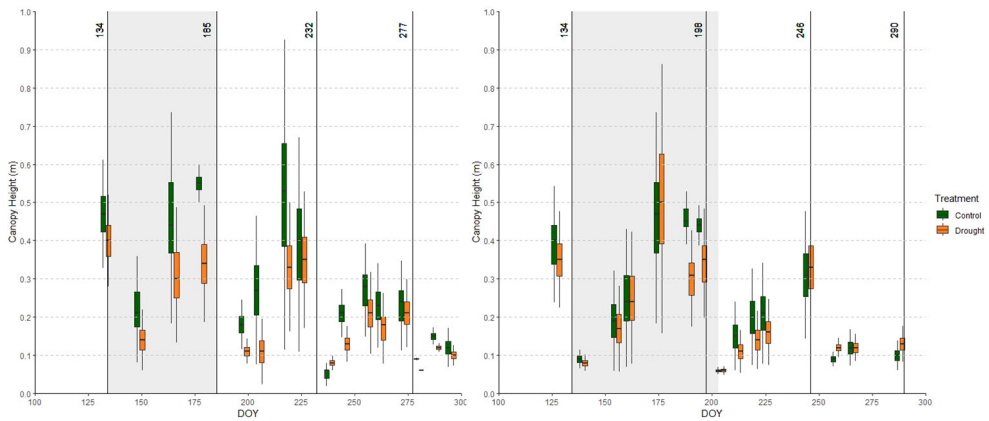


FIGURE 5
Average canopy height (CH) over all accessions in the relevant part of the growing season in year 1 (left) and year 2 (right) in the relevant part of the growing the control (green) and drought (orange) fields. Adjacent green and orange boxplots indicate measurements in the control and drought fields obtained on the same day. Cuts are indicated with vertical lines and drought treatments are shaded.

10.3389/fpls.2023.1304411

negative effect of drought on CC in the growing period following the treatment (between the 2nd and 3rd cut) in both years. In year 1, drought also tended to affect CC in the growing period between the 3rd and 4th cut, yet this effect had disappeared in year 2. The Yr indices for CC (Table 2) confirm this trend: in both years, the relative performance of the drought field gradually decreased as the drought period progressed, and remained lower during the recovery phase.

The effects of drought on CH are shown in Figure 5. Before the drought treatment in year 1, the control field slightly outperformed

TABLE 2 Relative performance indices (Yr) for CC, CH and CWSI, averaged over all accessions, for different observation moments in the growing seasons of year 1 and year 2.

Year	DOY	Average Yr index		
		Yr_CC	Yr_CH	Yr_CWSI
2019	133	-0.009***	0.148***	
2019	149	0.057***	0.351***	
2019	165	0.178***	0.311***	
2019	178	0.438***	0.373***	
2019	198	0.303***	0.351***	
2019	205	0.280***	0.593***	
2019	206			-0.735***
2019	218	0.155***	0.349***	
2019	225	0.221***	0.069***	
2019	238	0.123***	-0.586***	
2019	245	0.185***	0.394***	
2019	256	0.146***	0.231***	
2019	262	0.144***	0.227***	
2019	273	0.092***	0.106***	
2019	280	-0.328***	0.289***	
2019	288	0.031***	0.186***	
2019	295	-0.046***	0.095***	
2020	127	0.008***	0.114***	
2020	155	0.154***	0.061***	
2020	161	0.202***	-0.007	
2020	175	0.258***	-2.750***	
2020	189	0.275***	0.332***	
2020	194			-1.280***
2020	195	0.462***	0.225***	
2020	204	0.568***	0.004	
2020	212	0.454***	0.269***	
2020	220	0.291***	0.305***	
2020	225	0.174***	0.206***	
2020	245	0.013***	-0.043***	
2020	258	0.059***	-2.750***	
2020	266	0.017**	-0.005	
2020	288	-0.054***	-0.301***	

DOY, day-of-year; CC, canopy cover; CH, canopy height; CWSI, crop water stress index; Yr, relative performance index; SD, standard deviation. Mowing events are indicated with dashed lines and scissors. Drought periods are shaded. Asterisks indicate if Yr values differ statistically from 0 according to Z-test: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Asterisk indicate if Yr values differ significantly from 0 according to Z-tests.

the drought field in terms of CH. This is also represented in the Yr index (Table 2), which was higher than 0 at the start of the season in year 1. Nonetheless, drought caused clear reductions in CH, especially in year 1: 32% and 37% at CH_165 and CH_178 in year 1, and 9% and 9% at CH_175 and CH_195 in year 2. Similar to CC, a legacy effect of drought was observed in the two growing periods following the drought in year 1 (less mature swards), but only in the first growing period after drought in year 2 (more mature swards). As the stress intensity in the two years was different, it is difficult to separate the effects of sward age and stress intensity.

In both years, the Yr indices for CH increased towards the end of the drought period. Similar to CC, a legacy effect was observed in the drought field: CH in the drought field remained substantially lower than in the control field until the 4th cut.

Whereas the CC data for each time point generally approached a normal distribution, CH displayed a bimodal distribution in the control field in the growing periods between cuts 1 and 2, and cuts 2 and 3, i.e. in late spring and summer. More precisely: in year 1 at CH_149, CH_165, CH_205, CH_218 and CH_225, and in year 2 at CH_139, CH_155, CH_161, CH_175, CH_212, CH_220, CH_225, and CH_245 (Supplementary Figure S2). This segregation between rapidly and slowly re-growing accessions was investigated further in our PCA.

Due to technical reasons, less thermal UAV flights could be performed than initially intended. We obtained CWSI data only on two occasions. In year 1 CWSI data were obtained 22 days after drought relief (CWSI_206), and in year 2 at the end of the drought period (CWSI_194). Average CWSI values for the control and drought field were 0.39 vs. 0.67 (CWSI_206 in year 1), and 0.16 vs. 0.37 (CWSI_194 in year 2). This shows that on the days these measurements were done there was a clear difference between control and drought fields in both years. In other words, at both occasions stress levels experienced by the plants were higher in the drought field compared to the control field.

Relations between variables

General relations among CC and CH variables were uncovered by our PCAs on the control field data. In the first step, PCAs were performed including all observations for CC and CH between the 1st and the 4th cut (i.e. the drought period and the recovery phase) (Figures 6A, B). The first two principal components explained 64.4% and 67.2% of the total variation in year 1 and year 2, respectively. Subsequently, the number of variables was reduced to 1 representative observation in the middle and 1 at the end of each growing period (Figures 6C, D). In the latter set of PCAs, the first two principal components explained 68.3% and 74.7% of the variation, respectively in year 1 and year 2. We decided to continue with the reduced set of PCAs, as this set contained sufficient variables to describe the growth responses in our panel of accessions. Generally, the relations between variables in the control field were similar, but not identical in both years. In year 1, CC measurements at different dates were correlated among themselves, and largely independent of CH measurements. In year 2, the separation between CC and CH was less distinct than in year 1 (e.g., CC_161 and CH_161 were highly correlated).

Clustering according to maturity, geographic origin, and variety type is represented in Figure 7. A clear clustering according to maturity was found: in both years, early flowering accessions generally attained higher CH, while late flowering accessions had higher CC values. This is likely due to the formation of flowering stems on relatively small plants in early accessions, while late accessions produced more leafy branches before initiating flowering. Clustering patterns for maturity and geographic origin corresponded well: as expected, accessions from Central and South Europe were mostly early flowering, accessions from North Europe were mostly late flowering, and most West European and non-European accessions were situated in-between. No clear clustering was found according to variety type (ecotype, landrace, cultivar or breeding material), which indicates that relations among the investigated variables are similar in all variety types.

A final set of PCAs was performed on the Yr indices to reveal key variables and time-points related to drought responses. The general patterns of differentiation between Yr_CC, Yr_CH, and Yr_CWSI were similar in year 1 and year 2 (Figure 8), indicating similar responses to drought stress in both trial years. In both years, Yr_CC and Yr_CH observations clustered into two rather independent groups: observations during drought (Yr_CC and Yr_CH observations on DOY 165 and 178 in year 1, and on DOY 161 and 195 in year 2), and during the recovery phase after drought (Yr_CC and Yr_CH observations on DOY 205, 225, 256 and 273 in year 1, and on DOY 220, 245, 266 and 288 in year 2). In other words, plant responses during the drought period and in the recovery phase were largely independent. In general, CC and CH observations acquired on the same day were fairly well correlated in both years (e.g., Yr_CC_256 and Yr_CH_256 in year 1), except in late spring and early summer (e.g. Yr_CC and Yr_CH on DOY 165 and 178 in year 1), probably due to the effect of flowering which causes an increase in CH but not necessarily in CC. Hence, Yr_CC and Yr_CH could be used interchangeably to phenotype drought responses in late summer and autumn. In both years, Yr_CWSI observations were inversely correlated to CC and CH measurements acquired in the same period. (e.g., CWSI_206 vs. Yr_CC_205, Yr_CH_225 in year 1, and Yr_CWSI_194 vs. Yr_CC_195, Yr_CH_195 and Yr_CC_161 in year 2).

Relation between growth patterns under control and drought conditions

When we ranked the accessions according to the number of observations for CC or CH for which they attained value in the top 25%, it became possible to identify the 50 (or slightly more in case of accessions with equal performance) accessions with best overall performance, i.e. the 'top' accessions, in the control and drought fields (Table 3). Subsequently, we assessed how large the overlap was between both groups, i.e. how many accessions ranked in the top groups of both fields. Generally, nearly half of the accessions from the top group of the drought field also ranked top the control field, and vice versa. For CC, the percentage of top accessions in the

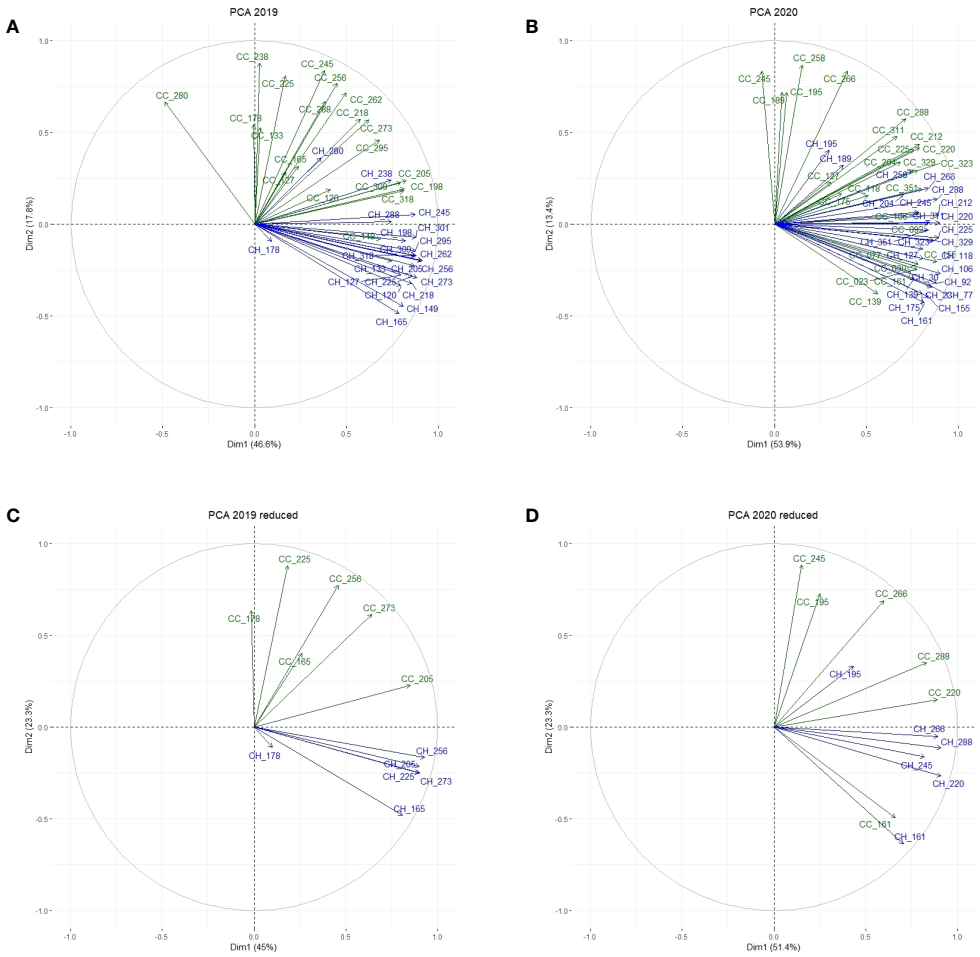


FIGURE 6
PCA on the control field displaying the relations between all observations for CC (green) and CH (blue) between the first and last cut in year 1 **(A)** and year 2 **(B)**, and between observations in the middle and at the end of each growing period between the 1st and 4th cut in year 1 **(C)** and year 2 **(D)**.

control field that ranked top in the drought field was lower in year 1 compared to year 2 (18/60 = 35% vs. 27/52 = 46%, respectively). For CH, the overlap was similar in both years (25/52 = 46% in year 1 vs. 27/50 = 54% in year 2).

Discussion

Meteorological conditions during the trial

Years 1 and 2 (2019 and 2020) were classified by the KMI (Royal Meteorological Institute, Belgium) as 'relatively dry' and 'dry' (KMI, 2020; KMI, 2021). Year 2 inhibited higher average temperatures and less precipitation, and imposed more soil moisture stress than year 1. In contrast to what we intended, the control field in year 2 also suffered above-normal soil moisture stress, in spite of the (modest) irrigations applied. In retrospect, we should have irrigated the control field more in year 2 to simulate normal soil moisture conditions. Nonetheless, the drought field suffered even more soil moisture stress, rendering Yr observations for year 2 meaningful.

Performance under control conditions

The control field showed similar patterns of CC and CH during both growing seasons (Figures 4, 5). However, in year 2 CC was lower before the 4th cut, and CH before each cut remained lower by approximately 10 cm compared to year 1. These differences are presumably due to the (unintended) soil moisture stress in the control field in year 2, which appear to have impacted plant growth. Furthermore, we observed that CC and CH behaved more independently from each other in year 1 compared to year 2 (Figure 6). This could also be due to the drier conditions in year 2, or it may be explained by the fact that the perennial plants in our plots were older in year 2 than in year 1.

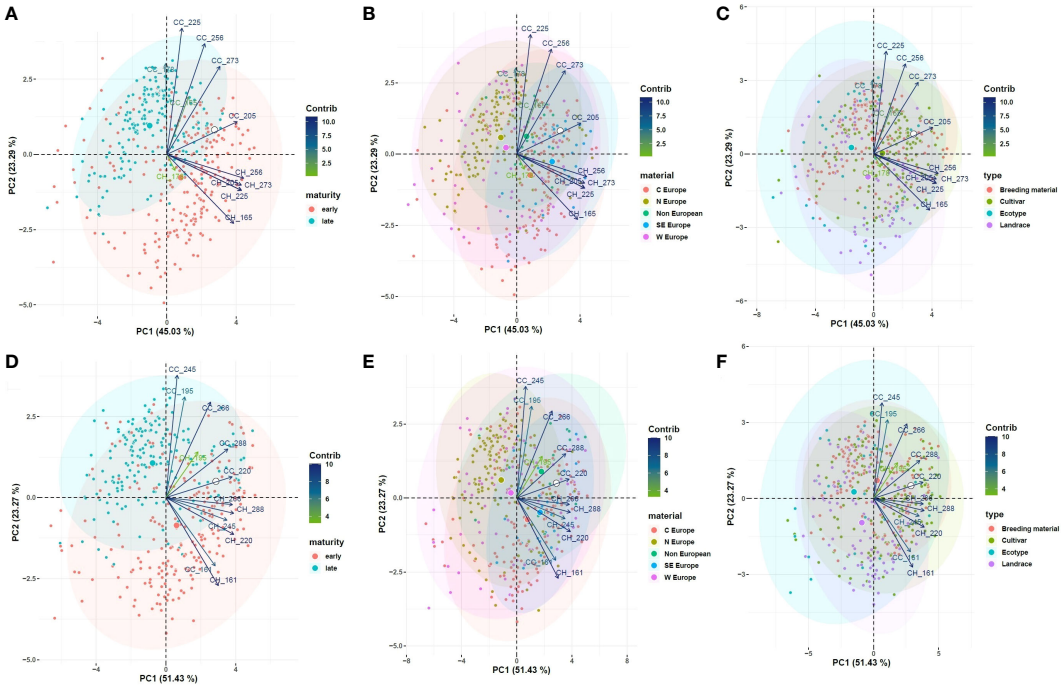


FIGURE 7
PCA on the control field for relevant observations of CC and CH, displaying the clustering according to maturity **(A)**, geographic origin **(B)** and variety type **(C)** in year 1 and in year 2 (**D–F**, respectively). Contrib: contribution (%) of each variable to the principal components.

How to phenotype red clover drought responses?

Implementation of High-Throughput Field Phenotyping (HTFP) using a UAV-based phenotyping protocol using RGB and thermal sensors was extremely advantageous in a study where plants need to be phenotyped in a dynamic way over the course of an entire growing season, as in the present work. Getting a similar dataset using destructive measurements of biomass yield or manual measurements of CC or CH would have been extremely time-consuming. Although we did not validate the CC observations using an destructive method (cutting and weighing), we have shown that UAV-derived CH observations were highly correlated with manual measurements (CH_rpm) using a rising plate meter. Similar findings were obtained in forage grasses by Borra-Serrano et al. (2019). This suggests that our UAV protocol to measure CH can be applied in a reliable way in red clover.

All variables studied – CC, CH and CWSI – were clearly affected by drought. Reductions in CC and CH presented themselves two weeks after the onset of the drought treatment, and remained

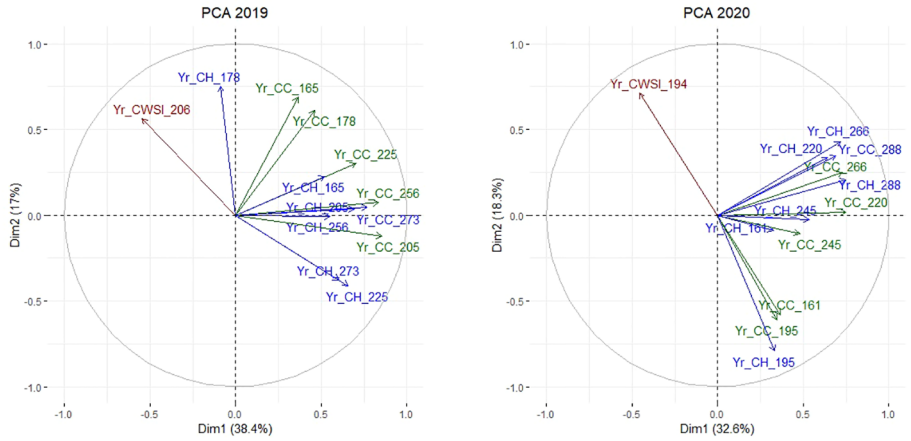


FIGURE 8
PCA on the Yr indices in year 1 (left) and year 2 (right) displaying relations between CC (green) and CH (blue) in the middle and at the end of each growing period, and CWSI (red).

TABLE 3 Accessions with overall best performance in terms of CC and CH in the control field, in the drought field, and in both fields for both years.

Year	Variable	# Accessions classified as top ranking in		
		Control field	Drought field	Both fields
2019	CC	60	51	18
2020	CC	52	59	27
2019	CH	52	54	25
2020	CH	50	50	27

observable until the 4th cut in year 1, and until the 3rd cut in year 2. The PCA revealed that CC and CH were largely independent variables that may describe different genetic variation.

During the drought period, relative reductions in CH were smaller than for CC. A first explanation is that CH at the end of a growth period can be affected by lodging, which can 'conceal' actual differences between control and drought-treated plants. Presumable lodging effects were observed in the control field in year 1 before the 2nd (CH_178) and 3rd cut (CH_225), and in year 2 before the 2nd cut (CH_195). In the drought field, on the contrary, no lodging was observed and CH continued to slowly increase towards the end of each growth period. Secondly, flowering can interfere with CH measurements. Numerous accessions flowered towards the end of the drought periods. In the control field, plants formed multiple flowering stems that were branched and rich in leaves and flower heads (personal observation – data not shown). In contrast, drought-treated plants generally formed fewer, more slender, but often reasonably high flowering stems. When measuring CH manually, it is possible to 'straighten' plants and bypass the effects of lodging and/or flowering. Although it is possible to flatten/smooth UAV-derived CH data and omit single high stems, our data displayed moderate reductions in CH in the drought-treated plants, while in fact the canopy density was remarkably lower than in the control field.

In addition, we have shown that measurements of canopy temperature using a thermal sensor mounted on a drone are very useful to determine the response of red clover plots to drought. A prerequisite is that thermal images are obtained on a warm and sunny day around solar noon, which is not always straightforward (Maes and Steppe, 2012). Due to technical failures and practical limitations, we could only obtain reliable canopy temperature data from two UAV flights. Nonetheless, CWSI values are a highly relevant variable to phenotype drought responses: as they reflect transpiration rates, they present a complementary source of information for selection. In year 1, CWSI data obtained after drought relief indicated that the drought-treated plants had not yet recovered from drought stress at that time and that their physiological performance was yet affected by the drought treatment. In year 2, CWSI data obtained at the end of the drought period indicated clearly higher stress levels in drought-treated vs/control plots. Both observations are in agreement with our observations for CC and CH. Congruently, previous work in forage grasses also found good correlations between CWSI values and breeders scores for plant vigor at the end of the drought period (De Swaef et al., 2021).

In conclusion, CC and CWSI are highly suitable variables to phenotype drought responses in red clover. Although CH is prone to bias, it can be used after correction for lodging and/or flowering. The most suitable time points to assess the response to drought is towards the end of the drought period, and/or during the recovery phase 2 to 3 weeks after drought relief, as reductions in CC and CH were maximal in these periods. Breeders aiming to phenotype drought responses in red clover can reduce time and resources by limiting the number of observations to these time points.

Do we need drought trials to identify drought-tolerant plants?

When studying the 'top' group of accessions in the drought field, roughly half of them also ranked in the top group in the control field. This was true for CC and CH, but more pronounced for CH. In year 1, this is a remarkably finding, given the contrasting soil moisture stress levels in the control and drought fields. In year 2, a better relation can be expected, as the control field also imposed soil moisture stress, although less intense than the drought field. This finding implies that, when breeders select in normal soil moisture conditions, roughly half of the material that they select will also perform well under drought stress. In other words, even without applying drought conditions, drought responses will improve (to a certain extent), merely by selecting well-performing plant material.

Associations between drought responses and accession characteristics

When a diverse panel of accessions is compared for a stress factor in a field trial, it is difficult to separate the stress response from confounding factors such as the accessions' general adaptation, the variety type or the maturity type. A first factor that may interfere with stress responses is the accessions' general adaptation, i.e. adaptation to the climate, soil type, mowing regime, local disease pressure, and other factors. Poorly adapted accessions may have suffered more background stress in our trial, which could interfere with their drought responses. However, our PCAs did not clearly show different responses in terms of CC and CH in material from regions different than West Europe (Figure 7).

A second factor is the variety type: ecotypes, landraces, cultivars and breeding material. Our PCA revealed no clustering of Yr indices according to the variety type, suggesting a similar response to drought in all groups. Ecotypes generally have poorly productive, short, and early-flowering phenotypes, while landraces, cultivars and breeding material develop taller, bushier and more productive plants (Taylor and Quesenberry, 1996). For breeders aiming to improve adaptation to drought in red clover, it is important not only to observe relative reductions in productivity after drought treatment, but also to monitor the absolute

productivity of their plant material. Annicchiarico and Pagnotta (2012) compared various Italian red clover accessions for drought responses in a Mediterranean climate. Natural populations that evolved in regions with severe summer drought showed increased yield and persistence, but had no yield disadvantage relative to the best-performing landrace or cultivar.

A third interfering factor is the maturity type, which determines the capacity to re-grow after mowing. We observed bimodal distributions in CH in the control field between cuts 1 and 2 in late spring, and between cuts 2 and 3 in summer (Supplementary Figure S2), which coincided with the accessions' maturity. After mowing, the rapidly re-growing and early flowering 'double-cut types' quickly formed flowering stems, whereas the 'single-cut types', most of them of Nordic origin, displayed slower regrowth and later flowering, with lower CH values as a consequence. This bimodal pattern was less apparent in the drought field, probably due to the effects of drought. Our PCA revealed that Yr_CC and Yr_CH observations at the end of the drought period acted largely independently from observations during the recovery phase. Simply put, accessions that reduced their growth during the drought period, did not necessarily reduce growth during the recovery phase, and vice versa. Previous studies in red clover have proposed two strategies: 'drought tolerance' and 'drought survival', which seem to coincide largely with the maturity type. Loucks et al. (2018) observed that double-cut red clover plants maintained growth longer during drought, but showed earlier and higher mortality than single-cut plants. In other words, double-cut types express more drought tolerance, while single-cut types focus more on survival (Loucks et al., 2018). As our PCA indicated, both strategies appear to be largely independent mechanisms in red clover, as proposed by Aslam et al. (2015). A similar pattern exists in alfalfa. Non autumn-dormant alfalfa accessions exhibit high drought tolerance, as they maintain productivity during drought, but display more mortality and slower recovery after severe drought stress (Pembleton and Sathish, 2014). Autumn-dormant alfalfa accessions show high drought recovery: they cease growing when drought sets in, but display low mortality and a quick re-growth after drought relief (Pembleton and Sathish, 2014). While under well-watered conditions non-dormant types often outperform dormant types, the opposite occurs under moderate to severe drought stress (Pembleton and Sathish, 2014). Our results also confirm the independent strategies and their possible association with the maturity type in red clover. However, it is difficult to make statements on plant mortality, as the number of plants per plot was not monitored in our trial.

Consequences of a two-year experiment

Drought was imposed to the perennial plants in two subsequent years, and year 2 exhibited a higher soil moisture stress (CWD) than year 1. As the stress intensity in the two years was different, it is difficult to separate the effects of sward age and stress intensity. On the one hand, the soil moisture stress in the control field in year 2 may have concealed differences between the control and drought treatments. This reasoning may explain (1) why reductions in CH were larger in year 1 compared to year 2 (37% and 9% at the end of the drought period, respectively), and (2) why the legacy effect of drought on canopy cover and canopy height appeared larger and lasted longer in year 1 compared to year 2. Alternatively, plants in year 2 were older, and have had more time to develop deep roots, enabling them to access deep soil water. In the drought field, the drought in year 1 may even have triggered plants to increase their rooting depth, which undoubtedly helped them through the drought period in year 2. Especially well-adapted accessions that withstood the drought stress in year 1 well, would have had this opportunity. Additionally, it is likely that some selection for drought adaptation has occurred in year 1: the most drought-sensitive plants may have perished, allowing more drought-tolerant neighboring plants in the same plot to fill the gaps. This may explain why even the 'objective' index of drought stress (Yr_CWSI) was less pronounced in year 2 than in year 1.

Conclusions and future perspectives

In the present paper, we characterized the responses to drought stress in a diverse panel of red clover accessions, and we identified promising accessions that could be used as source for breeding. We validated the use of UAV-derived measurements for CH in phenotyping drought responses. We observed largely similar responses to drought stress in the first and second production year, and we found evidence for two independent strategies to cope with drought stress, drought tolerance and drought recovery, which largely coincide with the maturity type. Finally, we pinpointed variables and time-points that are helpful to breeders aiming to create more drought-resilient red clover varieties. We further found that a large proportion of the accessions able to perform well under well-watered conditions were also the ones which were less affected by drought. However, it remains to be investigated which physiological mechanisms contribute to improved drought responses in red clover. Analyzing the multispectral data available for our trial could reveal additional characteristics such as the leaf density of the canopy, which could help to uncover physiological mechanisms behind drought responses. Furthermore, understanding the genetic basis of drought responses may provide additional insights and could reveal candidate genes associated with adaptation to drought stress. A next step could be the development of molecular markers, which would allow breeders to further optimize their breeding methods for adaptation to drought in red clover. In future research, we plan to perform GWAS for drought responses on the EUCLEG red clover collection, using the information generated in this study.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

TV: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Supervision,

Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. RD: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. HM: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Visualization, Writing – review & editing. IB-S: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing. PL: Conceptualization, Data curation, Formal Analysis, Methodology, Writing – review & editing. TD: Conceptualization, Data curation, Formal Analysis, Methodology, Visualization, Writing – review & editing. IR-R: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work has received funding from the European Union's Horizon 2020 Program for Research and Innovation under grant agreement no. 727312 (project: "EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency"); http://www.eucleg.eu.

Acknowledgments

We sincerely thank the technical teams at the ILVO Plant Sciences Unit for establishing and maintaining the field trial and assisting in the data collection. We thank the team of Emphasis-Belgium @ILVO (granted IRI project, FWO) for the technical support related to High-Throughput Field Phenotyping. We thank Paul Quataert for his guidance and assistance with the statistical analysis.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1304411/full#supplementary-material

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., and Marshall, A. H. (2015). Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* 34, 327–380. doi: 10.1080/07352689.2014.898462

Annicchiarico, P., and Pagnotta, M. A. (2012). Agronomic value and adaptation across climatically contrasting environments of Italian red clover landraces and natural populations. *Grass Forage Sci.* 67, 597–605. doi: 10.1111/j.1365-2494.2012.00887.x

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

Aslam, M., Maqbool, M. A., and Cengiz, R. (2015). "Mechanisms of drought resistance," in *Drought Stress in Maize (Zea mays L.)* (Springer, Cham: SpringerBriefs in Agriculture). doi: 10.1007/978-3-319-25442-5

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67 (1), 1–48. doi: 10.18637/jss.v067.i01

Bellague, D., M'Hammedi-Bouzina, M., and Abdelguerfi, A. (2016). Measuring the performance of perennial alfalfa with drought tolerance indices. *Chil. J. Agric. Res.* 76 (3), 273–284. doi: 10.4067/S0718-58392016000300003

Borra-Serrano, I., De Swaef, T., Muylle, H., Nuyttens, D., Vangeyte, J., Mertens, K., et al. (2019). Canopy height measurements and non-destructive biomass estimation of *Lolium perenne* swards using UAV imagery. *Grass Forage Sci.* 74, 356–369. doi: 10.1111/gfs.12439

Borra-Serrano, I., De Swaef, T., Quataert, P., Aper, J., Saleem, A., Saeys, W., et al. (2020). Closing the phenotyping gap: high resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sens.* 12 (10), 1644. doi: 10.3390/rs12101644

Boucher, K. (2016). *Breeding for improved drought tolerance in red clover (Trifolium pratense L.)* (Ontario, Canada: Master Degree Thesis. University of Guelph), 58.

Dellar, M., Topp, C. F. E., Banos, G., and Wall, E. (2018). A meta-analysis on the effects of climate change on the yield and quality of European pastures. *Agric. Ecosyst. Environ.* 265, 413–420. doi: 10.1016/j.agee.2018.06.029

De Swaef, T., Maes, W. H., Aper, J., Baert, J., Cougnon, M., Reheul, D., et al. (2021). Applying RGB- and thermal-based vegetation indices from UAVs for high-throughput field phenotyping of drought tolerance in forage grasses. *Remote Sens.* 13, 147. doi: 10.3390/rs13010147

Fang, Y., and Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cell. Mol. Life Sci.* 72, 673–689. doi: 10.1007/s00018-014-1767-0

Frey, L. A., Vleugels, T., Ruttink, T., Schubiger, F. X., Pégard, M., Skøt, L., et al. (2022). Phenotypic variation and quantitative trait loci for resistance to southern anthracnose and clover rot in red clover. *TAG* 135, 4337–4349. doi: 10.1007/s00122-022-04223-8

Hanson, A., Xu, L., Boe, A., Johnson, P. S., Gates, R. N., and Wu, Y. (2015). Identification and characterization of drought-tolerant alfalfa (*Medicago sativa* subsp. falcata) germplasm. *Proc. S. Dak. Acad. Sci.* 94, 263–272.

Hejduk, S., and Knot, P. (2010). Effect of provenance and ploidy of red clover varieties on productivity, persistence and growth pattern in mixture with grasses. *Plant Soil Environ.* 56 (3), 111–119. doi: 10.17221/164/2009-PSE

IPCC (2023). "Climate change 2023: synthesis report," in *Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Geneva, Switzerland: IPCC). doi: 10.59327/IPCC/AR6-9789291691647.001

Kassambara, A., and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Available at: https://CRAN.R-project.org/package=factoextra.

Kjærgaard, T. (2003). A plant that changed the world: the rise and fall of clover 1000-2000. *Landsc. Res.* 28 (1), 41–49. doi: 10.1080/01426390306531

Koninklijk Meteorologisch Instituut (KMI) (2020). *Klimatologisch jaaroverzicht jaar 2019* (Ukkel, Belgium: KMI-IRM), 12.

Koninklijk Meteorologisch Instituut (KMI) (2021). *Klimatologisch jaaroverzicht jaar 2020* (Ukkel, Belgium: KMI-IRM), 12.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25 (1), 1–18. doi: 10.18637/jss.v025.i01

Loucks, C. E. S., Deen, W., Gaudin, A. C. M., Earl, H. J., Bowley, S. R., and Martin, R. C. (2018). Genotypic differences in red clover (*Trifolium pratense* L.) response under severe water deficit. *Plant Soil* 425, 401–414. doi: 10.1007/s11104-018-3594-9

Maes, W. H., and Steppe, K. (2012). Estimating evapotranspiration and drought stress with ground-based thermal remote sensing in agriculture: a review. *J. Exp. Bot.* 63 (13), 4671–4712. doi: 10.1093/jxb/ers165

Parra, A., Ramírez, D. A., Resco, V., Velasco, A., and Moreno, J. M. (2012). Modifying rainfall patterns in a Mediterranean shrubland: System design, plant responses, and experimental burning. *Int. J. Biometeorol.* 56, 1033–1043. doi: 10.1007/s00484-011-0517-3

Pembleton, K. G., and Sathish, P. (2014). Giving drought the cold shoulder: a relationship between drought tolerance and fall dormancy in an agriculturally important crop. *AoB Plants* 6, plu012. doi: 10.1093/aobpla/plu012

Pembleton, K. G., Volenec, J. J., Rawnsley, R. P., and Donaghy, D. J. (2010). Partitioning of taproot constituents and crown bud development are affected by water deficit in regrowing alfalfa (*Medicago sativa* L.). *Crop Sci.* 50, 989–999. doi: 10.2135/cropsci2009.03.0140

Quan, W., Liu, X., Wang, H., and Chan, Z. (2016). Comparative physiological and transcriptional analyses of two contrasting drought tolerant alfalfa varieties. *Front. Plant Sci.* 6, 1256. doi: 10.3389/fpls.2015.01256

R Core Team (2022). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).

RStudio Team (2022). *RStudio: Integrated Development for R* (PBC, Boston, MA, USA: RStudio).

Saleem, A., Aper, J., Muylle, H., Borra-Serrano, I., Quataert, P., Lootens, P., et al. (2022). Response of a diverse European soybean collection to "short duration" and "long duration" drought stress. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.818766

Surault, F., Roy, E., Mahieu, S., Combes, D., Ghesquière, M., and Julier, B. (2021). *UAV to measure canopy height and plot biomass in a lucerne variety trial* (Germany: Symposium of the European Grassland Federation).

Taylor, N. L., and Quesenberry, K. H. (1996). *Red Clover Science* (Dordrecht, Netherlands: Kluwer Academic Publishers), 228.

Vaseva, I., Akiscan, Y., Demirevska, K., Anders, I., and Feller, U. (2011). Drought stress tolerance of red and white clover–comparative analysis of some chaperonins and dehydrins. *Sci. Hortic.* 130, 653–659. doi: 10.1016/j.scienta.2011.08.021

Yates, S. A., Swain, M. T., Hegarty, M. J., Chernukin, I., Lowe, M., Allison, G. G., et al. (2014). *De novo* assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. *BMC Genom.* 14, 453. doi: 10.1186/1471-2164-15-453

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fpls.2024.1328690

Check for updates

OPEN ACCESS

EDITED BY
Åshild Ergon,
Norwegian University of Life Sciences, Norway

REVIEWED BY
Wolfgang Link,
University of Göttingen, Germany
Susanne Windju,
Graminor, Norway

*CORRESPONDENCE
Natalia Gutierrez
natalia.gutierrez.leiva@juntadeandalucia.es

RECEIVED 27 October 2023
ACCEPTED 26 February 2024
PUBLISHED 13 March 2024

CITATION
Gutierrez N, Pégard M, Solis I, Sokolovic D, Lloyd D, Howarth C and Torres AM (2024) Genome-wide association study for yield-related traits in faba bean (*Vicia faba* L.). *Front. Plant Sci.* 15:1328690. doi: 10.3389/fpls.2024.1328690

COPYRIGHT
© 2024 Gutierrez, Pégard, Solis, Sokolovic, Lloyd, Howarth and Torres. 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.

Genome-wide association study for yield-related traits in faba bean (*Vicia faba* L.)

Natalia Gutierrez[1]*, Marie Pégard[2], Ignacio Solis[3], Dejan Sokolovic[4], David Lloyd[5], Catherine Howarth[5] and Ana M. Torres[1]

[1]Área de Mejora Vegetal y Biotecnología, IFAPA Centro "Alameda del Obispo", Córdoba, Spain, [2]INRA, Centre Nouvelle-Aquitaine-Poitiers, UR4 (URP3F), Lusignan, France, [3]Agrovegetal S.A., Sevilla, Spain, [4]Institute for Forage Crops, Kruševac, Serbia, [5]Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Aberystwyth, United Kingdom

Yield is the most complex trait to improve crop production, and identifying the genetic determinants for high yield is a major issue in breeding new varieties. In faba bean (*Vicia faba* L.), quantitative trait loci (QTLs) have previously been detected in studies of biparental mapping populations, but the genes controlling the main trait components remain largely unknown. In this study, we investigated for the first time the genetic control of six faba bean yield-related traits: shattering (SH), pods per plant (PP), seeds per pod (SP), seeds per plant (SPL), 100-seed weight (HSW), and plot yield (PY), using a genome-wide association study (GWAS) on a worldwide collection of 352 homozygous faba bean accessions with the aim of identifying markers associated with them. Phenotyping was carried out in field trials at three locations (Spain, United Kingdom, and Serbia) over 2 years. The faba bean panel was genotyped with the Affymetrix faba bean SNP-chip yielding 22,867 SNP markers. The GWAS analysis identified 112 marker–trait associations (MTAs) in 97 candidate genes, distributed over the six faba bean chromosomes. Eight MTAs were detected in at least two environments, and five were associated with multiple traits. The next step will be to validate these candidates in different genetic backgrounds to provide resources for marker-assisted breeding of faba bean yield.

KEYWORDS

yield, heritability, population structure, linkage disequilibrium, GWAS, MTA markers, candidate genes, faba bean

Introduction

Faba bean (*Vicia faba* L.) has been cultivated since the beginning of agriculture (Cubero, 1973) and, at present, is the fourth most widely grown cool season legume after chickpea, pea, and lentil (FAOSTAT, 2020). Its nutrient-rich seeds are used as a protein source for both human consumption and as a feed grain. It is well adapted to a wide range

of climatic areas and has one of the highest yield potentials (Cernay et al., 2016) and protein content (average ~29%) of the grain legumes (Warsame et al., 2018).

Legumes play a crucial role in ensuring worldwide food security, promoting agroecosystem resilience, and facilitating sustainable agriculture. The key benefits of legumes include the biological fixation of atmospheric nitrogen as a result of a symbiotic relationship with soil bacteria known as rhizobia. Among grain legumes, faba bean is the most efficient nitrogen fixer with values ranging between 50 and 200 kg N ha^{-1} (Schwenke et al., 1998; Carranca et al., 1999; Lopez-Bellido et al., 2006). This unique ability reduces the dependence of farmers on extensive use of fertilizers and protects soil and groundwater quality. Furthermore, implemented in crop rotations, faba bean functions as a break crop, decreasing the occurrence of weeds, pests and diseases and enhancing soil physical conditions. This leads to a higher yield of subsequent crop(s) while reducing the amount of fertilizers and biocides required (Köpke and Nemecek, 2010; Cernay et al., 2016).

Despite these nutritional and multiple environmental services, faba bean represents a minor part of European agricultural systems that have developed a high dependence on imports of grain legumes. This is partly due to a lack of economically competitive grain legume crops that can match cereals in terms of farmer profits. For the faba bean to become an economically more attractive crop and thus increase the area under cultivation, major advances in both yield and yield stability need to be achieved. Currently, faba bean yield is considered to be highly unreliable due to a significant level of genotype × environment interaction (Cernay et al., 2016). A greater understanding of the main biological and environmental factors affecting plant growth and the identification of the main yield components are critical to improve adaptation and yield in this crop. While biotic and abiotic stresses have received considerable attention in faba bean (Adhikari et al., 2021; Khazaei et al., 2021), plant architecture and yield-related traits are still poorly understood in this crop. Faba bean has lagged behind cereals in the genetic improvement of yield, due to a combination of less investment and limited genomic resources available (Adhikari et al., 2021). Recent advancements in genomic tools, such as a reference genome sequence (Jayakodi et al., 2023), enable a genomics-based breeding platform for assisting conventional breeding and accelerate the release of high-yielding and stable faba bean cultivars. Quantitative trait loci (QTLs) associated with yield-related traits have been identified previously, but the results were limited due to the use of anonymous markers and low-density maps (Ramsay et al., 1995; Avila et al., 2005). Stable QTLs for three reproductive traits were identified on chromosomes (Chr.) II and V (pod length) and VI (number of ovules per pod and number of seeds per pod), by evaluating the recombinant inbred line (RIL) population Vf6 × Vf27 (Cruz-Izquierdo et al., 2012) during two seasons. More recently, 11 plant architecture and yield-related traits were recorded in the same population across four seasons (Ávila et al., 2017), confirming the QTLs for pod length, number of ovules per pod, and number of seeds per pod and further identifying stable QTLs for hundred-seed weight on Chr. V and VI. Despite these valuable advances, the genetic architecture and determinants of faba bean yield remain uncertain.

The resolution and accuracy with which QTL mapping can identify the direct causal gene(s) controlling a trait are limited by the large confidence intervals and the relatively low number of recombination events existing in biparental mapping populations (Beji et al., 2020). With the recent development of powerful faba bean SNP array platforms (Khazaei et al., 2021), genome-wide association studies (GWAS) are now an additional tool to dissect polygenic traits by utilizing the genetic diversity and historical recombination events present in wide germplasm collections. GWAS aims at identifying markers strongly associated with quantitative traits by using the linkage disequilibrium (LD) between candidate genes and markers. Compared with family-based QTL mapping, GWAS significantly increases mapping resolution and enables minor effect genes to be detected (Abdurakhmonov et al., 2008). In recent years, GWAS studies have been reported in a range of legume crops such as soybean (Hwang et al., 2014), pigeon pea (Varshney et al., 2017), common bean (Raggi et al., 2019), chickpea (Varshney et al., 2019), red clover (Zanotto et al., 2023), alfalfa (Pégard et al., 2023), and the model legume *Medicago truncatula* (Bonhomme et al., 2014). In faba bean, only a few GWAS studies have been reported so far which have identified candidate genes associated with frost tolerance (Sallam et al., 2016); resistance to *Ascochyta fabae* (Faridi et al., 2021); tolerance to herbicides (Abou-Khater et al., 2022); drought, heat, and freezing tolerance (Ali et al., 2016; Maalouf et al., 2022; Gutiérrez et al., 2023); agronomic traits (Skovbjerg et al., 2023); and seed size (Jayakodi et al., 2023).

The aim of this work was to identify for the first time genomic regions controlling yield-related traits using a panel of 400 faba bean lines grown in a range of environments and using the recently available 60K Axiom Vfaba_v2 SNP array (O'Sullivan et al., 2019; Khazaei et al., 2021). The traits analyzed were pod shattering (SH), seeds per pod (SP), pods per plant (PP), seeds per plant (SPL), 100 seed weight (HSW), and plot yield (PY). The results obtained contribute to the discovery of new genomic regions associated with grain yield characters and the identification of candidate genes associated with the SNPs to accelerate future molecular marker-assisted breeding in this crop.

Materials and methods

Plant material

The faba bean EUCLEG collection consists of 400 accessions from Africa (25 accessions); North, Central, and South America (10); Asia (59); and Europe (185). Moreover, it also consists of 121 accessions with unknown origin. Europe, with 22 countries, is the most represented continent in the panel, followed by Asia, America, and Africa (13, 4, and 5 countries, respectively). Spain accounts for the highest number of accessions (68). The panel includes 81 breeding and advanced materials, 47 varieties, 7 parental inbred lines from different mapping populations, and 265 accessions from different germplasm banks (Supplementary Table S1). The selection was made in collaboration with public institutes including ICARDA, IFAPA, IFVCNS, INIA, and INRA; the universities of

Ghent and Göttingen; and the following gene banks: ESP004, ESP046, FRA043, SWE054, SYR002, and NordGen. Private sector contributions consisted of lines/varieties from the companies Agrovegetal, NPZ, Batlle, and Fitó. Since the faba bean is a partially allogamous species, prior to genotyping, all the lines from Spain had been selfed at IFAPA, Córdoba, in the field for at least four generations using insect-proof cages with the remaining accessions selfed for two generations.

Phenotypic data analysis

Phenotypic traits and experimental design

The faba bean collection was phenotyped in three geographic locations: 1) Escacena del Campo, Huelva (Spain; 2019 and 2020) by the company Agrovegetal; 2) Aberystwyth (United Kingdom (UK); 2019) by IBERS; and 3) Krusevac (Serbia; 2020) by IKBKS. The growing season in southern Spain was from November to June, while in the rest of the countries, it was from April to September. The location descriptors are as follows: Escacena del Campo (37°30′ N, 6°22′W), Aberystwyth (52°41′N, 4°06′W), and Krusevac (43°58′ N, 21°20′E).

In each location and year, the accessions were arranged in the field following an augmented design MAD type 2 (Lin and Poushinsky, 1983). The experimental trials included 440 plots (3 m^2 each), distributed in 22 rows and 20 columns. Plots consisted of 4 rows of 2 m length with 0.5 m row spacing and seeding distance of 10 cm (80 seeds/plot). Twenty accessions were chosen as checks and distributed in four incomplete blocks with different numbers of repetitions (5 checks repeated six times and 15 twice), and the remaining 380 accessions were unreplicated and randomly distributed within and between blocks.

The plants were measured and scored for six agronomic traits affecting the final yield response. Pods per plant (PP), seeds per pod (SP), and seeds per plant (SPL) were recorded as the mean value from 10 plants in the central rows. Plot yield (PY) was determined as the weight of the seeds from a whole plot in kilograms (kg). The seeds were used to determine 100 seed weight (HSW) in grams (g). Shattering (SH) was determined just before harvesting only at Escacena del Campo, Spain, in 2019 using a 0 to 3 scale as follows: indehiscent (0), fissured with valves slightly open along the ventral suture (1), dehiscent (DH) with non-twisting valves (2), and dehiscent with twisting valves (3). PP and SPL were not scored in the UK and the rest of the traits were recorded in all locations.

Genomic data analysis

SNP genotyping and quality control

For DNA extraction, young leaf tissue was collected from a single plant per accession. Leaf samples were frozen and stored at −80°C until used. Genomic DNA was extracted using a DNeasy Plant Mini Kit (Qiagen Ltd., UK), and DNA quality was assessed as described in Gutiérrez et al. (2023). Pure and good-quality DNA samples with an average concentration of 40 ng/µl were used for genotyping using the Vfaba_v2 Axiom SNP array with 60K SNP (O'Sullivan et al., 2019; Khazaei et al., 2021) from Affymetrix (Thermo Fisher Scientific), University of Reading, UK. After quality control, 26 accessions with poor DNA quality and 22 revealing a low number of SNPs were excluded from the analysis. SNP calling was performed on the remaining 352 samples, and SNP markers with a call rate below 97% were discarded from the final genotyping database. The SNPs were filtered to remove those with a minor allele frequency (MAF)<0.05 and >15% of heterozygotes. The missing values were imputed with the minor allele frequency (Badke et al., 2014), and it represented less than 1% of all genomic data. The allelic genotyping matrix was transformed into a numeric format (0, for the reference allele; 1, for the heterozygous allele; and 2, for the alternative allele).

Genomic relationship matrix

The genomic relationship matrix (*G*) was constructed based on VanRaden (VanRaden, 2008) (Equation 1):

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)} \quad (1)$$

where the matrix **Z** was calculated as (**M** − **P**). **M** is a matrix of minor allele counts (0, 1, and 2 for the reference, heterozygote, and alternative, respectively) with *m* columns (one for each marker) and *n* rows (one for each accession). **P** is a matrix that contains the minor allele frequency, expressed as a difference from 0.5 and multiplied by 2, such that column *i* of P is $2(p_i - 0.5)$.

Analysis of phenotypic data

For phenotypic analysis, environments were designed as follows: *Spain.2019* (Spain, season 2018/2019), *Spain.2020* (Spain, season 2019/2020), *Spain.Global* (combined data in Spain for two seasons), *UK.2019* (United Kingdom, year 2019), *Serbia.2020* (Serbia, year 2020), and *Global* (all environments combined). The check accessions were used to capture the spatial heterogeneity at the plot level, and all traits were independently adjusted for field microenvironmental heterogeneity using a mixed linear model (MLM) by the function restricted maximum likelihood (REML) with the "breedR" package (Muñoz and Sanchez, 2020). A random effect was fitted using the tensor product of two B-spline bases with a covariance structure for the random knot effects (RKE) to account for spatial variability along the rows and the columns of the field design (Cantet et al., 2005; Cappa and Cantet, 2007; Cappa et al., 2015). The genomic estimated breeding values (GEBVs) for each trait were determined with the genomic best linear unbiased prediction-based model (GBLUP) (Whittaker et al., 2000; Meuwissen et al., 2001; Cantet et al., 2005). The following basic model was considered:

$$y = \mu + Zu + Ws + \varepsilon(\text{within the environment}) \quad (2)$$

$$y = \mu + X\beta + Zu + Ws + \varepsilon(\text{across environments}) \quad (3)$$

where y is the raw phenotypes; μ the global mean; u the vector of random additive effects following the distribution $N(0, G\sigma_a^2)$, where σ_a^2 is the additive variance and G the genomic relationship matrix between accessions; s is the vector of random spatial effects

containing the parameters of the B-spline tensor product following $N(0, S\sigma_s^2)$, where σ_s^2 is the variance of the RKE for the rows and columns, while S represents the covariance structure in two dimensions; and ε is the vector of residual effects following $N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance. The design matrices Z and W were identity matrices relating the plot to the random effects. For analyses across environments (*Spain.Global* and *Global*), β (Equation 2) is a fixed effect of the year in the same location or the effect of interaction among locations and X is the design matrix relating the plot to the fixed effect. The method used to obtain the covariance structure S was (Equation 3) the following: bi-splines were anchored at a given number of knots for rows and columns, and a higher number of knots smooths out the surfaces. "*breedR*" optimized the knot numbers by an automated grid search based on the Akaike information criterion (AIC). The microenvironmental individual effect was subtracted from the observed phenotype to obtain a spatially adjusted individual phenotype used to conduct the GWA studies. A genotypic mean of the spatially adjusted phenotypes was calculated for each trait and used for the GWAS. All measurements were tested for deviations from normality by a randomized quantile–quantile (Q–Q) plot.

Heritability and correlation

Narrow sense heritability (h^2) was estimated from the variance components of each model after phenotypic adjustment. For the individual environments (*Spain.2019*, *Spain.2020*, *UK2019*, and *Serbia.2020*), the following formula was used (Equation 4):

$$h^2 = V_g/(V_g + V_{sp} + V_{res}) \quad (4)$$

(within the environment)

where V_g is the additive genetic variance component, V_{sp} is the spatial variance component, and V_{res} is the residual variance component.

To calculate the h^2 for combined environments (*Spain.Global* and *Global*), the formula used was the following (Equation 5):

$$h^2 = V_g/(V_g + V_{spn}) + V_{res}/n \quad (5)$$

(across environments)

where V_{spn} is the spatial variance component for each environment and n is the mean number of replicates (checks) for each accession per environment.

To understand the extent of the relationship between traits and after adjusting phenotypic data, a correlation coefficient analysis was performed using the Pearson method for all the traits across environments. Additionally, the genetic correlation was assessed using a multitrait model on adjusted phenotypes with the "breedR" package (Muñoz and Sanchez, 2020). For each random effect, including genetic and spatial effects, a full covariance matrix is estimated. The "cov2cor" function in the R package "stats" was used to compute the genetic correlation between traits from the additive genetic variance–covariance matrix. This information was based on the research by Calus and Veerkamp (2011). The genetic correlation among traits was extracted and compared with the phenotypic correlation. Descriptive analysis and correlations of the phenotypic data were conducted with the R 4.2.3 software (R Core Team, 2022).

Alignment of SNP markers to the *Vicia faba* reference genome

To identify the chromosomal location of SNP markers, their flanking sequences were aligned to the *Vicia faba* reference genome (Jayakodi et al., 2023) using the "map to reference" option implemented in Geneious v.7.1.9. For the genomic position of the SNP markers, the information provided by Skovbjerg et al. (2023) was used. To facilitate data analysis, the extremely large chromosome I (>3 Gbp) was divided at the centromere (position 1,574,527,093 bp) by the faba bean genome consortium to form Chr1S and Chr1L. Functional annotation was done using eggNOG-mapper v.2 with the eukaryotic database (Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). The associated genes were categorized by the Clusters of Orthologous Group (COG) and plotted using the "ggplot2" R package (Wickham, 2016).

Estimation of linkage disequilibrium

To calculate the linkage disequilibrium (LD), only SNP markers with physical position and chromosomal location within the *V. faba* genome were used. So, a genotyping matrix of 19,741 SNP markers was filtered for a MAF of 5% and a numerical imputation with the LD-kNNi method (Money et al., 2015) implemented in TASSEL v5.2.88 (Bradbury et al., 2007). LD was estimated for each chromosome and for the whole genome, by computing the squared allele frequency correlations (r^2) for each pairwise combination of markers distanced within 1 Mbp in PLINK v.1.9 (Purcell et al., 2007). LD was plotted against the genomic distance between markers in kbp, and a curve was fitted using the LOESS regression model and R. The LD decay was estimated per chromosome and the whole genome as the point where the fitted curve reached half of its maximum value.

Population structure and phylogeny

To infer the population structure of the SNP marker panel, a Bayesian-based clustering analysis was performed using fastSTRUCTURE v. 1.0 (Raj et al., 2014). fastSTRUCTURE was run with default settings and 10-fold cross-validation using K values ranging from 1 to 10. The most likely number of subpopulations (K) was identified by plotting the marginal likelihood of each model as a function of K and determining when the graph begins to plateau. The choice of K was further supported by applying a discriminant analysis of principal components (DAPC)-based procedure for clustering using the "fviz_pca" function in the "factoextra" R package (Kassambara and Mundt, 2020). The resulting admixture proportions were graphically displayed using the distruct.py script provided by fastSTRUCTURE.

A phylogenetic tree was constructed with the neighbor-joining method applying a bootstrap test with 1,000 replications, using MEGA 11 (Tamura et al., 2021). The R package "ggtree" was then used to visualize a circular phylogenetic tree (Xu et al., 2021, 2022).

Genome-wide association analysis

The GWAS analyses were performed using the multi-locus mixed model method (MLMM) (Segura et al., 2012), which

accounts for the genetic structure of the faba bean collection within the genomic relationship matrix (*G*), using the R package "mlmm.gwas" (Bonnafous et al., 2019). The MLMM method uses a stepwise mixed-model regression approach with forward inclusion of the SNPs as co-factors and a backward elimination. The variance components of the model were re-estimated at each step (maximum 10 steps) and then used to calculate *p*-values for the association of each SNP with the trait in the study. MLMM implements two model selection methods to determine the optimal mixed model from the set of stepwise models calculated: the extended Bayesian information criterion and the Bonferroni criterion. The Bonferroni method is considered the most stringent for selecting a threshold *p*-value (Kaler and Purcell, 2019) and may result in a loss of power and of true positives. For this reason, in this study, we have considered both the eBIC with a lambda value of 0.60 and the Bonferroni test (0.05 divided by the number of SNPs) as significant cutoffs. Associated markers were visualized with a *p*-value distribution (expected vs. observed on a –log10 scale) with a Manhattan plot and a Q–Q plot. The percentage of phenotypic variation explained by each QTL was obtained by subtracting the R^2 of a linear model with all the QTL as fixed effects and the genomic relationship matrix (*G*) as random effect to the R^2 of the same model but without the focused QTL.

The genomic regions (72 bp) harboring associated SNPs for each trait in different environments were represented on the *V. faba* physical map using the Pretzel platform (Keeble-Gagnère et al., 2019) (http://pulses.plantinformatics.io/mapview).

Potential candidate gene identification

The genome sequence (Jayakodi et al., 2023) of the associated SNPs was blasted against the NCBI *Medicago truncatula* reference genome (MtrunA17r5.0-ANR) to annotate the potential candidate genes underlying the causal variants. Gene locations were determined using the Genome Data Viewer (GDV).

Results

SNP calling

Genotyping of the EUCLEG collection with the Vfaba 60K Axiom array revealed a total of 34,354 SNP markers (57% of the total 59,871 SNPs present on the array) with a call rate above 97%. Following the Axiom Best Practices Genotyping Workflow, these SNPs were classified into three quality classes according to their clustering performance: polymorphic high resolution (PHR, 71%), monomorphic high resolution (MHR, 17%), and no minor homozygous (NMH, 12%). PHR refers to polymorphic SNPs exhibiting all three highly resolved clusters (two homozygous and one heterozygous), MHR to not informative SNPs with only one of the homozygous clusters, and NMH to SNPs with good resolution lacking one of the homozygous clusters. After the quality control, 48 accessions were removed for further analysis, 26 due to poor-quality DNA, and 22 for revealing a low number of SNP markers (7,656 SNPs). The average reference allele frequency was 44%, and for the alternative allele, it was 56%. The number of missing values in the genotyping matrix was low (0.89%). After quality control, the final matrix consisted of 352 accessions genotyped for 22,867 high-quality SNPs (MAF above 5% and without missing data) and was kept for further GWAS analysis.

Genomic distribution of SNP markers

Markers with known chromosomal position, present in a window of 10 Mb, were used to develop the high-density SNP-based map (Figure 1). Of the total number of SNP markers (22,867), 93% (21,271) were well-distributed across the six chromosomes after assembling against the faba bean reference genome (Jayakodi et al., 2023), and 19,741 of them (86%) were mapped to a genomic location (Table 1). The 1,596 SNP markers not assigned to

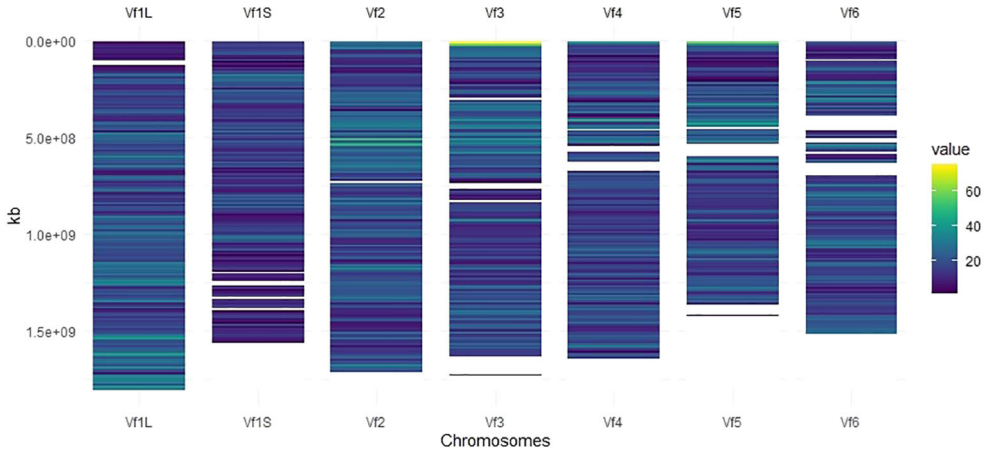


FIGURE 1
SNP density map across each faba bean chromosome representing the number of SNPs after quality control within a 10-Mbp window size. The color blue corresponds to the lowest density, while the color yellow corresponds to the highest density.

chromosomes (named as Chr0) together with the ones without genomic positions (1,530 SNPs) were, however, included in the GWAS analysis. The number of SNPs on each chromosome ranged from 3,847 on Chr1L to 2,296 on Chr1S (Table 1), and the total genetic coverage was 11.4 Gbp. The highest average density of SNPs was 20 SNPs per 10 Mbp in Chr1L and Chr2 and the lowest was 14 SNPs per 10 Mbp in Chr1S. Chr3 was the one with the maximum local density of SNPs (75 SNPs per 10 Mbp) and the maximum gap between them (95,593.1 kbp). The higher average distance between two adjacent SNPs was 738.5 kbp in Chr1S and the lowest was 506.4 kbp in Chr1L.

Estimation of linkage disequilibrium

Estimates of the linkage disequilibrium (using r^2) for each chromosome as well as for the whole genome are presented in Figure 2. LD values showed an inverse relationship with distance, and the LD decay, estimated as the distance for which r^2 decreases to half of its maximum level (0.131), was 139.2 kbp for the whole genome. Considering the chromosomes individually, the r^2 values ranged from 0.125 in Chr3 to 0.140 for Chr1L and decreased to half at reaching 139.2 kbp and 151.8 kbp, respectively (Figure 2; Supplementary Table S2).

Population structure

We analyzed the population structure of the EUCLEG collection using two approaches: fastSTRUCTURE and DAPC. The marginal likelihood of the fastSTRUCTURE output from K = 2 to K = 10 was represented, and a delta-K peak at K = 3 was determined (Figure 3A). The accessions were divided into three groups (P1, P2, and P3) with clear differences in geographic origin. Based on the results of the fastSTRUCTURE analysis, accessions with membership probabilities ≥0.50 were considered to belong to the same group. The DAPC supported the choice of K = 3 using the first two PCs. PC1 distinguished the population P1 from P2 and P3, whereas PC2 further distinguished P2 and P3 populations (Figure 3B).

Group P1 contained 148 accessions, of which 38% had an unknown geographic origin and 44% were associated with North European countries, with Finland (22), Sweden (15), and France (11) being the most highly represented. Population P2 included 177 accessions of which 46% belonged to Mediterranean countries and 26% had an unknown origin. Spain was the country with the highest number of accessions (61), followed by Egypt (9), Ethiopia (7), and Syria (6). P3 was the smallest group, with 22 accessions associated with countries from Asia, with China (15) being the predominant country, followed by Japan (1), Nepal (1), and Thailand (1), and the rest of the accessions (4) were of unknown geographic origin (Supplementary Table S1). Five of the 352 accessions were found to be admixed and were not assigned to a specific group.

According to the fastSTRUCTURE approach, the neighbor-joining tree generated with 352 faba bean accessions and 22,867 SNP markers further suggested the three main clades (Figure 3C) although populations P2 and P3 did not show such a clear division.

Phenotypic variation, heritability, and correlation

Descriptive analysis, heritability, and correlations of the phenotypic data were estimated after phenotypic adjustment. Except for SH, scored only at *Spain.2019*, all traits showed a high variability across the six environments (Supplementary Figure S1) and followed a normal distribution with a positive skewness (except for SH where the deviation of the data was toward indehiscent plants). SP showed the highest phenotypic mean in *Serbia.2020* (3.69 ± 0.50), followed by *Spain.2020* (3.22 ± 0.66). The highest phenotypic mean for PP was in *Spain.Global* [*Spain.2019* (22.70 ± 9.94) and *Spain.2020* (22.97 ± 10.13)], while SPL had a higher mean value in *Spain.2020* (61.08 ± 25.59). HSW showed the highest mean value in *Spain.2019* (74.97 ± 37.97) and *Spain.2020* (71.37 ± 32.77), while the lowest value was in *Serbia.2020* (43.52 ± 16.12). PY revealed the highest mean value in *Spain.Global* [*Spain.2019* (1.18

TABLE 1 SNP distribution and coverage per individual chromosome of 352 faba bean accessions against the *Vicia faba* reference genome.

Chromosome	SNP number	Length (bp)	SNP density[a]	No. of max. SNPs[b]	SNP distance[c]	Max. gap[d]
Chr1S	2,124	1,568,008,521	14	43	738.5	28,576.9
Chr1L	3,559	1,804,809,984	20	35	506.4	31,207.6
Chr2	3,377	1,716,744,951	20	56	508.5	22,732.4
Chr3	3,069	1,733,023,265	18	75	564.8	95,593.1
Chr4	2,754	1,645,775,113	17	48	597.4	60,679.1
Chr5	2,403	1,429,386,368	17	58	594.8	68,652.8
Chr6	2,455	1,519,787,660	16	41	618.9	86,906.2
Total	19,741	11,417,535,862				

[a]Average number of SNPs per 10 Mbp.
[b]Maximum number of SNPs found on a 10-Mbp window.
[c]Average spacing between neighboring SNPs in kbp.
[d]Maximum distance between SNPs in kbp.

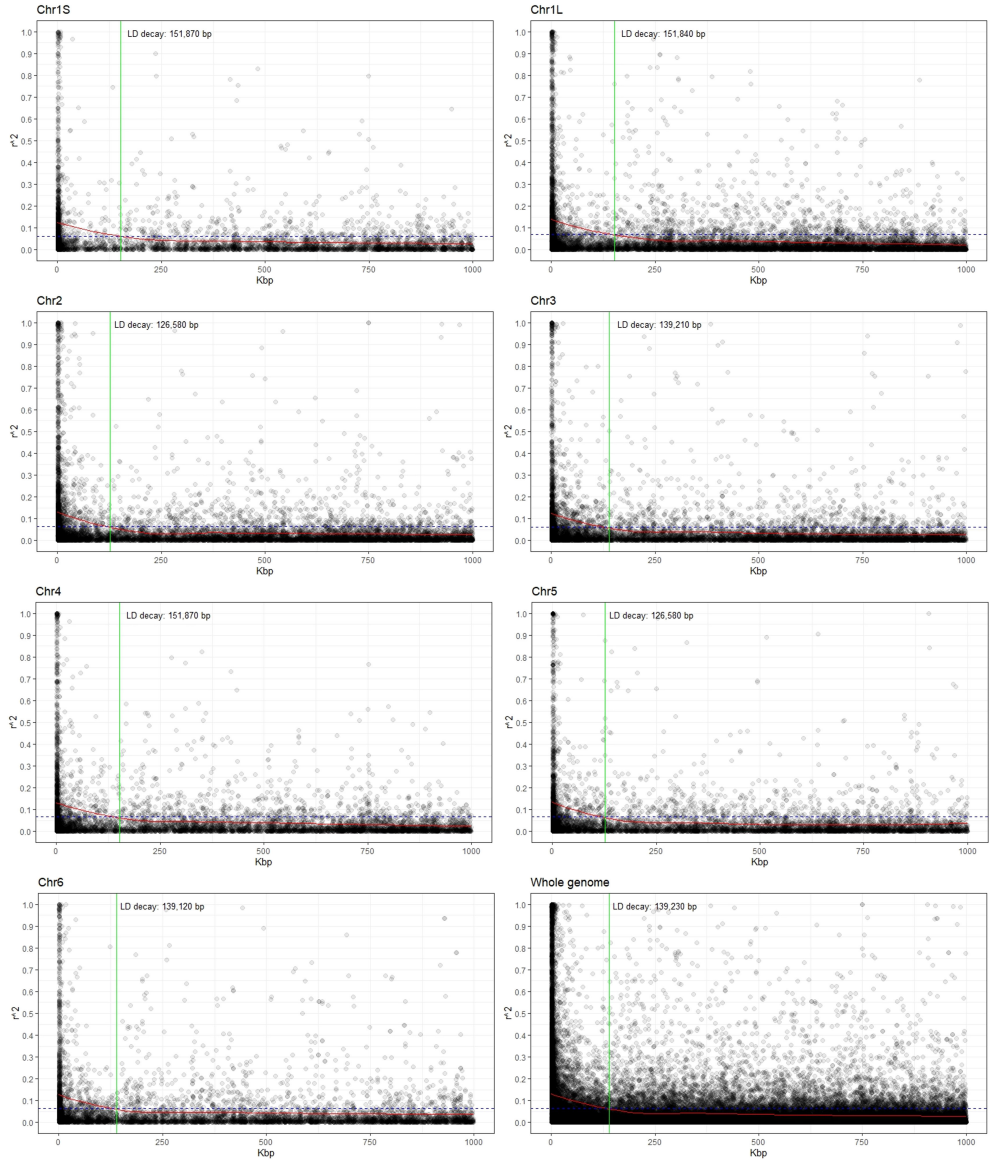


FIGURE 2
Scatter plot showing the linkage disequilibrium (LD) decay for each chromosome as well as for the whole genome. The values on the *Y*-axis represent the squared correlations of allele frequencies (r^2) between markers with a maximum distance of 1 Mbp. The *X*-axis shows the genomic distance in kbp. The intersection (green line) between the LOESS curve (red) and the threshold (half of the average value at the minimal distance; dashed blue line) indicates the extent of LD decay in base pairs (bp).

± 0.60) and *Spain.2020* (1.15 ± 0.63)], while in Serbia (0.42 ± 0.25) and the UK (0.27 ± 0.19), values were three and four times lower, respectively (Figure 4; Table 2).

The narrow sense heritability (h^2) values are shown in Table 2. On average, the highest heritability values were observed for HSW (0.9 to 0.98), whereas the lowest values were for SPL (0.65 to 0.85). *Serbia.2020* was the location with the highest mean heritability values for SP, PP, and SPL (0.93, 0.94, and 0.85, respectively), while SPL showed the lowest values in *Spain.2019* and *Spain.2020* (0.65 and 0.67). Furthermore, the proportion of the different components of the variance estimated from each environment is represented in Figure 5. Overall, the models combining multiple environments (*Global* and *Spain.Global*) displayed higher values for the residual variance component (*varE*).

The phenotypic correlation between traits and environments after phenotypic adjustment is shown in Figure 6. In general, similar performance patterns were observed in most of the traits across environments. Thus, SH evaluated only in *Spain.2019* revealed a high significance and a negative correlation with SPL, HSW, and PY (−0.22, −0.19, and −0.34, respectively). SP showed a positive correlation with SPL in all environments and a negative correlation with PP in *Spain.2020* (−0.15), *Spain.Global* (−0.15), and *Global* (−0.21). PP displayed a strong positive correlation with SPL and a negative correlation with HSW in all environments, and SPL presented a negative correlation with HSW. Finally, PY revealed a positive correlation with HSW, SPL (except in *Spain.2019*), and PP, although no correlation between PY and PP was detected in *Spain.Global* or *Global* and a negative correlation

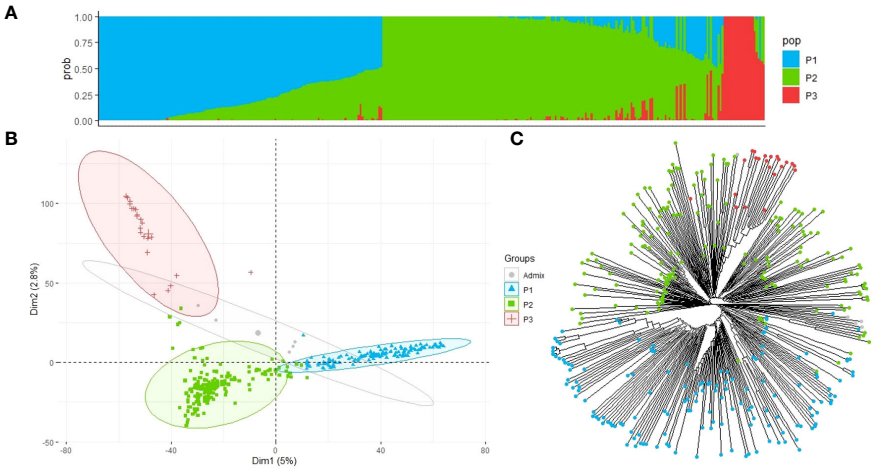


FIGURE 3
Population structure and principal component analysis of the 352 faba bean accessions. **(A)** Admixture vertical plot at $K = 3$; the vertical bars represent an individual accession, and each color corresponds to its assignment to one of the clusters based on its ancestry proportion. **(B)** Discriminant analysis of principal components (DAPC) for 352 faba bean genotypes revealing three clear groups. Each group is represented by circles with different colors and shapes. **(C)** Neighbor-joining tree of the faba bean accessions. The tips are highlighted with different colors according to the population groups.

was observed in *Spain.2019*. Only in *UK.2019*, *Serbia.2020*, and *Global* environments, PY showed a positive correlation with SP (0.35, 0.27, and 0.15, respectively).

The genetic correlation for each trait (Supplementary Table S3) was calculated using the data obtained in the different locations, with a genetic correlation close to one indicating a low G × E interaction and a value close to zero indicating a strong G × E interaction. The genetic correlation between traits was estimated between them when all the environmental variation was removed from the phenotype (named "Global"). It was noteworthy that the genetic correlation between SP in *Spain.2019* and in *Serbia.2020* (0.998) was high. Similarly, the correlation of HSW between *Serbia.2019*, *Spain.2019*, and *Spain.2020* varies between 0.928 and 0.950. However, PY was the least correlated trait between the *UK.2019* and the rest of the locations, showing the least (close to zero) correlated trait with *Spain.2020* (−0.049). PP correlations between Serbia and Spain ranged from 0.538 to 0.686, while SPL ranged from 0.571 to 0.776 between those locations. In accordance with the phenotypic correlation (Figure 6), PP and SPL showed the highest genetic correlation between them in all environments, and higher values were observed in *Serbia.2020* and *Spain.2020* (0.919 and 0.904, respectively). SH showed a negative genetic correlation

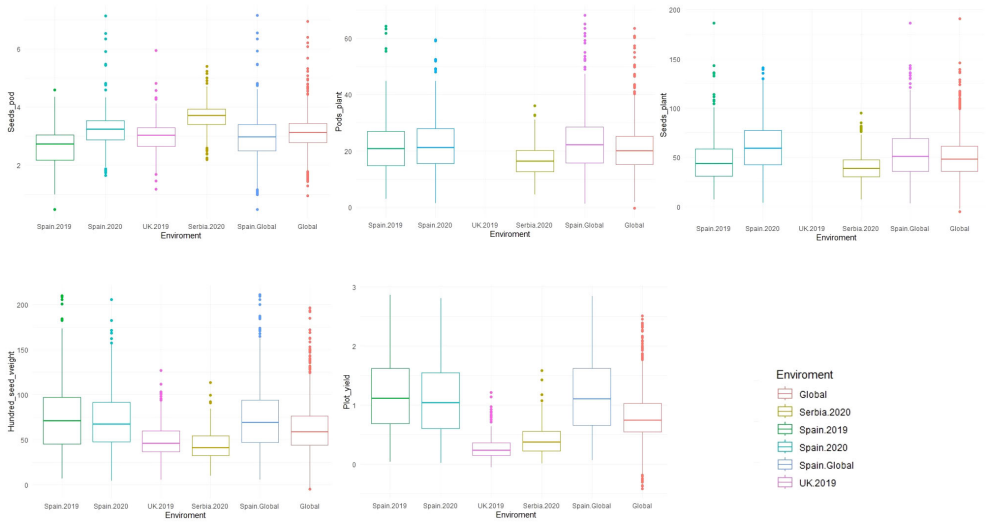


FIGURE 4
Boxplot of the phenotypic values for five yield-related traits measured in 352 faba bean accessions across different environments (*Spain.2019*, *Spain.2020*, *UK.2019*, *Serbia.2020*, *Spain.Global*, and *Global*). The traits were seeds per pod (SP), pods per plant (PP), seeds per plant (SPL), hundred-seed weight (HSW), and plot yield (PY).

10.3389/fpls.2024.1328690

TABLE 2 Descriptive statistics, heritability, and proportion of phenotypic variance explained in each environment after adjusting trait phenotypic values as described in the methods.

Trait	Environments		Mean	SD	Min	Max	Range	h^2	*varG*	*varS*	*varE*
Shattering (SH)	Spain.2019		1.01	1.31	−0.09	3.08	3.16	0.76	2.583	0.0025	0.8097
Seeds_pod (SP)	Spain.2019		2.56	0.76	0.48	4.59	4.11	0.60	0.6011	0.0013	0.3981
	Spain.2020		3.22	0.66	1.65	7.12	5.48	0.81	0.8039	0.0007	0.1826
	UK.2019		2.97	0.54	1.17	5.93	4.76	0.91	0.6892	0.0004	0.0667
	Serbia.2020		3.69	0.50	2.20	5.40	3.20	0.93	0.6975	0.0007	0.0552
	Spain.Global	Spain.2019	2.56	0.76	0.49	4.60	4.11	0.76	0.5187	0.0117	0.3321
		Spain.2020	3.22	0.66	1.65	7.13	5.48			0.0021	
	Global	Spain.2019	3.09	0.75	0.95	5.07	4.12	0.90	0.4652	0.0003	0.2240
		Spain.2020	3.09	0.65	1.54	6.94	5.39			0.0005	
		UK.2019	3.09	0.54	1.30	6.08	4.78			0.0005	
		Serbia.2020	3.08	0.50	1.60	4.79	3.20			0.0003	
Pods_plant (PP)	Spain.2019		21.57	9.58	2.91	64.29	61.38	0.65	111.2	0.1169	60.460
	Spain.2020		22.44	9.89	1.43	59.33	57.90	0.69	135.3	0.1472	61.300
	Serbia.2020		16.54	5.54	4.55	35.99	31.44	0.94	90.38	0.0702	5.204
	Spain.Global	Spain.2019	22.70	9.94	1.25	68.00	66.75	0.74	107.7	4.9100	67.280
		Spain.2020	22.97	10.13	2.50	61.67	59.17			3.1010	
	Global	Spain.2019	20.69	9.46	1.75	63.49	61.75	0.85	71.25	0.0547	52.570
		Spain.2020	20.67	9.84	−0.39	57.23	57.62			0.1383	
		Serbia.2020	20.70	5.55	8.50	40.24	31.74			0.0867	
Seeds_plant (SPL)	Spain.2019		46.45	24.58	7.24	186.46	179.23	0.65	720.8	1.0360	391.2
	Spain.2020		61.08	25.59	3.53	140.77	137.24	0.67	837.3	0.8801	420.5
	Serbia.2020		39.63	14.93	7.07	95.18	88.11	0.85	426.8	0.5247	76.53
	Spain.Global	Spain.2019	46.56	24.62	7.00	186.25	179.25	0.73	578.2	2.6050	457
		Spain.2020	61.17	25.63	3.33	141.00	137.67			3.0300	
	Global	Spain.2019	49.89	24.48	10.19	190.55	180.36	0.82	377.3	0.5867	369.4
		Spain.2020	49.75	25.34	−5.33	129.08	134.41			0.4953	
		Serbia.2020	49.94	14.95	17.09	105.29	88.20			0.3124	
H_Seed_weight (HSW)	Spain.2019		74.97	37.86	7.23	209.34	202.11	0.91	2,764	2.6	258.3
	Spain.2020		70.89	32.34	4.37	205.17	200.80	0.92	1,996	1.166	177.7
	UK.2019		49.09	18.02	5.81	126.61	120.81	0.98	755.2	0.2728	16.5
	Serbia.2020		43.53	16.12	10.18	113.31	103.12	0.97	518.1	0.2353	13.37
	Spain.Global	Spain.2019	74.97	37.97	6.11	210.44	204.33	0.90	1,670	4.361	380
		Spain.2020	71.37	32.77	5.50	209.72	204.22			13.09	
	Global	Spain.2019	61.52	37.45	−4.75	193.16	197.91	0.91	771.8	0.5101	308.2
		Spain.2020	61.71	32.21	−2.76	195.93	198.70			0.5388	
		UK.2019	61.52	18.05	17.66	139.20	121.54			0.5532	
		Serbia.2020	62.05	16.15	28.54	131.70	103.16			0.3986	

(Continued)

TABLE 2 Continued

Trait	Environments		Mean	SD	Min	Max	Range	h^2	*varG*	*varS*	*varE*
Plot_yield (PY)	Spain.2019		1.15	0.58	0.04	2.86	2.82	0.95	0.6082	0.0002	0.0351
	Spain.2020		1.13	0.62	0.02	2.80	2.79	0.93	0.7990	0.0004	0.0612
	UK.2019		0.27	0.19	−0.05	1.22	1.27	0.71	0.0398	0.0000	0.0165
	Serbia.2020		0.42	0.25	0.02	1.59	1.57	0.87	0.1236	0.0002	0.0185
	Spain.Global	Spain.2019	1.18	0.60	0.09	2.84	2.75	0.83	0.3867	0.0140	0.1257
		Spain.2020	1.15	0.63	0.07	2.83	2.76			0.0113	
	Global	Spain.2019	0.79	0.58	−0.37	2.46	2.82	0.78	0.1135	0.0001	0.1373
		Spain.2020	0.79	0.61	−0.42	2.52	2.93			0.0001	
		UK.2019	0.79	0.19	0.50	1.81	1.31			0.0002	
		Serbia.2020	0.79	0.25	0.38	1.96	1.58			0.0001	

with the other traits, while HSW had it with PP, SPL, and SH, as well as PP and SP between them (Supplementary Table S3).

Genome-wide association mapping

Association analyses were performed for the six yield-related traits using individual site data as well as the *Spain.Global* and *Global* values. This identified 112 MTAs in 97 candidate genes, of which 77 had functional annotation (Table 3). Of them, 40 harbored markers were significantly associated with the traits [Bonferroni threshold −log10 (p) > 5.66]. The Manhattan and their corresponding Q–Q plots are shown in Supplementary Figure S2.

A total of 8, 17, 25, 17, 14, and 16 unique MTAs were detected for SH, SP, PP, SPL, HSW, and PY, respectively (Table 3; Supplementary Figure S2). Among the associated SNP markers, eight were detected and validated in *Global* environments or in at least two environments [AX-181439008 (Chr1L), AX-416789762 (Chr1L), AX-416724190 (Chr3), AX-416783339 (Chr4), AX-416765420 (Chr4), AX-181163860 (Chr5), AX-416822980 (Chr6), AX-416730393 (Chr6)], and five markers were found to be associated with multiple traits (Supplementary Table S4). Thus, markers AX-181460360 (Chr1L), AX-416788562 (Chr4), AX-416724626 (Chr5), and AX-416730393 (Chr6) were associated with both PP and SPL, which showed the highest phenotypic correlation in all environments, and AX-181205104 (Chr3) was significant for SPL and SH, which revealed a reasonable high phenotypic correlation as well. No common associated markers were identified for PY, while PP was the trait sharing more markers in different environments or with correlated traits, and all of them were further significant in the global analyses (Table 3; Supplementary Table S4).

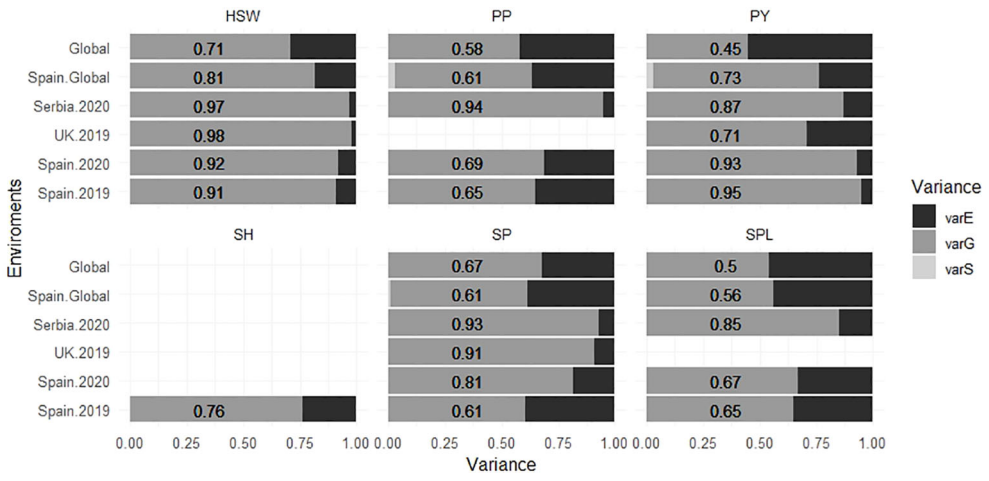


FIGURE 5
Proportion of the different components of the variance for six faba bean yield-related traits calculated for each environment. The numbers correspond with the value of the *varG*. *varE* is the residual variance component, *varG* is the additive genetic variance component, and *varS* is the spatial variance component for each environment. The traits were SH (shattering), SP (seeds per pod), SPL (seeds per plant), HSW (hundred-seed weight), PP (pods per plant), and PY (plot yield).

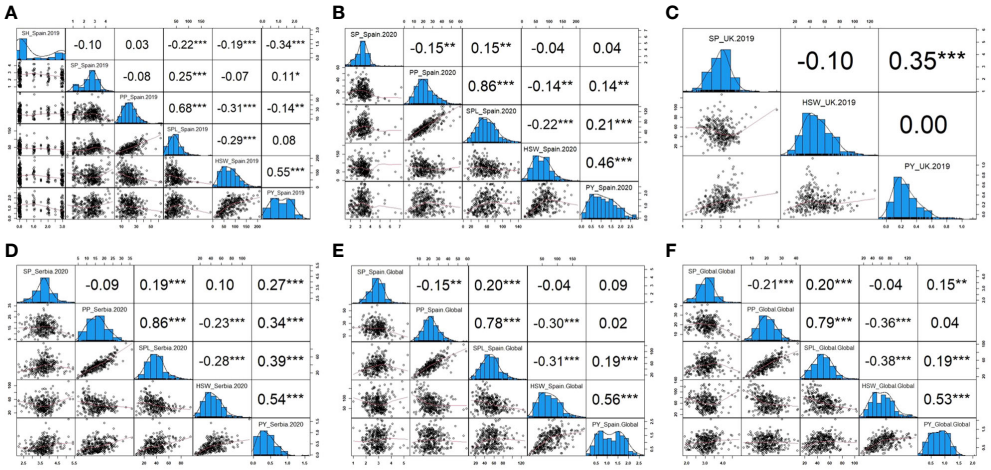


FIGURE 6
Pearson correlations and histograms showing the distribution of the yield-related traits in each environment. **(A)** Phenotypic frequency and correlation between SH (shattering), SP (seeds per pod), PP (pods per plant), SPL (seeds per plant), HSW (hundred-seed weight), and PY (plot yield) corresponding to Spain in 2019. **(B)** Phenotypic frequency and correlation between SP, PP, SPL, HSW, and PY corresponding to Spain in 2020. **(C)** Phenotypic frequency and correlation between SP, HSW, and PY corresponding to the United Kingdom in 2019. **(D)** Phenotypic frequency and correlation between SP, PP, SPL, HSW, and PY corresponding to Serbia in 2020. **(E)** Phenotypic frequency and correlation between SP, PP, SPL, HSW, and PY corresponding to Spain combined across years. **(F)** Phenotypic frequency and correlation between SP, PP, SPL, HSW, and PY corresponding to the Global analysis combining all locations. *, **, and *** significance at $p< 0.05$, $p< 0.01$, and $p< 0.001$, respectively.

TABLE 3 List of candidate genes associated with yield traits in different environments.

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
Shattering (SH)	Spain.2019	AX-416812583	15.4	T	Vf4	4g097400	LOC11444286 - MTR_8g070520	Actin-related protein 8 (ARP)	Chr8: 32,420,418–32,427,339
	Spain.2019	AX-181205104	14.9	T	Vf3	3g118040	LOC11414164 - MTR_1g094730	Protein LYK5	Chr1: 46,228,212–46,230,102
		AX-416731109	12.2	T	Vf4	4g045480	LOC25500825 - MTR_8g028565	WRKY transcription factor 22	Chr8: 10,975,117–10,977,932
		AX-181188996	9.3	F	Vf1L	1g440520	LOC25486300 - MTR_2g029540	30-kDa cleavage and polyadenylation specificity factor 30	Chr2: 11,149,134–11,161,704
		AX-416745732	8.6	F	Vf6	6g023600	LOC11444615 - MTR_4g054920	Cytochrome P450 94C1	Chr4: 27,970,823–27,972,919
		AX-416735297	7	F	Vf3	3g084880	–	–	–
		AX-181148136	6.6	F	Vf2	2g045960	LOC11419536 - MTR_3g055920	Probable cytokinin riboside 5′-monophosphate phosphoribohydrolase LOGL3	Chr3: 25,411,325–25,415,851
		AX-416817919	6.2	F	Vf1L	1g317840	LOC11413965 - MTR_5g038450	Probable serine/threonine-protein kinase PIX7	Chr5: 16,797,826–16,802,888
Seeds_pod (SP)	Spain.2019	AX-416729362	11.2	F	Vf2	2g239680	LOC11421272 - MTR_4g024600	UDP-glucuronate 4-epimerase 3	Chr4: 8,834,452–8,837,002
		AX-181153939	10.4	T	Vf5	5g054520	LOC11408637 - MTR_7g067460	Pentatricopeptide repeat-containing protein At4g28010 (PRP)	Chr7: 31,443,108–31,450,139
		AX-416740100	9.4	T	Vf4	4g124760	LOC120577622	Uncharacterized LOC120577622	

(Continued)

10.3389/fpls.2024.1328690

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
									Chr8: 36,889,875–36,900,957
		AX-181203726	8.6	T	Vf3	3g080360	LOC11410019 - MTR_1g080340	Histone-lysine N-methyltransferase SUVR4	Chr1: 39,567,264–39,578,478
		AX-416759496	6	F	Vf1L	1g443800	LOC11440205 - MTR_4g108150	14 kDa proline-rich protein DC2.15 (PRP)	Chr4: 52,817,542–52,818,514
		AX-416724190	5.7	F	Vf3	3g038960	LOC120577666	40S ribosomal protein S17-like (RP)	Chr8: 37,485,907–37,486,648
		AX-416790750	4.4	F	Vf1S	1g069440	LOC11416615 - MTR_2g083210	COBRA-like protein 10	Chr2: 41,038,841–41,043,276
	UK.2019	AX-181489146	19.5	T	Vf1S	1g039440	LOC25487826 - MTR_2g094970	Helicase protein MOM1	Chr2: 46,655,993–46,663,142
		AX-181166098	14.3	T	Vf1L	1g227200	LOC11426315 - MTR_5g071550	Probable protein phosphatase 2C 2 (PP2C)	Chr5: 31,700,909–31,703,736
		AX-416778732	13.2	T	Vf1S	1g065120	LOC25487504 - MTR_2g084715	Splicing factor U2af small subunit B	Chr2: 42,015,122–42,019,580
		AX-416816331	7.7	F	Vf3	3g205240	LOC11425698 - MTR_1g012590	UDP-N-acetylglucosamine transporter UGNT1	Chr1: 2,584,219–2,590,273
		AX-181493417	7.3	T	Vf4	4g047120	LOC25500757 - MTR_8g028075	Cytochrome b561 and DOMON domain-containing protein At3g07570	Chr8: 10,468,557–10,475,734
		AX-416732455	6.7	F	Vf1S	1g159880	–	–	–
		AX-416741800	5.7	F	Vf1L	1g387400	LOC11439490 - MTR_5g010000	Endoglucanase	Chr5: 2,586,159–2,589,390
		AX-181438122	4.3	F	Vf2	2g131080	LOC11425191 - MTR_3g091610	Phytolongin Phyl2.2	Chr3: 45,158,382 - 45,160,038
	Spain.Global	AX-181149329	8.1	F	Vf3	3g197400	LOC11423171 - MTR_1g008300	Rhomboid-like protein 11, chloroplastic	Chr1: 881,733–885,768
		AX-416724190	3.5	T	Vf3	3g038960	LOC120577666	40S ribosomal protein S17-like (RP)	Chr8: 37,485,907–37,486,648
	Global	AX-416736976	11.7	F	Vf0	Ung029360	LOC11444270	Rhamnogalacturonate lyase-like (RGL)	Chr8: 30,863,233–30,871,957
		AX-416724190	4.5	T	Vf3	3g038960	LOC120577666	40S ribosomal protein S17-like (RP)	Chr8: 37,485,907–37,486,648
Pods_plant (PP)	Spain.2019	AX-416730393	8.6	F	Vf6	6g200160	LOC11416749 - MTR_8g107380	Protection of telomeres protein 1b (POT1)	Chr8: 49,513,374–49,517,113
		AX-416765420	8	T	Vf4	4g124760	LOC120577622	Uncharacterized LOC120577622	Chr8: 36,889,875–36,900,957

(Continued)

10.3389/fpls.2024.1328690

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
		AX-181163476	7	F	Vf1S	1g090480	LOC11422251 - MTR_6g078300	54S ribosomal protein L37, mitocondrial	Chr6: 36,744,596–36,747,938
		AX-416822980	6.4	F	Vf6	6g138400	LOC11411398 - MTR_8g088150	Carboxyl-terminal-processing peptidase 1, chloroplastic (CTP)	Chr8: 40,577,160–40,583,029
		AX-416807594	5.9	F	Vf3	3g200600	LOC25481877 - MTR_1g009860	Protein ACTIVITY OF BC1 COMPLEX KINASE 1, chloroplastic	Chr1: 1,599,423–1,610,255
		AX-416799595	3.9	F	Vf1L	1g328120	LOC11433776 - MTR_5g043880	Probable WRKY transcription factor 57	Chr5: 19,193,017–19,206,704
	Spain.2020	AX-181460360	10.2	T	Vf1L	1g400320	LOC11437498 - MTR_6g088180	Cleavage stimulation factor subunit 50 (CstF-50)	Chr6: 40,794,459–40,808,637
		AX-416733063	8.8	F	Vf1L	1g346160	LOC11440716 - MTR_4g124030	DNA-directed RNA polymerase I subunit 2	Chr4: 59,293,111–59,319,043
		AX-181439008	7.6	T	Vf1L	1g283160	LOC11423535 - MTR_5g096830	Cytochrome c oxidase assembly protein COX15	Chr5: 43,546,294–43,551,833
		AX-416812479	7.5	T	Vf1S	1g009440	LOC11410541 - MTR_7g010950	Cyclic dof factor 2	Chr7: 3,056,313–3,060,692
		AX-416783339	5.4	F	Vf4	4g181200	LOC11438181 - MTR_4g115970	Pyrophosphate-energized vacuolar membrane proton pump	Chr4: 56,052,505–56,058,129
		AX-416760647	5.4	F	Vf1S	1g123640	LOC25495647 - MTR_6g018920	Floral homeotic protein AGAMOUS	Chr6: 7,501,850–7,503,098
		AX-416775893	4.7	F	Vf1L	1g243680	LOC11408143 - MTR_5g081250	Folylpolyglutamate synthase	Chr5: 36,003,484–36,010,018
	Serbia.2020	AX-416730393	13.2	T	Vf6	6g200160	LOC11416749 - MTR_8g107380	Protection of telomeres protein 1b (POT1)	Chr8: 49,513,374–49,517,113
		AX-416724626	10.6	T	Vf5	5g008520	LOC11436354 - MTR_2g101410	Protein JINGUBANG (JGB)	Chr2: 49,715,905–49,718,673
		AX-416748507	8.5	F	Vf1S	1g078200	LOC25487288 - MTR_2g078540	Uncharacterized LOC25487288	Chr2: 38,871,191–38,875,866
		AX-416739028	7.1	T	Vf5	5g095400	LOC11411488 - MTR_7g088360	Probable protein S-acyltransferase 3	Chr7: 41,196,893–41,199,796
		AX-416763724	6.7	F	Vf1S	1g069400	LOC11415615 - MTR_2g083220	Rab GTPase-activating protein 22	Chr2: 41,042,418–41,049,391
		AX-416754106	5.8	F	Vf3	3g238080	LOC11437395 - MTR_3g114030	Thaumatin-like protein 1b	Chr3: 56,596,596–56,602,725
		AX-416793421	4.4	F	Vf1S	1g017520	LOC25497623 - MTR_7g017900	6-phosphogluconate dehydrogenase, decarboxylating 2	Chr7: 6,009,237–6,013,808
	Spain.Global	AX-416765420	11.9	T	Vf4	4g124760	LOC120577622	Uncharacterized LOC120577622	

(Continued)

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
									Chr8: 36,889,875–36,900,957
		AX-416788562	8.3	T	Vf4	4g204800	LOC11418607 - MTR_4g124560	DNA repair protein RAD51 homolog	Chr4: 59,677,804–59,681,527
		AX-416777278	8.1	T	Vf5	5g112120	–	–	–
		AX-416772251	7.8	T	Vf2	2g084120	LOC25489423 - MTR_3g074050	Phosphatidylinositol transfer protein 3	Chr3: 36,901,813–36,904,344
		AX-181163860	6.7	F	Vf5	5g030840	LOC11420563 - MTR_7g058430	E3 ubiquitin-protein ligase PRT1	Chr7: 27,468,773–27,476,588
		AX-181497612	6.3	F	Vf2	2g241320	LOC11423056 - MTR_4g023560	ATP-dependent DNA helicase 2 subunit KU80	Chr4: 8,492,295–8,500,374
		AX-416822980	6.3	F	Vf6	6g138400	LOC11411398 - MTR_8g088150	Carboxyl-terminal-processing peptidase 1, chloroplastic (CTP)	–
		AX-181189847	4.9	F	Vf4	4g231800	LOC11435791 - MTR_4g132340	U3 small nucleolar RNA-associated protein 6 homolog	Chr8: 40,577,160–40,583,029
	Global	AX-181439008	6.2	F	Vf1L	1g283160	LOC11423535 - MTR_5g096830	Cytochrome c oxidase assembly protein COX15	Chr5: 43,546,294–43,551,833
		AX-416765420	6.1	F	Vf4	4g124760	LOC120577622	Uncharacterized LOC120577622	Chr8: 36,889,875–36,900,957
		AX-181163860	5.9	F	Vf5	5g030840	LOC11420563 - MTR_7g058430	E3 ubiquitin-protein ligase PRT1	Chr7: 27,468,773–27,476,588
		AX-416783339	4.8	F	Vf4	4g181200	LOC11438181 - MTR_4g115970	Pyrophosphate-energized vacuolar membrane proton pump	Chr4: 56,052,505–56,058,129
Seeds_plant (SPL)	Spain.2019	AX-416783780	7.8	T	Vf0	–	–	–	–
		AX-416757363	7.7	T	Vf2	–	–	–	–
		AX-416745161	7.4	T	Vf5	5g082040	LOC11429113 - MTR_7g083720	V-type proton ATPase subunit C	Chr7: 38,993,699–39,001,736
		AX-181186344	7.2	F	Vf6	6g117120	LOC25493041 - MTR_4g087905	Nudix hydrolase 15, mitocondrial	Chr4: 42,551,506–42,555,292
		AX-181205104	6.9	T	Vf3	3g118040	LOC11414164 - MTR_1g094730	Protein LYK5	Chr1: 46,228,212–46,230,102
		AX-181148593	6.4	F	Vf2	2g002120	LOC11416313 - MTR_3g010180	GPI-anchored protein LLG1 (GPI-AP)	Chr3: 2,335,419–2,337,669
		AX-416739501	4.6	F	Vf4	4g214600	LOC11440584 - MTR_2g036860	Villin-4	Chr2: 16,084,673–16,101,032
	Spain.2020	AX-181460360	12.6	T	Vf1L	1g400320	LOC11437498 - MTR_6g088180	Cleavage stimulation factor subunit 50 (CstF-50)	Chr6: 40,794,459–40,808,637

(Continued)

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
		AX-181147789	8.7	T	Vf3	3g220120	LOC11431072 - MTR_1g018320	Strigolactone esterase RMS3	Chr1: 5,255,290–5,256,875
		AX-416761617	8.6	F	Vf4	4g180280	LOC11442878 - MTR_4g115640	Putative chloride channel-like protein CLC-g	Chr4: 55,897,598–55,903,091
		AX-181204824	7	F	Vf1L	1g340840	LOC11407099 - MTR_5g026910	Uncharacterized LOC11407099	Chr5: 11,164,632–11,168,640
		AX-416794410	6.6	F	Vf5	–	–	–	–
		AX-416790652	5.2	F	Vf3	3g184840	LOC11421623 - MTR_1g031780	Putative receptor-like protein kinase At4g00960 (RLK)	Chr1: 11,241,206–11,247,464
		AX-416788562	5	F	Vf4	4g204800	LOC11418607 - MTR_4g124560	DNA repair protein RAD51 homolog	Chr4: 59,677,804–59,681,527
		AX-416781866	4.8	F	Vf6	6g049840	LOC11440200 - MTR_4g071880	Fructose-bisphosphate aldolase 1, chloroplastic (FBA)	Chr4: 35,238,383–35,247,955
	Serbia.2020	AX-416724626	9.2	T	Vf5	5g008520	LOC11436354 - MTR_2g101410	Protein JINGUBANG (JGB)	Chr2: 49,715,905–49,718,673
		AX-181493754	8.6	F	Vf6	6g046000	LOC25492796 - MTR_4g073840	Putative disease resistance RPP13-like protein 3	Chr4: 36,019,614–36,023,252
		AX-416817126	8.2	F	Vf2	2g090040	LOC11413114 - MTR_3g077080	Copper methylamine oxidase	Chr3: 38,099,287–38,107,913
		AX-416730393	8	T	Vf6	6g200160	LOC11416749 - MTR_8g107380	Protection of telomeres protein 1b (POT1)	Chr8: 49,513,374–49,517,113
		AX-416752104	7.3	F	Vf1S	1g198960	LOC11417190 - MTR_1g018030	Serotonin N-acetyltransferase 2, chloroplastic	Chr1: 5,212,095–5,214,935
		AX-181204029	6.8	F	Vf6	6g091000	LOC11437903 - MTR_4g096900	Aspartate carbamoyltransferase 3, chloroplastic (ATCase)	Chr4: 46,486,489–46,491,286
		AX-416778114	5.2	F	Vf3	3g080280	LOC11407475 - MTR_1g080320	Uncharacterized LOC11407475	Chr1: 39,552,506–39,557,764
H_seed_weight (HSW)	Spain.2020	AX-416773275	16.8	F	Vf6	6g042880	LOC25492681 - MTR_4g068457	Protein JINGUBANG (JGB)	Chr4: 33,844,582–33,847,451
		AX-181448659	10.3	F	Vf4	4g151200	LOC11412050 - MTR_4g104350	Protein STICHEL-like 2	Chr4: 51,062,681–51,069,264
		AX-416722650	6.8	F	Vf4	4g114560	LOC11408165 - MTR_8g076150	V-type proton ATPase 16 kDa proteolipid subunit	Chr8: 34,714,373–34,716,329
		AX-416816211	2.7	F	Vf5	5g181520	LOC11430562 - MTR_7g113920	Phosphoglycerate mutase-like protein AT74H	Chr7: 54,000,124–54,003,242
	UK.2019	AX-416789762	16.9	T	Vf1L	–	–	–	–
	Serbia.2020	AX-416792429	15.8	T	Vf5	–	–	–	–

(Continued)

10.3389/fpls.2024.1328690

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
		AX-181497672	14	F	Vf1L	1g225400	LOC11405494-MTR_5g032380	Ubiquitin carboxyl-terminal hydrolase 17	Chr5: 13,805,006–13,815,781
	Global	AX-181197475	20.5	T	Vf1L	–	LOC11436753-MTR_5g009700	Hydroxyproline O-galactosyltransferase HPGT1	Chr5: 2,415,935–2,423,973
		AX-416789762	15.9	T	Vf1L	–	–	–	–
		AX-181483657	15.2	F	Vf1L	1g441080	LOC11414408-MTR_2g029040	BRISC and BRCA1-A complex member 2	Chr2: 10,878,648–10,887,239
		AX-416813674	12.7	T	Vf1S	–	–	–	–
		AX-181471800	7.3	F	Vf4	4g000600	LOC11435395-MTR_8g005270	Mannan endo-1,4-beta-mannosidase 6	Chr8: 230,980–234,251
		AX-416816053	5.5	F	Vf3	3g171360	LOC25485595-MTR_1g112320	UPF0307 protein PMI3641	Chr1: 54,554,225–54,559,444
		AX-181187483	3.9	F	Vf2	2g269560	–	–	–
		AX-416761691	2.8	F	Vf5	5g039080	LOC25498271-MTR_7g056233	Oleoyl-acyl carrier protein thioesterase 1, chloroplastic	Chr7: 26,103,028–26,112,036
Plot_yield (PY)	Spain.2019	AX-416744497	21.9	T	Vf2	–	–	–	–
		AX-416760651	19.8	T	Vf1S	1g089640	LOC11422248-MTR_6g077860	DNA-binding protein ROOT HAIRLESS 1 gene (RHL1)	Chr6: 36,585,521–36,596,508
		AX-181494841	14.5	T	Vf1S	1g043400	LOC11417361-MTR_2g093100	Uncharacterized LOC11417361	Chr2: 45,759,498–45,770,253
		AX-181459267	12.1	F	Vf4	4g161640	LOC11440572-MTR_4g107830	Hypersensitive-induced response protein 1 (HIR1)	Chr4: 52,667,925–52,682,756
		AX-416814439	12	F	Vf4	4g180440	LOC11446673-MTR_4g115620	Probable sucrose-phosphate synthase (SPS)	Chr4: 55,872,671–55,880,769
		AX-416824665	10	T	Vf1S	1g119680	LOC11418954-MTR_6g023340	Basic endochitinase	Chr6: 8,380,403–8,381,786
		AX-416801028	8.3	F	Vf3	3g038240	LOC25483703-MTR_1g057150	Protein SMAX1-LIKE 3	Chr1: 29,007,041–29,012,213
		AX-416751279	7.9	F	Vf6	6g199840	LOC11415767-MTR_8g107250	Tubulin beta chain	Chr8: 49,453,661–49,456,244
	UK.2019	AX-416786350	14.5	T	Vf1L	1g079040	LOC11442148-MTR_5g080260	Uncharacterized LOC11442148	Chr5: 35,537,265–35,538,015
		AX-416803636	10.8	F	Vf2	2g127080	LOC11427826-MTR_3g090480	LRR receptor-like serine/threonine-protein kinase RCH1	Chr3: 44,517,008–44,521,950
		AX-181440082	10.4	T	Vf5	5g039640	LOC11436813-MTR_7g046250	E3 ubiquitin-protein ligase BRE1-like 1	Chr7: 22,277,729–22,300,671
		AX-416749659	8.9	F	Vf2	2g274360		Metal tolerance protein C4	

(Continued)

TABLE 3 Continued

Trait	Location	Axiom_Vfaba ID[a]	% R2	Bonf	Chr_Vf	Reference genome Vf	MT_Ortholog	Gene annotation	Mt locations[b]
							LOC25491241 - MTR_4g008150		Chr4: 1,626,201–1,632,456
		AX-416739576	7.7	F	Vf6	6g038240	LOC11427645 - MTR_4g064740	Serine/arginine repetitive matrix protein 1	Chr4: 32,339,258–32,347,116
		AX-416740534	3.6	F	Vf1L	1g306120	LOC11413521 - MTR_5g033800	Protein trichome birefringence-like 2	Chr5: 14,494,614–14,500,404
		AX-181485567	3.2	F	Vf4	4g125640	LOC25493470 - MTR_4g094848	Filament-like plant protein 7	Chr4: 47,043,537–47,050,784
		AX-181456171	2.5	F	Vf3	–	–	–	–

The ID of the SNP markers in the *Vicia faba* Axiom, percentage of phenotypic variation explained (%R2) and Bonferroni threshold (T: true and F: false), location of the SNP in the faba bean chromosomes, contig in the reference genome and orthologous genes, annotation, and location in Medicago truncatula.
[a]In gray, the common associated loci among traits across environments.
[b]Gene locations were determined using the Genome Data Viewer (GDV).

Among environments, *Spain.2019* was the one revealing the highest number of associations for all the traits, except HSW (Table 3). Thus, the eight MTAs found for both SH and PY jointly explained 80.2% and 100% of the total phenotypic variation, respectively. In *Spain.2020*, seven, eight, and four MTAs accounted for 49.6%, 58.5%, and 36.6% of the variation for PP, SPL and HSW, respectively. The MTAs detected in *Serbia.2020* explained 56.3%, 53.3%, and 29.8% of the variation in PP, SPL and HSW, respectively. Finally, in *UK.2019*, the markers associated with SP, PY, and HSW accounted for 78.7%, 61.6%, and 16.9% of the respective trait variation. The global analyses *Spain.Global* or the combination of all the environments *Global* identified new candidates and confirmed some of the MTAs previously detected (Table 3).

The recent availability of the genome sequence (Jayakodi et al., 2023) enabled us to locate the significant MTAs on the faba bean physical map and visualize the genomic regions harboring multiple associations for the different traits and environments (Figure 7; Supplementary Table S5). The MTAs were distributed across the genome, but a few genomic regions harbored multiple associated SNPs (Figure 7). Four pleiotropic MTAs controlling multiple traits co-localized on Chr1L (PP-SPL), Chr3 (SH-SPL), Chr4 (PP-SPL), and Chr5 (PP-SPL). Seven stable MTAs expressed in multiple environments were identified for PP in Chr1L, Chr4, Chr5, and

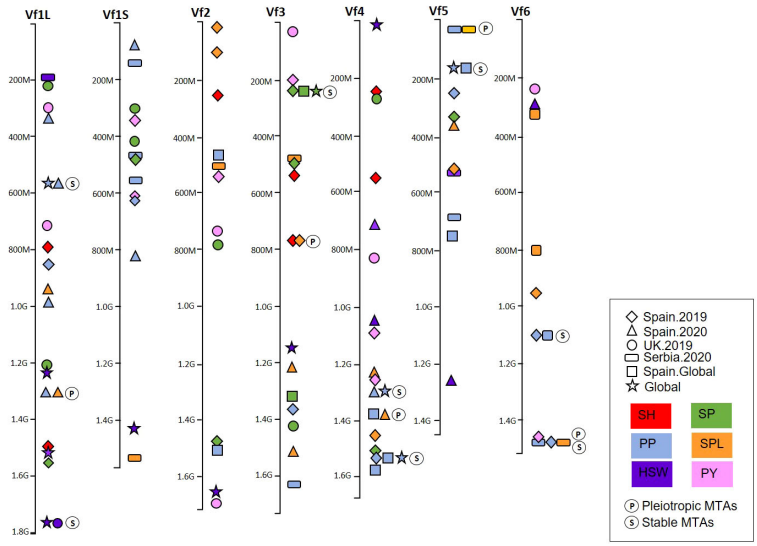


FIGURE 7
Physical map of the marker–trait associations (MTAs) of six yield-related traits detected by GWAS in a faba bean panel (352 accessions). The traits were represented by colors and the environments by shapes. (P) represents the MTAs with pleiotropic effect and (S) the MTAs stable across different environments. SH, shattering; SP, seeds per pod; PP, pods per pod; SPL, seeds per plant; HSW, hundred-seed weight; PY, plot yield.

Chr6; for HSW in Chr1L; and SP in Chr3. Finally, we found one MTA showing pleiotropic effects between PP and SPL and stable among environments on Chr6 (Figure 7). The nucleotide sequence of each associated SNP is available in Supplementary Table S6.

Postulation of candidate genes

To understand the potential role of MTAs in faba bean yield, we searched for homologs in *M. truncatula* (Table 3). Functional annotation of their molecular functions helps to predict candidate genes associated with the traits studied. The COG showed that 20 of them were unknown and 16 of them had no significant similarity with previous annotated sequences (Figure 8; Supplementary Table S7). The remaining candidates were involved in a wide variety of functions such as carbohydrate metabolism and transport, signal transduction, chaperone functions, transcription, replication and repair process, and cytoskeleton (Figure 8).

For shattering (SH), three significant MTAs were identified accounting for 15.4%, 14.9%, and 12.2% of the phenotypic variation, respectively. The candidate genes harboring the significant SNPs were "actin-related protein 8 (ARP)," "protein LYK5," and "WRKY transcription factor 22" (Table 3).

For seeds per pod (SP), the two main candidate genes containing the associated MTAs identified in *Spain.2019* were "UDP-glucuronate 4-epimerase 3" and "pentatricopeptide repeat-containing (PPR) protein" which explained 11.2% and 10.4% of the phenotypic variation, respectively. In *UK.2019*, "helicase protein MOM1," "protein phosphatase 2C 2 (PP2C)," and "splicing factor U2af small subunit B" harbored significant SNPs that explained 19.5%, 14.3%, and 13.2% of the trait variation, respectively. Two MTAs corresponding to "rhomboid-like protein" and "40S ribosomal protein S17-like (RP)" were also associated with *Spain.Global* and accounted for 8.1% and 3.5% of the phenotypic variation, respectively. Finally, the *Global* analysis identified MTAs corresponding to "rhamnogalacturonate lyase-like gene (RGL)" in addition to "40S ribosomal protein S17-like (RP)" (Table 3).

Pods per plant (PP) was the trait sharing the highest number of associated SNPs across environments and yield-related traits (SP), and most of the corresponding candidates were further significant in the global analyses (Table 3). Thus, "protection of telomeres protein 1b" (POT1) was associated with PP in *Spain.2019* and with PP and SP in *Serbia.2020* and was the significant marker explaining the highest percentage of variation (13.2%) of the trait. The MTA identified in the uncharacterized LOC120577622 was significant in *Spain.2019*, *Spain.Global*, and *Global*, explaining 8%, 11.9%, and 6.1% of the variation in PP, respectively, a fact supported by the high positive phenotypic correlations between environments ($r >$ 0.76) (Supplementary Figure S3). Similarly, the significant MTA present in "carboxyl-terminal peptidase (CTP)" was consistently associated with PP in *Spain.2019* and *Spain.Global*. Moreover, "cleavage stimulation factor subunit 50" (CstF-50) contained a significant MTA for PP and SPL that explained 10.2% and 12.6% of the variation, respectively. This outcome was further supported by the high positive phenotypic correlations existing between both traits (r = 0.86) (Figure 6). "Protein JINGUBANG" (JGB) also contained a significant MTA significant for PP and SPL that explained 10.6% and 9.2% of the variation, respectively. In *Spain.Global*, four associated MTAs uncovered new candidates, an uncharacterized gene was further confirmed in the *Global* analysis, "DNA repair protein RAD51 homolog" was consistently identified in PP and SP, and "E3 ubiquitin-protein ligase PRT1" was further validated in the *Global* analysis. As mentioned above, the pooled dataset analysis (*Global*) validated the genes harboring the associated SNPs (COX15, uncharacterized LOC120577622, E3 ubiquitin-ligase, and vacuolar proton pyrophosphatase) in all the environments.

For seeds per plant (SPL), the five candidate genes harboring significant MTAs consistently identified in different yield-related traits were LYK5 (*Spain.2019* for SPL and SH), CstF-5 and RAD51 (*Spain.2020* for SPL and PP), and JGB and POT1 (*Serbia.2020* for

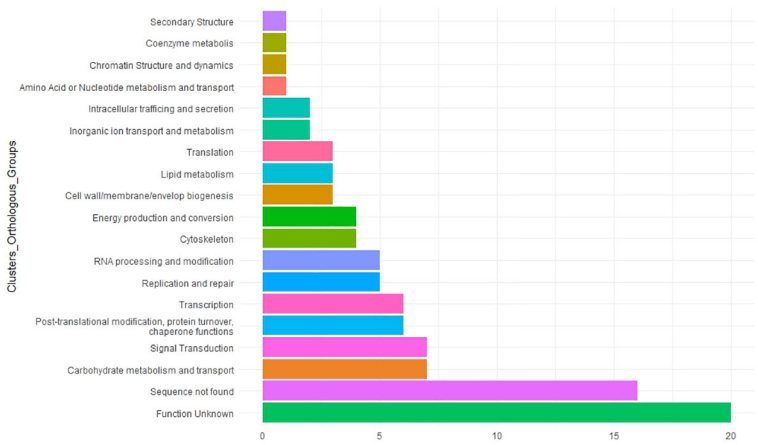


FIGURE 8
Functional analysis of 97 significant marker–trait associations (MTAs). Candidate genes were classified by the Clusters of Orthologous Groups (COGs). The *X*-axis indicates the number of genes in a category. The *Y*-axis shows the 17 functional COG categories found.

SPL and PP). Except for CstF-50, none of the MATs accounted for more than 10% of the phenotypic variation of this trait (Table 3).

In hundred-seed weight (HSW), five of the associated SNP markers corresponded to candidate gene sequences with no significant similarity and accounted for high values of the phenotypic variation (one of them common in the *UK.2019* and *Global* analyses). In *Spain.2020*, the candidate genes were "JGB" protein (different from the one detected in PP and SPL) and "protein STICHEL-like 2" and explained 16.8% and 10.3% of the total variation. In *Serbia.2020*, "ubiquitin carboxyl-terminal hydrolase 17" explained 14% of the trait variation. The associated SNPs in the *Global* analysis revealed "hydroxyproline O-galactosyltransferase HPGT1" accounting for 20.5% of the variation and BRISC and BRCA1-A complex member 2 (explaining 15.2%) (Table 3).

Plot yield (PY) was the only trait that did not share candidate genes among environments or with other correlated traits. In *Spain.2019*, six significant MTAs explained more than 90% of the phenotypic variation and corresponded to "DNA-binding protein ROOT HAIRLESS 1 gene (RHL1)," "hypersensitive-induced response protein 1 (HIR1)," "sucrose-phosphate synthase (SPS)," and "basic endochitinase." In *UK.2019*, the study revealed three main candidates: one of them uncharacterized, "LRR receptor-like serine/threonine-protein kinases RCH1," and "E3 ubiquitin-protein ligase BRE1-like 1," and the associated SNPs explained 14.5%, 10.8%, and 10.4% of the variation (Table 3). The remaining candidates have not yet been functionally characterized.

Discussion

Faba bean is an important crop for global food security, ecosystem resilience, and sustainable agriculture. To turn this crop into an economically attractive proposition for farmers and to increase the area under cultivation, major advances in yield and yield stability need to be achieved. Grain yield, however, is a complex trait controlled by many genes and strongly affected by the environment. Genetic improvement can be achieved, but the difficulty is in knowing which trait combinations should be selected to produce stable high-yielding genotypes. This study is a comprehensive effort to exploit a wide collection of important yield-related traits. We performed a GWAS analysis on a world collection of 400 faba bean accessions to assess its genetic diversity and population structure and to identify MTAs associated with yield using the Vfaba_v2 Axiom 60K SNP array. The panel was phenotyped for six yield-related traits (SH, SP, PP, SPL, HSW, and PY) in different locations from Spain, Serbia, and the United Kingdom. For phenotypic adjustment, we considered six models (*Spain.2019*, *Spain.2020*, *Spain.Global*, *UK.2019*, *Serbia.2020*, and *Global*). Overall, there was better performance of the genotypes in Spain than in Serbia or the UK, with better phenological development and greater range for all traits.

The high and positive phenotypic correlations detected between most of the traits in the different locations suggest that it is possible to improve multiple yield-related traits through genomics-assisted breeding. The exception was HSW, which was positively correlated with PY and negatively correlated with PP and SPL in all the environments. This outcome could be the result of genetic and/or environmental interactions with the availability and remobilization of photoassimilates during seed filling (Mangini et al., 2018). However, this explanation does not exclude the hypothesis of clusters of tightly linked genes with opposite effects and/or single genes with pleiotropic effects. On the other hand, the relatively low genetic correlation observed in PY among environments highlights a high G × E interaction with a significant effect on the final yield which is not shared with the other yield components studied. Thus, the bimodal distribution of PY in Spain (Supplementary Figure S1) suggests that the MTAs detected might differentiate between Mediterranean-adapted and not-adapted genotypes and could be related to other adaptive loci that respond to environmental factors. These outcomes highlight the fact that identifying MTAs for yield components offers more scope for future breeding than those associated with PY itself.

A high heritability was observed for all the traits in all environments, suggesting a strong genetic control and low microenvironmental effects, a fact desired for rapid progress in breeding. HSW was the least affected trait across environmental conditions, followed by PY and SP. Different studies have reported high narrow sense heritability for SP, PP, HSW, and PY (Toker, 2004; Zhou et al., 2021; Khan et al., 2022; Singh et al., 2022). Traits with high heritability will help breeders shorten the breeding cycles and can result in faster and higher genetic gain (Singh et al., 2022).

Based on the results of the population structure, the delta-*K* peak suggested the presence of three faba bean groups with clear differences in their geographic origin which was supported by the DAPC and the phylogenetic analysis. Three subpopulations differentiated by geographical origin were also found recently in a genetic analysis of a worldwide collection of 2,678 faba bean genotypes, including the EUCLEG panel (Skovbjerg et al., 2023) which also identified the EUCLEG collection being the most diverse among the panels analyzed. As indicated by the diagonal linear shape in the Q–Q plots (Supplementary Figure S2), the approach used controlled the population stratification, thus supporting the reliability of the GWAS detected. This collection with diverse phenotypic and molecular parameters constitutes a valuable resource for future breeding and high-resolution gene mapping, including candidate gene discovery for a wide range of traits (Govindaraj et al., 2015).

The adequacy of association studies for complex traits depends critically on the existence of LD between functional alleles and the surrounding SNP markers. LD values dropped from 0.140 to 0.125, with the increase of physical distance from 126.6 kbp to 151.8 kbp. Low LD in faba bean was previously reported (Skovbjerg et al., 2023; Zhang et al., 2023) when comparing different faba bean diversity panels. Thus, the EUCLEG collection was the one showing lower LD blocks (higher recombination), an expected outcome in outbreeding species with high genetic diversity (Pégard et al., 2023; Skovbjerg et al., 2023; Zanotto et al., 2023; Zhang et al., 2023). The rapid LD decay observed in this study (within roughly 150 kbp) revealed the variation present in a highly allogamous panmictic population and suggested that the MTAs identified (or the closely linked genes on either flanking side of the significant

SNPs) were nearly or completely independent from each other and, hence, were sufficient for association mapping in faba bean.

Concerning the candidate genes harboring the significant SNPs, we will mainly discuss the ones detected consistently in different environments and traits (stable and pleiotropic MTAs), explaining more than 10% of the trait variation (Table 3). Three main candidates were associated with shattering (SH): "actin-related protein 8 (ARP)," "protein LYK5," and "WRKY transcription factor 22." No single specific role has been defined for nuclear ARPs that are involved in many cellular physiological processes including plant growth and development (Szymanski and Staiger, 2018). Of particular interest is the role of actin as a sensor mechanism for chemical and physical perturbations in the intracellular and extracellular environment (Porter and Day, 2016), as what may happen in mature faba bean pods. The protein LYK5 has been reported to be involved in cell wall integrity maintenance mechanisms (Baez et al., 2022). In the signaling response of *Arabidopsis* to fungal chitin, the LYK5 plays a direct role in chitin signaling and plant innate immunity (Bacete and Hamann, 2020). Interestingly, LYK5 was pleiotropic with the correlated trait SPL. A WRKY transcription factor has been reported to modulate floral and seed development, lignin deposition, and shattering process in sorghum (Tang et al., 2013; Ogutcen et al., 2018). In our study, the protein LYK5 showed a pleiotropic effect with SH and SPL in *Spain.2019*, supported by the significant negative correlation between SH and SPL in this environment ($r = -0.22$), thus suggesting the importance of this candidate for yield improvement through marker-assisted selection (MAS).

The seeds per pod (SP) main candidates bearing the associated MTAs were "UDP-glucuronate 4-epimerase 3" and "pentatricopeptide repeat-containing (PPR) protein." Several authors (Duan et al., 2022; Li et al., 2022) reported that gene seed thickness 1 (ST1), encoding UDP-d-glucuronate 4-epimerase 6, influences soybean seed morphology via the pectin biosynthesis pathway. In addition, genes encoding PPR proteins have been shown to play prominent roles in seed development in different crops. Thus, a PPR protein was partly responsible for the increased seed size and weight during domestication in peanut (Li et al., 2021), and it affected photosynthesis and grain filling in maize (Liu et al., 2013; Huang et al., 2020) and regulated pod number in chickpea (Das et al., 2016). Other significant MTAs, although explaining less percentage of variation, corresponded to "histone-lysine N-methyltransferase SUVR4" with functions in plant growth and development, including pollen and female gametophyte development, flowering, and responses to stresses. The other three significant MTAs for SP corresponded to "helicase protein MOM1," "protein phosphatase 2C 2 (PP2C)," and "splicing factor U2af small subunit B." The *in-vivo* role of many helicases has not been well investigated in plants; however, through indirect evidence, it has been suggested that they play critical roles in biological pathways encompassing all aspects of cell biology, organismal physiology, development, and stress physiology (Sami et al., 2021). Plant PP2Cs have emerged as major players in stress signaling. A PP2C is a major signaling component in the ABA-dependent signaling cascade that regulates seed germination in rice (Bhatnagar et al., 2017) or seed dormancy abscisic and abscisic acid-activated protein kinases in *Arabidopsis* (Kim et al., 2013). The regulation of MAPK activities by PP2Cs in plants indicates that protein phosphatases may act as specificity determinants in MAPK signaling (Umezawa et al., 2009). The next candidate is the splicing factor U2af responsible for removing introns from precursor mRNAs (pre-mRNAs) in all eukaryotes. Park et al. (2019) showed that normal plant development, including floral transition, and male gametophyte development in *Arabidopsis* require two U2af isoforms. However, the specific molecular mechanisms related to the regulation of splicing for the control of plant growth, reproduction, and stress response are still unknown. The next significant SNP corresponded to "40S ribosomal protein S17-like (RP)" and was the only candidate validated in the different environments (*Spain.2019*, *Spain.Global*, and *Global*). RPs are indispensable in ribosome biogenesis and protein synthesis and play a crucial role in diverse developmental processes (Ban et al., 2000; Klein et al., 2004). In rice, RPs have been demonstrated to be involved in inflorescence development and grain filling (Saha et al., 2017), while in foxtail millet, ribosomal proteins were identified to be significantly increased during drought (Pan et al., 2018). Finally, the analysis combining all the environments also identified "rhamnogalacturonate lyase-like gene (RGL)." RGLs have been shown to have a key role during pollen tube growth and defense against pathogens in tomato (Ochoa-Jiménez et al., 2018), cell expansion and growth and plant development in cotton (Naran et al., 2007), and involvement in fruit softening and ripening (Vicente et al., 2007).

Pods per plant (PP) was the trait showing a higher number of MTAs stable among environments and/or pleiotropic with the highly correlated SPL trait. Stable candidates were "protection of telomeres protein 1b (POT1)," "carboxyl-terminal peptidase (CTP)," "cytochrome c oxidase assembly protein COX15," "pyrophosphate-energized vacuolar membrane proton pump," and "E3 ubiquitin-protein ligase PRT1." Telomeres are nucleoprotein complexes that physically cap the ends of chromosomes preventing them from rapid degradation. "Carboxyl-terminal peptidase (CTP)" represents an unusual and poorly understood class of serine proteases found in a broad range of organisms, controlling multiple cellular processes (binding, posttranslational modifications, and trafficking), through posttranslational modification of proteins (Carroll et al., 2014). COX is the last enzyme of the mitochondrial respiratory chain, playing a key role in the regulation of aerobic production of energy. The lack of mutant plants in COX components, due to embryonic lethality, highlights the importance of COX activity in plants (Mansilla et al., 2018). Next, vacuolar proton pyrophosphatases facilitate auxin biosynthesis, transport, signaling, conjugation, and catabolism during seed development (Cao et al., 2020). Auxin is a regulator of yield contributing to ovule and seed growth, morphogenesis, and progression through different reproductive stages in *Arabidopsis*, rice, or maize among other examples (Shirley et al., 2019). Finally, E3 ubiquitin-ligases are involved in the regulation of plant innate immunity (Duplan and Rivas, 2014) and also affect the induction of flowering in angiosperms, ensuring their formation when conditions are optimal for pollination and

that seeds or subsequent seedlings have time to develop. Pleiotropic candidates were "cleavage stimulation factor subunit 50 (CstF-50)," "protein JINGUBANG (JGB)," "DNA repair protein RAD51 homolog," and "protection of telomeres protein 1b (POT1)" already mentioned above. CstF-50 has a regulatory role that is indispensable for the biogenesis of mRNA and participates in the control of gene expression under DNA-damaging conditions by regulating polyadenylation/deadenylation (Cevher et al., 2010). JGB is a pollen-specific protein containing seven WD40 repeats that regulate pollen germination and tube growth ensuring pollination in moist environments (Ju et al., 2016). Further analysis of the group revealed that JGB interacts with the transcription factor TCP4 to control pollen jasmonic acid synthesis. Interestingly, the consistent and pleiotropic effect of JGB with three yield components (PP, SP, and HSW) may be crucial for molecular breeding of yield-related traits in faba bean. Finally, RAD51 proteins contribute to genome stability by repairing DNA damage after replication, transcription, or cellular metabolic activities and play a direct role in the control of immune responses (Wang et al., 2010). Moreover, the phenotypic characterization of RAD51 in *Arabidopsis* revealed that this gene has an essential function in male and female meiosis (Li et al., 2004).

Seeds per plant (SPL) was the trait revealing more unknown or uncharacterized proteins. Five candidates were pleiotropic with other yield-related traits: LYK5 (with SPL and SH) and CstF-5, JGB, POT1, and RAD51 (with PP and SPL) explained above.

For hundred-seed weight (HSW), apart from JINGUBANG (JGB), the four SNPs with corresponding candidate genes, explaining a relevant percentage of the variation, were "protein STICHEL-like 2," whose role in seed size modulation is still not known; "ubiquitin carboxyl-terminal hydrolase 1"; "hydroxyproline O-galactosyltransferase HPGT1"; and "BRISC and BRCA1-A complex member 2." Several components of the ubiquitin pathway have been found to play critical roles in the regulation of seed and organ size in *Arabidopsis* and rice (Li and Li, 2014) and in cacao (Bekele et al., 2022). "Hydroxyproline O-galactosyltransferase HPGT1" has a functional role in various aspects of plant growth, development, and fertility (e.g., germination, seed set, seed size and morphology, and silique length) in *Arabidopsis* (Kaur et al., 2021). Finally, "BRISC and BRCA1-A complex member 2" has been shown to be involved in DNA repair (Block-Schmidt et al., 2011). Other candidates explaining a lower percentage but putatively involved in seed development were "mannan endo-1,4-beta-mannosidase 6", "UPF0307 protein PMI3641," and "oleoyl-acyl carrier protein thioesterase 1."

Finally, in the case of plot yield (PY), nine of the candidates harboring the significant SNPs explained more than 10% of the variation. "DNA-binding protein RHL1" (ROOT HAIRLESS 1 gene) encodes a nuclear protein required for root hair initiation in *Arabidopsis* (Schneider et al., 1998) and for ploidy-dependent cell growth (Sugimoto-Shirasu et al., 2005). The next candidate was "hypersensitive-induced response protein 1" (HIR1). A highly conserved interaction between receptor-like protein kinase (LRR1) and "hypersensitive-induced response protein 1" (HIR1) homologs is a common mechanism in the defense response of both monocots and dicots (Zhou et al., 2009). "Sucrose-phosphate synthase (SPS)" is a key enzyme in the plant sugar metabolic pathways with functions on growth, development, and yield. Several authors have reported that enhancement of SPS activity provides a higher carbon partitioning that increases growth and grain yield in rice (Sharkey et al., 2000; Mulyatama et al., 2022) or potato (Ishimaru et al., 2008). The next candidate is "basic endochitinase" reported to enhance the defense system of plants (as they act on chitin, the major component of the cell wall) and also to improve plant growth and yield in different crops (Gongora and Broadway, 2002; Jeong et al., 2010). Next, "LRR receptor-like serine/threonine-protein kinases RCH1" has been reported to play vital roles in plant growth and development and the responses to environmental stress (Liu et al., 2017), while "E3 ubiquitin-protein ligase BRE1-like 1" and the ubiquitin system in general affect plant health, reproduction, and responses to the environment, processes that impact important agronomic traits such as the induction of flowering, yield, and pathogen responses (Linden and Callis, 2020).

The relatively large amounts of MTAs associated with yield components identified in this study are promising candidates for follow-up studies on the validation of genes controlling faba bean production. Although most of the MTAs were locally relevant in response to the conditions of a particular location or year, 12 of them (Supplementary Table S4) were stable across different environments and/or were associated with multiple traits. The MTAs identified or the closely linked genes on either flanking side of the significant SNPs are likely to represent significant candidates for the molecular breeding of faba bean yield-related traits.

Conclusion

Genetic dissection of the genomic regions controlling faba bean yield is of great interest for the development of highly productive varieties. By phenotyping a worldwide collection of 400 faba bean accessions in different environments and using a GWAS analysis, this study provides a comprehensive genomic resource for the genetic dissection of yield components in this crop. A wide variability, a high heritability, and a high positive correlation were observed for most of the traits (except for the final PY) in all the environments studied. Overall, the panel revealed 112 associated MTAs linked with six traits. Several clusters of associated markers were distributed along the genome and highlighted important genomic associations. The identification of consistent MTAs that are stable across different environments is of great value to MAS in breeding genotypes adapted to diverse ecological environments. Five of these candidates were also pleiotropic and co-localized with different highly phenotypic correlated traits (i.e., SH, SP, and SPL). Pleiotropic effects are also beneficial in the breeding process as they allow breeders to simultaneously select for multiple traits. Gene annotation showed that the highest percentages of candidates identified have unknown functions or were not found in the current version of the faba bean genome. The detection of new gene sequences in very large and highly repetitive genomes such as that of faba bean remains a significant challenge due to the presence of significant gaps, highlighting the need for a high-quality genome assembly. Even though these candidates or the closely linked genes flanking the significant SNPs remain to be validated in different

genetic backgrounds, the identified markers provide a valuable genetic resource for future marker-assisted selection and fine mapping of the genes underlying yield improvement in this crop.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author/s.

Author contributions

NG: Conceptualization, Formal Analysis, Resources, Writing – original draft, Methodology. MP: Formal Analysis, Software, Supervision, Writing – review & editing. IS: Data curation, Visualization, Writing – review & editing. DS: Data curation, Visualization, Writing – review & editing. DL: Data curation, Visualization, Writing – review & editing. CH: Data curation, Visualization, Writing – review & editing. AT: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project has received funding from the European Union's Horizon 2020 Programme for Research and Innovation under grant agreement n° 727312 (EUCLEG Project), PID2020-114952RR-I00 and PR.AVA23.INV2023.009 co-financed by ERDF.

Acknowledgments

We thank the seed companies (NORGEN, Boreal, and NPZ) and the institutes that provided us with faba bean seeds from their accessions: Universidad de Göttingen (Germany), IFVCNS (Institute of Field and Vegetable Crops, Serbia), Research Institute for Fodder Crops (Czech Republic), University of Ghent (Belgium), INRA Dijon (Jean-Bernard Magnin-Robert), ICARDA (Fouad Maalouf), and IFAPA. The technical staff of IFAPA, "centro Alameda del Obispo" is thanked for their deep investment to multiply accessions and collect leaves for genotyping. Grace Gay and Ellen Sizer Coverdale are thanked for their hard work in phenotyping the field trial in Aberystwyth. The authors are grateful to Dragan Milic and Miroslav Zoric for setting up the experimental augmented design.

Conflict of interest

Author IS was employed by the company Agrovegetal S.A. Agrovegetal S.A. was a partner of the EUCLEG project.

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 handling editor ÅE declared a past collaboration with the author MP.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1328690/full#supplementary-material

References

Abdurakhmonov, I. Y., Kohel, R. J., Yu, J. Z., Pepper, A. E., Abdullaev, A. A., Kushanov, F. N., et al. (2008). Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* 92, 478–487. doi: 10.1016/j.ygeno.2008.07.013

Abou-Khater, L., Maalouf, F., Jighly, A., Rubiales, D., and Kumar, S. (2022). Adaptability and Stability of Faba Bean (*Vicia faba* L.) Accessions under Diverse Environments and Herbicide Treatments. *Plants* 11 (3), 251. doi: 10.3390/plants11030251

Adhikari, K. N., Khazaei, H., Ghaouti, L., Maalouf, F., Vandenberg, A., Link, W., et al. (2021). Conventional and molecular breeding tools for accelerating genetic gain in faba bean (*Vicia faba* L.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.744259

Ali, M. B. M., Welna, G. C., Sallam, A., Martsch, R., Balko, C., Gebser, B., et al. (2016). Association analyses to genetically improve drought and freezing tolerance of Faba bean (*Vicia faba* L.). *Crop Sci.* 56, 1036–1048. doi: 10.2135/cropsci2015.08.0503

Ávila, C. M., Ruiz-Rodríguez, M. D., Cruz-Izquierdo, S., Atienza, S. G., Cubero, J. I., and Torres, A. M. (2017). Identification of plant architecture and yield-related QTL in *Vicia faba* L. *Mol. Breed.* 37, 88. doi: 10.1007/s11032-017-0688-7

Avila, C. M., Satovic, Z., Sillero, J. C., Rubiales, D., Moreno, M. T., and Torres, A. M. (2005). QTL detection for agronomic traits in faba bean (*Vicia faba* L.). *Agriculturae Conspectus Scientificus* 70, 65–73.

Bacete, L., and Hamann, T. (2020). The role of mechanoperception in plant cell wall integrity maintenance. *Plants* 9 (5), 574. doi: 10.3390/plants9050574

Badke, Y. M., Bates, R. O., Ernst, C. W., Fix, J., and Steibel, J. P. (2014). Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. *G3 (Bethesda)* 4, 623–631. doi: 10.1534/g3.114.010504

Baez, L. A., Tichá, T., and Hamann, T. (2022). Cell wall integrity regulation across plant species. *Plant Mol. Biol.* 109, 483–504. doi: 10.1007/s11103-022-01284-7

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 289, 905–920. doi: 10.1126/science.289.5481.905

Beji, S., Fontaine, V., Devaux, R., Thomas, M., Negro, S. S., Bahrman, N., et al. (2020). Genome-wide association study identifies favorable SNP alleles and candidate genes for frost tolerance in pea. *BMC Genomics* 21, 536. doi: 10.1186/s12864-020-06928-w

Bekele, F. L., Bidaisee, G. G., Allegre, M., Argout, X., Fouet, O., Boccara, M., et al. (2022). Genome-wide association studies and genomic selection assays made in a large sample of cacao (*Theobroma cacao* L.) germplasm reveal significant marker-trait associations and good predictive value for improving yield potential. *PloS One* 17, e0260907. doi: 10.1371/journal.pone.0260907

Bhatnagar, N., Min, M.-K., Choi, E.-H., Kim, N., Moon, S.-J., Yoon, I., et al. (2017). The protein phosphatase 2C clade A protein OsPP2C51 positively regulates seed germination by directly inactivating OsbZIP10. *Plant Mol. Biol.* 93, 389–401. doi: 10.1007/s11103-016-0568-2

Block-Schmidt, A. S., Dukowic-Schulze, S., Wanieck, K., Reidt, W., and Puchta, H. (2011). BRCC36A is epistatic to BRCA1 in DNA crosslink repair and homologous recombination in Arabidopsis thaliana. *Nucleic Acids Res.* 39, 146–154. doi: 10.1093/nar/gkq722

Bonhomme, M., André, O., Badis, Y., Ronfort, J., Burgarella, C., Chantret, N., et al. (2014). High-density genome-wide association mapping implicates an F-box encoding gene in *Medicago truncatula* resistance to *Aphanomyces euteiches*. *New Phytol.* 201, 1328–1342. doi: 10.1111/nph.12611

Bonnafous, F., Duhnen, A., Gody, L., Guillaume, O., Mangin, B., Pegot-Espagnet, P., et al. (2019). *mlmm.gwas: Pipeline for GWAS Using MLMM* (CRAN). Available at: https://CRAN.R-project.org/package=mlmm.gwas.

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

Calus, M. P. L., 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

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Cantet, R. J. C., Birchmeier, A. N., Canaza Cayo, A. W., and Fioretti, C. (2005). Semiparametric animal models *via* penalized splines as alternatives to models with contemporary groups. *J. Anim. Sci.* 83, 2482–2494. doi: 10.2527/2005.83112482x

Cao, J., Li, G., Qu, D., Li, X., and Wang, Y. (2020). Into the seed: auxin controls seed development and grain yield. *Int. J. Mol. Sci.* 21 (5), 1662. doi: 10.3390/ijms21051662

Cappa, E. P., and Cantet, R. J. C. (2007). Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model. *Can. J. For. Res.* 37, 2677–2688. doi: 10.1139/X07-116

Cappa, E. P., Muñoz, F., Sanchez, L., and Cantet, R. J. C. (2015). A novel individual-tree mixed model to account for competition and environmental heterogeneity: a Bayesian approach. *Tree Genet. Genomes* 11, 120. doi: 10.1007/s11295-015-0917-3

Carranca, C., Varennes de, A., and Rolston, D. (1999). Biological nitrogen fixation by faba bean, pea and chickpea, under field conditions, estimated by the 15N isotope dilution technique. *Eur. J. Agron.* 10, 49–56. doi: 10.1016/S1161-0301(98)00049-5

Carroll, R. K., Rivera, F. E., Cavaco, C. K., Johnson, G. M., Martin, D., and Shaw, L. N. (2014). The lone S41 family C-terminal processing protease in Staphylococcus aureus is localized to the cell wall and contributes to virulence. *Microbiol. (Reading Engl)* 160, 1737–1748. doi: 10.1099/mic.0.079798-0

Cernay, C., Pelzer, E., and Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Sci. Data* 3, 160084. doi: 10.1038/sdata.2016.84

Cevher, M. A., Zhang, X., Fernandez, S., Kim, S., Baquero, J., Nilsson, P., et al. (2010). Nuclear deadenylation/polyadenylation factors regulate 3' processing in response to DNA damage. *EMBO J.* 29, 1674–1687. doi: 10.1038/emboj.2010.59

Cruz-Izquierdo, S., Avila, C. M., Satovic, Z., Palomino, C., Gutierrez, N., Ellwood, S. R., et al. (2012). Comparative genomics to bridge *Vicia faba* with model and closely-related legume species: stability of QTLs for flowering and yield-related traits. *Theor. Appl. Genet.* 125, 1767–1782. doi: 10.1007/s00122-012-1952-1

Cubero, J. I. (1973). Evolutionary trends in *Vicia faba*. *Theor. Appl. Genet.* 43, 59–65. doi: 10.1007/BF00274958

Das, S., Singh, M., Srivastava, R., Bajaj, D., Saxena, M. S., Rana, J. C., et al. (2016). mQTL-seq delineates functionally relevant candidate gene harbouring a major QTL regulating pod number in chickpea. *DNA Res.* 23, 53–65. doi: 10.1093/dnares/dsv036

Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., et al. (2022). Natural allelic variation of GmST05 controlling seed size and quality in soybean. *Plant Biotechnol. J.* 20, 1807–1818. doi: 10.1111/pbi.13865

Duplan, V., and Rivas, S. (2014). E3 ubiquitin-ligases and their target proteins during the regulation of plant innate immunity. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00042

FAOSTAT (2020) FAOSTAT. Available online at: http://www.fao.org/faostat/en/#data (Accessed June 17, 2021).

Faridi, R., Koopman, B., Schierholt, A., Ali, M. B., Apel, S., and Link, W. (2021). Genetic study of the resistance of faba bean (*Vicia faba*) against the fungus *Ascochyta fabae* through a genome-wide association analysis. *Plant Breed.* 140, 442–452. doi: 10.1111/pbr.12918

Gongora, C. E., and Broadway, R. M. (2002). Plant growth and development influenced by transgenic insertion of bacterial chitinolytic enzymes. *Mol. Breeding.* 9, 123–135. doi: 10.1023/A:1026732124713

Govindaraj, M., Vetriventhan, M., and Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet. Res. Int.* 2015, 431487. doi: 10.1155/2015/431487

Gutiérrez, N., Pégard, M., Balko, C., and Torres, A. M. (2023). Genome-wide association analysis for drought tolerance and associated traits in faba bean (*Vicia faba* L.). *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1091875

Huang, J., Lu, G., Liu, L., Raihan, M. S., Xu, J., Jian, L., et al. (2020). The Kernel Size-Related Quantitative Trait Locus qKW9 Encodes a Pentatricopeptide Repeat Protein That affects Photosynthesis and Grain Filling. *Plant Physiol.* 183, 1696–1709. doi: 10.1104/pp.20.00374

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Hwang, E.-Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15, 1. doi: 10.1186/1471-2164-15-1

Ishimaru, K., Hirotsu, N., Kashiwagi, T., Madoka, Y., Nagasuga, K., Ono, K., et al. (2008). Overexpression of a maize SPS gene improves yield characters of potato under field conditions. *Plant Prod. Sci.* 11, 104–107. doi: 10.1626/pps.11.104

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednář, P., et al. (2023). The giant diploid faba genome unlocks variation in a global protein crop. *Nature* 615, 652–659. doi: 10.1038/s41586-023-05791-5

Jeong, J. S., Kim, Y. S., Baek, K. H., Jung, H., Ha, S.-H., Do Choi, Y., et al. (2010). Root-specific expression of OsNAC10 improves drought tolerance and grain yield in rice under field drought conditions. *Plant Physiol.* 153, 185–197. doi: 10.1104/pp.110.154773

Ju, Y., Guo, L., Cai, Q., Ma, F., Zhu, Q.-Y., Zhang, Q., et al. (2016). Arabidopsis JINGUBANG is a negative regulator of pollen germination that prevents pollination in moist environments. *Plant Cell* 28, 2131–2146. doi: 10.1105/tpc.16.00401

Kaler, A. S., and Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics* 20, 618. doi: 10.1186/s12864-019-5992-7

Kassambara, A., and Mundt, F. (2020). Factoextra: extract and visualize the results of multivariate data analyses. Available at: https://CRAN.R-project.org/package=factoextra.

Kaur, D., Held, M. A., Smith, M. R., and Showalter, A. M. (2021). Functional characterization of hydroxyproline-O-galactosyltransferases for Arabidopsis arabinogalactan-protein synthesis. *BMC Plant Biol.* 21, 590. doi: 10.1186/s12870-021-03362-2

Keeble-Gagnère, G., Isdale, D., Suchecki, R., Kruger, A., Lomas, K., Carroll, D., et al. (2019). Integrating past, present and future wheat research with Pretzel. *BioRxiv*. doi: 10.1101/517953

Khan, H., Krishnappa, G., Kumar, S., Mishra, C. N., Krishna, H., Devate, N. B., et al. (2022). Genome-wide association study for grain yield and component traits in bread wheat (*Triticum aestivum* L.). *Front. Genet.* 13. doi: 10.3389/fgene.2022.982589

Khazaei, H., O'Sullivan, D. M., Stoddard, F. L., Adhikari, K. N., Paull, J. G., Schulman, A. H., et al. (2021). Recent advances in faba bean genetic and genomic tools for crop improvement. *Legume Sci.* 3, e75. doi: 10.1002/leg3.75

Kim, W., Lee, Y., Park, J., Lee, N., and Choi, G. (2013). HONSU, a protein phosphatase 2C, regulates seed dormancy by inhibiting ABA signaling in Arabidopsis. *Plant Cell Physiol.* 54, 555–572. doi: 10.1093/pcp/pct017

Klein, D. J., Moore, P. B., and Steitz, T. A. (2004). The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* 340, 141–177. doi: 10.1016/j.jmb.2004.03.076

Köpke, U., and Nemecek, T. (2010). Ecological services of faba bean. *Field Crops Res.* 115, 217–233. doi: 10.1016/j.fcr.2009.10.012

Li, W., Chen, C., Markmann-Mulisch, U., Timofejeva, L., Schmelzer, E., Ma, H., et al. (2004). The Arabidopsis AtRAD51 gene is dispensable for vegetative development but required for meiosis. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10596–10601. doi: 10.1073/pnas.0404110101

Li, N., and Li, Y. (2014). Ubiquitin-mediated control of seed size in plants. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00332

Li, J., Zhang, Y., Ma, R., Huang, W., Hou, J., Fang, C., et al. (2022). Identification of ST1 reveals a selection involving hitchhiking of seed morphology and oil content during soybean domestication. *Plant Biotechnol. J.* 20, 1110–1121. doi: 10.1111/pbi.13791

Li, Z., Zhang, X., Zhao, K., Zhao, K., Qu, C., Gao, G., et al. (2021). Comprehensive transcriptome analyses reveal candidate genes for variation in seed size/weight during peanut (*Arachis hypogaea* L.) domestication. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.666483

Lin, C. S., and Poushinsky, G. (1983). A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. *Biometrics* 39, 553–561. doi: 10.2307/2531083

Linden, K. J., and Callis, J. (2020). The ubiquitin system affects agronomic plant traits. *J. Biol. Chem.* 295, 13940–13955. doi: 10.1074/jbc.REV120.011303

Liu, P.-L., Du, L., Huang, Y., Gao, S.-M., and Yu, M. (2017). Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. *BMC Evol. Biol.* 17, 47. doi: 10.1186/s12862-017-0891-5

Liu, Y.-J., Xiu, Z.-H., Meeley, R., and Tan, B.-C. (2013). Empty pericarp5 encodes a pentatricopeptide repeat protein that is required for mitochondrial RNA editing and seed development in maize. *Plant Cell* 25, 868–883. doi: 10.1105/tpc.112.106781

Lopez-Bellido, L., Lopez-Bellido, R. J., Redondo, R., and Benítez, J. (2006). Faba bean nitrogen fixation in a wheat-based rotation under rainfed Mediterranean conditions: Effect of tillage system. *Field Crop Res.* 98, 253–260. doi: 10.1016/j.fcr.2006.03.001

Maalouf, F., Abou-Khater, L., Babiker, Z., Jighly, A., Alsamman, A. M., Hu, J., et al. (2022). Genetic dissection of heat stress tolerance in faba bean (*Vicia faba* L.) using GWAS. *Plants* 11 (9), 1108. doi: 10.3390/plants11091108

Mangini, G., Gadaleta, A., Colasuonno, P., Marcotuli, I., Signorile, A. M., Simeone, R., et al. (2018). Genetic dissection of the relationships between grain yield components by genome-wide association mapping in a collection of tetraploid wheats. *PloS One* 13, e0190162. doi: 10.1371/journal.pone.0190162

Mansilla, N., Racca, S., Gras, D. E., Gonzalez, D. H., and Welchen, E. (2018). The complexity of mitochondrial complex IV: an update of cytochrome c oxidase biogenesis in plants. *Int. J. Mol. Sci.* 19 (3), 662. doi: 10.3390/ijms19030662

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, GY., and Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 (Bethesda)* 5 (11), 2383–90. doi: 10.1534/g3.115.021667

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. doi: 10.1093/genetics/157.4.1819

Mulyatama, R. A., Neliana, I. R., Sawitri, W. D., Sakakibara, H., Kim, K.-M., and Sugiharto, B. (2022). Increasing the activity of sugarcane sucrose phosphate synthase enhanced growth and grain yields in transgenic indica rice. *Agronomy* 12, 2949. doi: 10.3390/agronomy12122949

Muñoz, F., and Sanchez, L. (2020) BreedR: Statistical Methods for Forest Genetic Resources Analysts. Available online at: https://github.com/famuvie/breedR.

Naran, R., Pierce, M. L., and Mort, A. J. (2007). Detection and identification of rhamnogalacturonan lyase activity in intercellular spaces of expanding cotton cotyledons. *Plant J.* 50, 95–107. doi: 10.1111/j.1365-313X.2007.03033.x

Ochoa-Jiménez, V.-A., Berumen-Varela, G., Burgara-Estrella, A., Orozco-Avitia, J.-A., Ojeda-Contreras, Á.-J., Trillo-Hernández, E.-A., et al. (2018). Functional analysis of tomato rhamnogalacturonan lyase gene Solyc11g011300 during fruit development and ripening. *J. Plant Physiol.* 231, 31–40. doi: 10.1016/j.jplph.2018.09.001

Ogutcen, E., Pandey, A., Khan, M. K., Marques, E., Penmetsa, R. V., Kahraman, A., et al. (2018). Pod shattering: A homologous series of variation underlying domestication and an avenue for crop improvement. *Agronomy* 8, 137. doi: 10.3390/agronomy8080137

O'Sullivan, D. M., Angra, D., Harvie, T., Tagkouli, V., and Warsame, A. (2019). "A genetic toolbox for *Vicia faba* improvement," in *International Conference on Legume Genetics and Genomics (ICLGG)* (Dijon) pp. 157.

Pan, J., Li, Z., Wang, Q., Garrell, A. K., Liu, M., Guan, Y., et al. (2018). Comparative proteomic investigation of drought responses in foxtail millet. *BMC Plant Biol.* 18, 315. doi: 10.1186/s12870-018-1533-9

Park, H.-Y., Lee, H. T., Lee, J. H., and Kim, J.-K. (2019). Arabidopsis U2AF65 regulates flowering time and the growth of pollen tubes. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00569

Pégard, M., Barre, P., Delaunay, S., Surault, F., Karagić, D., Milić, D., et al. (2023). Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1196134

Porter, K., and Day, B. (2016). From filaments to function: The role of the plant actin cytoskeleton in pathogen perception, signaling and immunity. *J. Integr. Plant Biol.* 58, 299–311. doi: 10.1111/jipb.12445

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

Raggi, L., Caproni, L., Carboni, A., and Negri, V. (2019). Genome-wide association study reveals candidate genes for flowering time variation in common bean (Phaseolus vulgaris L.). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00962

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ramsay, G., Van de Ven, W., Waugh, R., Griffiths, D. W., and Powell, W. (1995). "Mapping Quantitative trait loci in faba beans," in *Improving production and utilization of grain legumes. Second European Conference on Grain Legumes*(Copenhagen, Denmark), 444–445.

R Core Team (2022). *A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing).

Saha, A., Das, S., Moin, M., Dutta, M., Bakshi, A., Madhav, M. S., et al. (2017). Genome-wide identification and comprehensive expression profiling of ribosomal protein small subunit (RPS) genes and their comparative analysis with the large subunit (RPL) genes in rice. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01553

Sallam, A., Arbaoui, M., El-Esawi, M., Abshire, N., and Martsch, R. (2016). Identification and verification of QTL associated with frost tolerance using linkage mapping and GWAS in winter faba bean. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01098

Sami, A. A., Arabia, S., Sarker, R. H., and Islam, T. (2021). Deciphering the role of helicases and translocases: A multifunctional gene family safeguarding plants from diverse environmental adversities. *Curr. Plant Biol.* 26, 100204. doi: 10.1016/j.cpb.2021.100204

Schneider, K., Mathur, J., Boudonck, K., Wells, B., Dolan, L., and Roberts, K. (1998). The ROOT HAIRLESS 1 gene encodes a nuclear protein required for root hair initiation in Arabidopsis. *Genes Dev.* 12, 2013–2021. doi: 10.1101/gad.12.13.2013

Schwenke, G. D., Peoples, M. B., Turner, G. L., and Herridge, D. F. (1998). Does nitrogen fixation of commercial, dryland chickpea and faba bean crops in north-west New South Wales maintain or enhance soil nitrogen? *Aust. J. Exp. Agric.* 38, 61–70. doi: 10.1071/EA97078

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

Sharkey, T. D., Laporte, M. M., and Kruger, E. L. (2000). "Will increased photosynthetic efficiency lead to increased yield in rice? *Stud. Plant Sci.* 7, 73–86. doi: 10.1016/S0928-3420(00)80007-5

Shirley, N. J., Aubert, M. K., Wilkinson, L. G., Bird, D. C., Lora, J., Yang, X., et al. (2019). Translating auxin responses into ovules, seeds and yield: Insight from Arabidopsis and the cereals. *J. Integr. Plant Biol.* 61, 310–336. doi: 10.1111/jipb.12747

Singh, L., Dhillon, G. S., Kaur, S., Dhaliwal, S. K., Kaur, A., Malik, P., et al. (2022). Genome-wide association study for yield and yield-related traits in diverse blackgram panel (Vigna mungo L. Hepper) reveals novel putative alleles for future breeding programs. *Front. Genet.* 13. doi: 10.3389/fgene.2022.849016

Skovbjerg, C. K., Angra, D., Robertson-Shersby-Harvie, T., Kreplak, J., Keeble-Gagnère, G., Kaur, S., et al. (2023). Genetic analysis of global faba bean diversity, agronomic traits and selection signatures. *Theor. Appl. Genet.* 136, 114. doi: 10.1007/s00122-023-04360-8

Sugimoto-Shirasu, K., Roberts, G. R., Stacey, N. J., McCann, M. C., Maxwell, A., and Roberts, K. (2005). RHL1 is an essential component of the plant DNA topoisomerase VI complex and is required for ploidy-dependent cell growth. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18736–18741. doi: 10.1073/pnas.0505883102

Szymanski, D., and Staiger, C. J. (2018). The actin cytoskeleton: functional arrays for cytoplasmic organization and cell shape control. *Plant Physiol.* 176, 106–118. doi: 10.1104/pp.17.01519

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Tang, H., Cuevas, H. E., Das, S., Sezen, U. U., Zhou, C., Guo, H., et al. (2013). Seed shattering in a wild sorghum is conferred by a locus unrelated to domestication. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15824–15829. doi: 10.1073/pnas.1305213110

Toker, C. (2004). Estimates of broad-sense heritability for seed yield and yield criteria in faba bean (*Vicia faba* L.). *Hereditas* 140, 222–225. doi: 10.1111/j.1601-5223.2004.01780.x

Umezawa, T., Sugiyama, N., Mizoguchi, M., Hayashi, S., Myouga, F., Yamaguchi-Shinozaki, K., et al. (2009). Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 17588–17593. doi: 10.1073/pnas.0907095106

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

Varshney, R. K., Saxena, R. K., Upadhyaya, H. D., Khan, A. W., Yu, Y., Kim, C., et al. (2017). Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* 49, 1082–1088. doi: 10.1038/ng.3872

Varshney, R. K., Thudi, M., Roorkiwal, M., He, W., Upadhyaya, H. D., Yang, W., et al. (2019). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat. Genet.* 51, 857–864. doi: 10.1038/s41588-019-0401-3

Vicente, A. R., Saladié, M., Rose, J. K., and Labavitch, J. M. (2007). The linkage between cell wall metabolism and fruit softening: looking to the future. *J. Sci. Food Agric.* 87, 1435–1448. doi: 10.1002/jsfa.2837

Wang, S., Durrant, W. E., Song, J., Spivey, N. W., and Dong, X. (2010). Arabidopsis BRCA2 and RAD51 proteins are specifically involved in defense gene transcription during plant immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22716–22721. doi: 10.1073/pnas.1005978107

Warsame, A. O., O'Sullivan, D. M., and Tosi, P. (2018). Seed storage proteins of Faba bean (*Vicia faba* L): current status and prospects for genetic improvement. *J. Agric. Food Chem.* 66, 12617–12626. doi: 10.1021/acs.jafc.8b04992

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genet. Res.* 75, 249–252. doi: 10.1017/S0016672399004462

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Editors: R. Gentleman, K. Hornik and G. Parmigiani. Springer. doi: 10.1007/978-3-319-24277-4

Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., et al. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* 38, 4039–4042. doi: 10.1093/molbev/msab166

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., et al. (2022). *Ggtree* : A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta*. doi: 10.1002/imt2.56

Zanotto, S., Ruttink, T., Pégard, M., Skøt, L., Grieder, C., Kölliker, R., et al. (2023). A genome-wide association study of freezing tolerance in red clover (*Trifolium pratense* L.) germplasm of European origin. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1189662

Zhang, H., Liu, Y., Zong, X., Teng, C., Hou, W., Li, P., et al. (2023). Genetic diversity of global faba bean germplasm resources based on the 130K TNGS genotyping platform. *Agronomy* 13, 811. doi: 10.3390/agronomy13030811

Zhou, L., Cheung, M.-Y., Zhang, Q., Lei, C.-L., Zhang, S.-H., Sun, S. S.-M., et al. (2009). A novel simple extracellular leucine-rich repeat (eLRR) domain protein from rice (OsLRR1) enters the endosomal pathway and interacts with the hypersensitive-induced reaction protein 1 (OsHIR1). *Plant Cell Environ.* 32, 1804–1820. doi: 10.1111/j.1365-3040.2009.02039.x

Zhou, X., Guo, J., Pandey, M. K., Varshney, R. K., Huang, L., Luo, H., et al. (2021). Dissection of the genetic basis of yield-related traits in the Chinese peanut mini-core collection through genome-wide association studies. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.637284

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 June 2024
DOI 10.3389/fpls.2024.1407609

Check for updates

OPEN ACCESS

EDITED BY
Ali M. Missaoui,
University of Georgia, United States

REVIEWED BY
Kaye Enid Basford,
The University of Queensland, Australia
João Ricardo Bachega Feijó Rosa,
RB Genetics & Statistics Consulting, Brazil

*CORRESPONDENCE
Leif Skøt
lfs@aber.ac.uk
Roland Kölliker
roland.koelliker@usys.ethz.ch

RECEIVED 27 March 2024
ACCEPTED 20 May 2024
PUBLISHED 10 June 2024

CITATION
Skøt L, Nay MM, Grieder C, Frey LA, Pégard M, Öhlund L, Amdahl H, Radovic J, Jaluvka L, Palmé A, Ruttink T, Lloyd D, Howarth CJ and Kölliker R (2024) Including marker x environment interactions improves genomic prediction in red clover (*Trifolium pratense* L.). *Front. Plant Sci.* 15:1407609. doi: 10.3389/fpls.2024.1407609

COPYRIGHT
© 2024 Skøt, Nay, Grieder, Frey, Pégard, Öhlund, Amdahl, Radovic, Jaluvka, Palmé, Ruttink, Lloyd, Howarth and Kölliker. 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.

Including marker x environment interactions improves genomic prediction in red clover (*Trifolium pratense* L.)

Leif Skøt[1*], Michelle M. Nay[2], Christoph Grieder[2], Lea A. Frey[3], Marie Pégard[4], Linda Öhlund[5], Helga Amdahl[6], Jasmina Radovic[7], Libor Jaluvka[8], Anna Palmé[9], Tom Ruttink[10,11], David Lloyd[12], Catherine J. Howarth[1] and Roland Kölliker[3*]

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom, [2]Division of Plant Breeding, Fodder Plant Breeding, Agroscope, Zurich, Switzerland, [3]Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland, [4]INRAE P3F, Lusignan, France, [5]Lantmännen Lantbruk, Svalöv, Sweden, [6]Graminor Breeding Ltd., Bjørke Forsøksgård, Norway, [7]Institute for Forage Crops (IKBKS), Kruševac, Serbia, [8]DLF Seeds, Hladké Životice, Czechia, [9]The Nordic Genetic Resource Centre, Plant Section, Alnarp, Sweden, [10]Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food, Melle, Belgium, [11]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [12]Germinal Horizon, Plas Gogerddan, Aberystwyth, United Kingdom

Genomic prediction has mostly been used in single environment contexts, largely ignoring genotype x environment interaction, which greatly affects the performance of plants. However, in the last decade, prediction models including marker x environment (MxE) interaction have been developed. We evaluated the potential of genomic prediction in red clover (*Trifolium pratense* L.) using field trial data from five European locations, obtained in the Horizon 2020 EUCLEG project. Three models were compared: (1) single environment (SingleEnv), (2) across environment (AcrossEnv), (3) marker x environment interaction (MxE). Annual dry matter yield (DMY) gave the highest predictive ability (PA). Joint analyses of DMY from years 1 and 2 from each location varied from 0.87 in Britain and Switzerland in year 1, to 0.40 in Serbia in year 2. Overall, crude protein (CP) was predicted poorly. PAs for date of flowering (DOF), however ranged from 0.87 to 0.67 for Britain and Switzerland, respectively. Across the three traits, the MxE model performed best and the AcrossEnv worst, demonstrating that including marker x environment effects can improve genomic prediction in red clover. Leaving out accessions from specific regions or from specific breeders' material in the cross validation tended to reduce PA, but the magnitude of reduction depended on trait, region and breeders' material, indicating that population structure contributed to the high PAs observed for DMY and DOF. Testing the genomic estimated breeding values on new phenotypic data from Sweden showed that DMY training data from Britain gave high PAs in both years (0.43–0.76), while DMY training data from Switzerland gave high PAs only for year 1

10.3389/fpls.2024.1407609

(0.70–0.87). The genomic predictions we report here underline the potential benefits of incorporating MxE interaction in multi-environment trials and could have perspectives for identifying markers with effects that are stable across environments, and markers with environment-specific effects.

KEYWORDS

genomic prediction, marker x environment interaction, population structure, predictive ability, red clover, *trifolium pratense*

1 Introduction

Increased use of high protein legume crops, such as red clover as a forage for ruminants, can aid in reducing the protein deficit in Europe and contribute to sustainable livestock production. Red clover (*Trifolium pratense* L.) is particularly valued in livestock agriculture for its ability to fix atmospheric nitrogen in symbiosis with soil bacteria of the genus *Rhizobium*, reducing the reliance on chemically produced fertiliser. It can enrich the soil and provide companion crops with a supply of fixed nitrogen (Frame et al., 1998). There is also evidence that red clover can improve soil structure and can be a useful component in crop rotations (McKenna et al., 2018). Red clover contains the enzyme polyphenol oxidase (PPO), which catalyses the formation of compounds that can form complexes with and moderate the breakdown of protein in the rumen. This can improve utilisation of nitrogen by the ruminant animals, thereby reducing losses of nitrogen and emissions of methane to the environment (Lee, 2014).

Ready access to chemically produced nitrogen fertiliser, challenges in maintaining the optimal ratio of clover to grass in mixtures, lack of persistency, and insufficient tolerance to grazing still limits the use of red clover in many countries (Lüscher et al., 2014). Additionally, there is an urgent need to produce varieties that are more resilient to climate change.

Genomic selection (GS) is one of the most promising methods of increasing the speed of new plant variety development. The main objectives of this work was to investigate the potential of GS in red clover by comparing different prediction models. The term GS was introduced by Meuwissen et al. (2001) and is based on the following principle: A training population, for which genome-wide molecular marker and phenotypic data are available, is used to estimate the effect of each marker on each phenotype. This information is then used in a test population with only the molecular marker information available to determine a genomic estimated breeding value (GEBV), which in turn is used to select individuals for further crossing in a breeding programme. GS or genomic prediction (GP) has shown its utility particularly in dairy cattle breeding (Schefers and Weigel, 2012; Hayes et al., 2013b; Wiggans et al., 2017). It is also being incorporated in plant breeding, notably in the major cereals such as maize (*Zea mays* L.) (Zhao et al., 2012), rice (*Oryza sativa* L.) (Cui et al., 2020; Xu et al., 2021), wheat (*Triticum* spp.) (Bassi et al., 2016; Juliana et al., 2020), barley (*Hordeum vulgare* L.) (Puglisi et al., 2021) and oats (*Avena sativa* L.) (Campbell et al., 2021).

Among forage crops, GP has been most intensively studied in perennial ryegrass (*Lolium perenne* L.) (Hayes et al., 2013a; Forster et al., 2014; Fè et al., 2015, 2016; Grinberg et al., 2016; Lin et al., 2016; Byrne et al., 2017; Lin et al., 2017; Arojju et al., 2018; Cericola et al., 2018; Faville et al., 2018; Pembleton et al., 2018; Arojju et al., 2020; Faville et al., 2020; Barre et al., 2022). Many of those studies indicate that predictive abilities (PAs) are highest for traits with high heritability such as heading date, and disease resistance, while PAs for DMY were moderate to low.

Genomic prediction in forage legumes has received less attention than in ryegrass. However, some work has been carried out on DMY in alfalfa (*Medicago sativa* L.) (Annicchiarico et al., 2015; Li et al., 2015; Annicchiarico et al., 2017), and white clover (*T. repens* L.) (Griffiths et al., 2022) with encouraging results. However, the perennialism and often outbreeding nature of most forage crops present some challenges. Varieties are usually synthetic populations derived from crossing a small number of parents after they have been selected among and/or within families based on observation of performance of their half-sib progeny (Barre et al., 2022). Either way, the varieties are populations consisting of related, but genetically distinct and heterozygous individuals. Genomic prediction models are thus often based on genotypes of the parents, and phenotypes of their half-sib progeny (Annicchiarico et al., 2015; Grinberg et al., 2016; Arojju et al., 2020), or based on genotypes and phenotypes of spaced plants for suitable traits (Byrne et al., 2017; Arojju et al., 2018).

An alternative is to genotype pools of plants at the level of family or population using methods such as Genotyping-By-Sequencing (GBS) (Elshire et al., 2011) or RAD-seq (Baird et al., 2008). Allele frequency data on a population or family level allows predictions to be made based on genotypic and phenotypic data from the same population, not based on genotypic data from parents and phenotypic data from their progenies. It has been used in ryegrass for prediction of flowering time, crown rust resistance, seed yield and fructan content (Byrne et al., 2013; Fè et al., 2016; Cericola et al., 2018). Recently, GP was also successfully explored in alfalfa using pooled samples for genotyping (Pégard et al., 2023). PA was above 0.75 for plant height and dormancy in some years, demonstrating the potential of GP for some traits.

Pooled genotyping has been used in red clover to identify regions under selection for survival in the field (Ergon et al., 2019), and for genome-wide identification of loci involved in timing of stem elongation and freezing tolerance (Ergon et al., 2022; Zanotto et al., 2023). Allele frequency data from pooled samples compared well with those obtained by genotyping individuals from the same population (Ergon et al., 2019).

GP has mostly been used in single environment contexts, largely ignoring genotype by environment interaction (GxE), i.e. effects of location and year, that greatly affect crop performance. However, in the last decade, models incorporating GxE have been described (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Lopez-Cruz et al., 2015). While Juliana et al. (2020) reported no advantage of using GxE models, many other studies found that it increased PA, e.g. in wheat (Lopez-Cruz et al., 2015; Crossa et al., 2016; Sukumaran et al., 2018), maize (Bandeira E Sousa et al., 2017), barley (Puglisi et al., 2021) and sesame (*Sesamum indicum* L.) (Sabag et al., 2023).

This work was aimed at investigating whether incorporating marker by environment interaction effects (MxE) in the prediction models could be advantageous, compared to analysis trait by trait. The focus was not on GP in a breeding programme. More specifically the objective was to assess GP for DMY, CP content and DOF in a diverse panel of red clover accessions using different prediction models. GxE was incorporated using the strategy described by Lopez-Cruz et al. (2015), as this often performed best in the investigations described above. Three models were compared: (1) SingleEnv, where PA was measured separately for each environment, (2) AcrossEnv, where marker effects were an average across environments, and (3) MxE where marker effects were divided into those that were stable across environments, and those that were environment-specific. Finally, we attempted to validate the models by predicting phenotypes derived from independent experiments.

2 Materials and methods

2.1 Plant materials

The data on which much of this work is based were obtained through the EUCLEG project (Horizon 2020 Programme for Research & Innovation, grant agreement no. 727312; http://www.eucleg.eu) and have been described in detail previously (Nay et al., 2023). Briefly, field experiments with 400 accessions were established in Switzerland (CHE: 47.480°N, 8.904°E) and in Czechia (CZE: 49.690°N, 17.960°E) in a partially replicated (p-rep) design. The field experiment was abandoned after year 1 in CZE, so no data from year 2 were available from this location. Field experiments with approximately 100 accessions were established in Britain (GBR: 52.427°N, 4.020°W) with 100, Norway (NOR: 60.757°N, 11.203°E) with 109, and Serbia (SRB: 43.583°N, 21.206°E) with 100. For the latter three locations, twenty accessions were included at every site, and between 12 and 17 of the remaining accessions overlapped between pairs of locations (Supplementary Table S1). From all five locations annual DMY from year 1 and 2 (DMY1 and DMY2), CP content from cut 1 and cut 2 in year 1 (CP1 and CP2), and date of flowering (DOF) from year 1 were used here. Supplementary Table S1 contains more detailed information about the accessions used. A p-rep design was also used in GBR. For the p-rep trials there were no complete blocks, so that each row and column was an incomplete block (IB1 and IB2, respectively). The trials from NOR and SRB were Alpha designs with two complete blocks containing all the accessions. Observation rows were used to record DOF with two complete blocks at all sites.

To test the GEBVs derived from the field trials described above, we used data obtained from field trials at three additional sites, all in Sweden. A total of 49 accessions were included in these field trials in which a randomised complete block design (RCBD) was used with two replicates (blocks) at Bjertorp (BJT: 58.250°N, 13.117°E) and Kölbäck (KLB: 58.433°N, 15.250°E), but three replicates at Svalöv (SVA: 55.900°N, 13.100°E) (Supplementary Figure S1). DMY1, DMY2, CP1 and CP2 data were obtained. Of the 49 accessions, 42 were part of the EUCLEG panel, so only those were genotyped.

2.2 Genotypic data

GBS was used to obtain allele frequency data from a total of 400 red clover accessions as previously described (Frey et al., 2022; Nay et al., 2023). A total of 200 seedlings per accession were germinated in a greenhouse. The leaf at the one-leaf stage was harvested from each seedling, and the leaves from each accession were combined. DNA was extracted from each sample using the QIAGEN DNeasy 96 Plant kit (QIAGEN, Citylabs 2.0, Manchester M13 0BH, UK), and the concentration normalized to 20 ng μl^{-1}. Genotyping was carried out by LGC Genomics (Berlin, Germany) using a *Pst*I-*Mse*I double-digest pool-GBS method followed by PE-150 Illumina sequencing. Sequencing data covered 10,609 unique loci with an average read depth of 288. SNP calling and calculations of allele frequencies were done as described in Keep et al. (2020). The parameters used are described in detail in the supplementary methods of Frey et al. (2022). SNPs were retained if allele frequencies of at least 10 accessions were between 0.05 and 0.95, and if mean allele frequencies across all accessions were between 0.05 and 0.95. SNP positions with more than 5% missing values were discarded. The remaining missing values were imputed by the mean allele frequency across all accessions. After quality control and filtering, allele frequency data of a total of 20,156 SNP markers in 392 accessions were retained.

2.3 Experimental design and phenotypic data

The phenotypic data from all locations except Sweden were analysed using the methods described earlier (Nay et al., 2023). The data used in the present work were based on scaled and normalised values of the best linear unbiased estimates (BLUEs) from each

10.3389/fpls.2024.1407609

location separately. The data were analysed as seen in Equation 1 using the following model:

$$y_{imno} = \mu + g_i + b_m + ib1_n + ib2_o + \epsilon_{imno} \quad (1)$$

where y_{imno} is the phenotype on a single plot, μ is the overall mean, g_i is the effect of the ith accession, b_m is the effect of the mth block, $ib1_n$ is the effect of incomplete block 1 (i.e. row n), $ib2_o$ is the effect of incomplete block 2 (i.e. column o) and ϵ_{imno} is the residual error. For the p-rep designs (CZE, CHE and GBR) the block term was omitted, while for NOR the $ib2_o$ term was omitted. Data were available only from one complete block in SRB, so no separate analysis was performed with data from this location. A linear mixed model analysis was performed with $ib1_n$ and $ib2_o$ as random effects, while g_i (and b_m where relevant) were treated as fixed effects to obtain the BLUE values for each accession. Best linear unbiased predictions (BLUP) values were subsequently obtained by treating g_i as a random effect. These were used to estimate heritabilities using the method of Walsh and Lynch (2018) by regressing BLUP on BLUE for each trait. The relevant correlations and heritability values are reproduced in Supplementary Table S2.

The layout of the field trials in Sweden was a randomised complete block design (RCBD), where BJT and KLB had two replicates and SVA three replicates or blocks. BLUE values of accessions from each location in Sweden were obtained using the following model:

$$y_{ij} = \mu + g_i + b_j + \varepsilon_{ij} \quad (2)$$

where y_{ij} is the phenotypic value of the ith accession in the jth block, μ is the overall mean, g_i is the fixed effect of the ith accession, b_j is the random effect of the jth block, and ε_{ij} is the residual (Equation 2). The data were also analysed by having the accessions as a random effect to obtain best linear unbiased predictions (BLUP).The broad sense heritability was calculated as follows:

$$h_b^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\varepsilon^2}{n_r}}, \quad (3)$$

where σ_g^2 is the genetic variance, σ_ε^2 is the residual variance and n_r is the number of replicates (Equation 3). The raw phenotypic data and the heritability values are given in Supplementary Tables S3 and S4.

2.4 Genomic predictions

Three models for GPs were used, the first is a single environment (SingleEnv) strategy, in which separate analyses were carried out for each environment. A linear model was used as follows:

$$\boldsymbol{y_j} = \mathbf{1}\boldsymbol{\mu_j} + \boldsymbol{X_j}\,\boldsymbol{\beta_j} + \boldsymbol{\varepsilon_j} \quad (4)$$

where $\boldsymbol{y_j}$ is a vector of phenotypes from the jth environment, $\boldsymbol{\mu_j}$ is the overall mean, $\boldsymbol{X_j}$ is a matrix of marker allele frequencies, all centred (by subtracting the mean allele frequency) and standardised (by dividing by the standard deviation), and $\boldsymbol{\beta_j}$ is a vector of marker effects, $N(\mathbf{0}, \boldsymbol{I}\sigma^2_{\beta j})$. Finally, $\boldsymbol{\varepsilon_j}$ is the residual, $N(\mathbf{0}, \boldsymbol{I}\sigma^2_{\varepsilon j})$. The model (Equation 4) assumes equal variance across environments. For genomic best linear unbiased prediction (GBLUP) this can be rewritten so that $\boldsymbol{u_j} = \boldsymbol{X_j}\,\boldsymbol{\beta_j}$, so that $\boldsymbol{u_j} \sim N(\mathbf{0}, \boldsymbol{G_j}\sigma^2_{uj})$, with $\boldsymbol{G_j} = \frac{X_j X_j'}{p}$, where p is the number of markers, and σ^2_{uj} is the marker variance for the jth environment.

The second strategy assumes that the marker effects are the same across the environments being compared (AcrossEnv). This means that the linear model, where n environments are compared, becomes:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{1}\,\mu_1 \\ \mathbf{1}\,\mu_2 \\ \vdots \\ \mathbf{1}\,\mu_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{u_1} \\ \boldsymbol{u_2} \\ \vdots \\ \boldsymbol{u_n} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix} \quad (5)$$

with $\boldsymbol{u_j} = \boldsymbol{X_j}\boldsymbol{\beta}$ and $\boldsymbol{u} = (\boldsymbol{u_1'}, \boldsymbol{u_2'}, \ldots \boldsymbol{u_u'})' \sim N(\mathbf{0}, \boldsymbol{G_0}, \sigma^2_u)$, for which σ^2_u is the variance for the main genetic effects across environments (Equation 5). $\boldsymbol{G_0}$ is a marker-derived genomic relationship matrix (Equation 6).

$$\boldsymbol{G_0} = \begin{bmatrix} \boldsymbol{X_1X_1'} & \boldsymbol{X_1X_2'} & \cdots & \boldsymbol{X_1X_n'} \\ \boldsymbol{X_2X_1'} & \boldsymbol{X_2X_2'} & \ldots & \boldsymbol{X_2X_n'} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{X_nX_1'} & \boldsymbol{X_nX_2'} & \ldots & \boldsymbol{X_nX_n'} \end{bmatrix} / p \quad (6)$$

In the work described here, the marker matrices $\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_n}$ are identical for each multi-environment analysis, so $\boldsymbol{X_j}$ reduces to $\boldsymbol{X}$.

The third strategy incorporates marker x environment (MxE) interaction effects. This is done by splitting the effect of the kth marker on the jth environment β_{jk} into two parts, one which is the same in all environments b_{0k} ($\boldsymbol{b_0} \sim N(\mathbf{0}, \boldsymbol{I}\,\sigma^2_{b_0})$) and one which is specific for each environment b_{jk} ($\boldsymbol{b_j} \sim N(\mathbf{0}, \boldsymbol{I}\,\sigma^2_{b_j})$). For GBLUP the model then can be written as shown in Equation 7:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{u_0} + \boldsymbol{u_1} + \boldsymbol{\varepsilon_1} \quad (7)$$

where $\boldsymbol{y} = (\boldsymbol{y_1} + \boldsymbol{y_2} + \ldots \boldsymbol{y_n})'$ is the response vector for which $\boldsymbol{y_j}$ is the vector of observations of the accessions in the jth environment, and.

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \\ \vdots \\ \mathbf{1}\mu_n \end{bmatrix}, \boldsymbol{u_0} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_n} \end{bmatrix} \boldsymbol{b_0}, \boldsymbol{u_1} = \begin{bmatrix} \boldsymbol{X_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{X_n} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \\ \vdots \\ \boldsymbol{b_n} \end{bmatrix}$$

in which μ is the overall mean, $\boldsymbol{u_0} \sim N(\mathbf{0}, \boldsymbol{G_0}\,\sigma^2_{uo})$, for which σ^2_{uo} represents the variance of the common effects across environments as described above for the AcrossEnv model, $\boldsymbol{u_1} \sim N(\mathbf{0}, \boldsymbol{G_I})$, $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \quad \boldsymbol{I}\sigma^2_\varepsilon)$. $\boldsymbol{G_I}$ is a relationship matrix, which, as seen in Equation 8, captures the environment specific effects:

$$\boldsymbol{G_1} = \begin{bmatrix} \sigma^2_{u1}\boldsymbol{X_1X_1'} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \sigma^2_{u2}\boldsymbol{X_2X_2'} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \sigma^2_{un}\boldsymbol{X_nX_n'} \end{bmatrix} / p \quad (8)$$

The Bayesian generalised linear regression (BGLR) package (Pérez and De Los Campos, 2014) and the code described by Lopez-Cruz et al. (2015) were used for calculations. The number of iterations was set to 25,000 with 2,500 for burn-in.

Two cross validation methods, CV1 and CV2 as described by Lopez-Cruz et al. (2015), were applied. Briefly, the PAs were based on the mean of 50 random partitions between the training set and test set with a default split of 70% to 30%, respectively. CV1: test accessions have not been assessed in any of the environments. Thus for multi-environmental analyses the test set consisted of the same accessions from all environments in the analysis. CV2: test accessions have been evaluated in some, but not all environments, and the test sets do not overlap. The methods are illustrated for a two-environment scenario in Supplementary Figure S2.

The SingleEnv analysis is straightforward as it produces PA values for each environment separately. For the AcrossEnv and MxE methods, multi-location analysis of trial sites was not possible for GBR, NOR and SRB, as they shared only few accessions. Only comparisons between CHE and CZE were meaningful as they shared all 392 accessions. Since the field experiment in CZE was abandoned after the first year, only data from year 1 could be included from this location. Most of the multi-environment analyses were thus pair-wise comparisons between DMY data from year 1 and 2 at the same location, or between CP data from cut1 and cut 2 in year 1. The multi-location analyses included a pair-wise comparison of DOF in CHE and CZE, a three-way comparison of CHE_DMY1, CHE_DMY2 and CZE_DMY1, a pairwise analysis of CHE_DMY1 and CZE_DMY1, and finally a four-way analysis of the CP content at cut1 and 2 at both locations. The AcrossEnv and the MxE strategies produces PA values for each environment, based on the joint analysis of the environments. It should also be noted that for the SingleEnv analyses the two CV methods are equivalent, as they both use 50 randomly selected training/test set combinations.

A third CV method was used which either left one of the five regions out from the training set, and used it as test set, or left one of the six breeders' material out and used that as test set (Supplementary Table S1). The number of accessions per region varied between 16 and 142, and per breeders' material it ranged from 18 to 61. For comparison the CV1 method was used to remove 142 accessions randomly (comparing with regions), or 61 accessions (comparison with breeders' material) from the training set. For DMY the MxE model was used for the data from CHE year 1 and 2 to obtain the GEBVs. For DOF the MxE model was used in the joint analysis of data from CHE and CZE, year 1.

The PA data from the output were analysed using analysis of variance (ANOVA) to identify significant differences between environments, models and cross validation methods. One-way ANOVA was used to assess the effect of each factor, and three-way ANOVA to assess the combined effect of all three factors. The R statistical software was used for this (R Core Team, 2023).

3 Results

Initially, a variance component analysis was performed for each phenotype to get information on how the SingleEnv, the AcrossEnv and MxE models fitted the data. Secondly, GP analyses were performed to compare the performance of the three models. The multi-environment analyses included joint analyses of pairs of environments at the same location (year 1 and 2) for DMY, and joint analyses of the two locations CHE (year 1 and 2) and CZE in year 1. For CP content joint analyses of pairs of environments at the same location (cut 1 and cut 2 in year 1) were performed, and joint analyses of the two locations CHE and CZE at cut 1 and cut 2.

3.1 Variance component analysis for DMY

The variance component estimations for the annual DMY data derived from SingleEnv analyses varied considerably among the environments (Table 1). The residual variance was much higher for CZE than CHE, hence the proportion of the variance explained by the markers was considerably higher for CHE. In the multi-environment analyses, the R^2 values were generally higher for the MxE model compared to the AcrossEnv model. In the joint analysis of CZE.1, CHE.1 and CHE. 2, the MxE interaction variance was very high for CZE.1. Consequently, the R^2 values were very high for the MxE model. The specific-environment variance for CZE.1 contributed 84.3% of the total variance (Supplementary Figure S3). For the other environments the proportions were between 15.2% and 41.9%. In all cases the R^2 values were higher for the MxE model than for the AcrossEnv model (Table 1). The DMY data from GBR, NOR and SRB were based on approximately 100 accessions. Due to the scant overlap between accessions at those three locations, multi-environment analyses were thus carried out only between the two years at each site, not between the locations. At all three of those locations the genetic effects in the MxE model accounted for a higher proportion of total variance than the AcrossEnv model (Table 1).

3.2 Variance component analysis for CP content

The residual variance proportions were generally high for CP content, which resulted in lower R^2 values than for DMY in all three models (Table 2). Again, the variance explained by the MxE model was higher than for the AcrossEnv model. These results are generally consistent with the low heritability recorded for CP content (Supplementary Table S2; Nay et al., 2023). The exception was NOR, where the R^2 values and heritability were higher than for the other locations, at least in the joint analysis (Table 2). Supplementary Figure S4 underlines this, showing that except for NOR, the residual variance contributes over 70% of the total variance, and the main effect variance was 10.1% or lower. The MxE model thus explains little of the phenotypic variation for this trait.

3.3 Variance component analysis for DOF

DOF was recorded during year 1, which meant that only one joint analysis was possible, namely between the locations CHE and CZE. The variance explained by the genetic markers was higher than for the CP content, but lower than for DMY (Table 3). In the SingleEnv analysis, the proportion of the variance explained by markers was

TABLE 1 Variance component estimation (SD) of annual DMY from field trials in Czechia (CZE), Switzerland (CHE), Britain (GBR), Norway (NOR) and Serbia (SRB) using the SingleEnv, AcrossEnv and the MxE interaction models.

	SingleEnv model			AcrossEnv model			MxE model			
Environment	Residual variance (σ^2_E)	Marker variance (σ^2_U)	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	M x E interaction (σ^2_{U1})	R^2
CZE_DMY1	0.737(0.068)	0.301(0.072)	0.289(0.057)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CHE_DMY1	0.111(0.014)	0.288(0.034)	0.722(0.040)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CHE_DMY2	0.170(0.026)	0.512(0.069)	0.748(0047)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
GBR_DMY1	0.172(0.040)	0.490(0.093)	0.737(0.060)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
GBR_DMY2	0.187(0.044)	0.492(0.098)	0.721(0.065)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
NOR_DMY1	0.386(0.102)	0.592(0.161)	0.598(0.108)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
NOR_DMY2	0.375(0.133)	0.604(0.207)	0.610(0.130)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
SRB_DMY1	0.526(0.115)	0.491(0.149)	0.481(0.105)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
SRB_DMY2	0.430(0.114)	0.607(0.179)	0.578(0.111)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CZE_DMY1 & CHE_DMY1	*n.a*	*n.a*	*n.a*	0.533(0.033)	0.212(0.032)	0.283(0.035)	0.178(0.028)	0.114(0.041)	1.611(0.267) 0.137(0.042)	0.903(0.032) 0.585(0.053)
CHE_DMY1 & CHE_DMY2	*n.a*	*n.a*	*n.a*	0.259(0.016)	0.271(0.034)	0.510(0.036)	0.101(0.011)	0.262(0.034)	0.066(0.017) 0.290(0.048)	0.763(0.030) 0.844(0.023)
CZE_DMY1 & CHE DMY1 & CHE DMY2	*n.a*	*n.a*	*n.a*	0.511(0.023)	0.198(0.027)	0.278(0.030)	0.138(0.016)	0.218(0.032	1.910(0.236) 0.065(0.018) 0.257(0.050)	0.938(0.016) 0.671(0.039) 0.773(0.036)
GBR_DMY1 & GBR_DMY2	*n.a*	*n.a*	*n.a*	0.149(0.020)	0.444(0.077)	0.745(0.043)	0.100(0.017)	0.411(0.078)	0.094(0.030) 0.095(0.032)	0.833(0.035) 0.833(0.035)
NOR_DMY1 & NOR_DMY2	*n.a*	*n.a*	*n.a*	0.544(0.087)	0.399(0.122)	0.417(0.095)	0.357(0.086)	0.317(0.152)	0.245(0.122) 0.269(0.130)	0.605(0.104) 0.609(0.111)
SRB_DMY1 & SRB_DMY2	*n.a*	*n.a*	*n.a*	0.512(0.072)	0.465(0.148)	0.470(0.086)	0.361(0.069)	0.486(0.148)	0.152(0.070) 0.179(0.080)	0.633(0.087) 0.642(0.084)

R^2 of models was calculated as the ratio of the sum of the main and interaction variance relative to the sum of the residual, the main and the interaction variance. n.a., not applicable.

10.3389/fpls.2024.1407609

TABLE 2 Variance component estimation (SD) of CP content of cut 1 (CP1) and cut 2 (CP2) in year 1 from field trials in Czechia (CZE), Switzerland (CHE), Britain (GBR), Norway (NOR) and Serbia (SRB) using the SingleEnv, AcrossEnv and the MxE interaction models.

	SingleEnv			AcrossEnv model			MxE model			
Environment	Residual variance (σ^2_E)	Marker variance (σ^2_U)	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	M x E interaction (σ^2_{U1})	R^2
CZE_CP1	0.863(0.076)	0.247(0.067)	0.221(0.053)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CZE_CP2	0.854(0.071)	0.206(0.053)	0.194(0.044)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a*
CHE_CP1	0.703(0.072)	0.349(0.087)	0.330(0.067)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CHE_CP2	0.828(0.071)	0.249(0.063)	0.230(0.050)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
GBR_CP1	0.522(0.127)	0.579(0.186)	0.737(0.060)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
GBR_CP2	0.786(0.148)	0.397(0.147)	0.333(0.099)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
NOR_CP1	0.596(0.129)	0.472(0.158)	0.439(0.114)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
NOR_CP2	0.226(0.049)	0.425(0.087)	0.650(0.074)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
SRB_CP1	0.692(0.149)	0.465(0.170)	0.398(0.115)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
SRB_CP2	0.818(0.151)	0.379(0.139)	0.315(0.095)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CZE_CP1 & CZE_CP2	*n.a*	*n.a*	*n.a*	0.774(0.059)	0.295(0.098)	0.274(0.075)	0.766(0.065)	0.243(0.105)	0.094(0.035) 0.082(0.028)	0.301(0.079) 0.294(0.079)
CHE_CP1 & CHE_CP2	*n.a*	*n.a*	*n.a*	0.864(0.050)	0.136(0.038)	0.135(0.034)	0.796(0.050)	0.088(0.036)	0.151(0.057) 0.115(0.042)	0.230(0.051) 0.203(0.045)
CZE_CP1 & CHE CP1 & CZE CP2 & CHE_CP2	*n.a*	*n.a*	*n.a*	0.935(0.037)	0.065(0.021)	0.065(0.020)	0.859(0.037)	0.040(0.021)	0.098(0.043) 0.145(0.054) 0.077(0.030) 0.113(0.042)	0.137(0.042) 0.176(0.045) 0.119(0.033) 0.150(0.042)
GBR_CP1 GBR_CP2	*n.a*	*n.a*	*n.a*	0.932(0.101)	0.115(0.044)	0.129(0.047)	0.769(0.100)	0.072(0.035)	0.234(0.112) 0.182(0.096)	0.281(0.084) 0.245(0.079)
NOR_CP1 NOR_CP2	*n.a*	*n.a*	*n.a*	0.577(0.065)	0.250(0.066)	0.302(0.063)	0.326(0.072)	0.174(0.071)	0.479(0.208) 0.190(0.072)	0.651(0.113) 0.527(0.083)
SRB_CP1 SRB_CP2	*n.a*	*n.a*	*n.a*	0.882(0.102)	0.162(0.075)	0.154(0.064)	0.809(0.108)	0.108(0.065)	0.176(0.093) 0.156(0.079)	0.259(0.084) 0.245(0.082)

R^2 of models was calculated as the ratio of the sum of the main and the interaction variance relative to the sum of the residual, the main and the interaction variance. n.a., not applicable.

10.3389/fpls.2024.1407609

10.3389/fpls.2024.1407609

TABLE 3 Variance component estimation (SD) of flowering time (DOF) from field trials in Czechia (CZE), Switzerland (CHE), Britain (GBR), Norway (NOR) and Serbia (SRB) using the SingleEnv, AcrossEnv and the MxE interaction models.

	SingleEnv model			AcrossEnv model			MxE model			
Environment	Residual variance	Marker variance	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	R^2	Residual variance (σ^2_E)	Marker variance (σ^2_{U0})	M x E interaction (σ^2_{U1})	R^2
CZE_DOF	0.346(0.041)	0.385(0.069)	0.525(0.063)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CHE_DOF	0.371(0.039)	0.360(0.061)	0.229(0.051)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a*
GBR_DOF	0.231(0.055)	0.590(0.120)	0.716(0.067)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
NOR_DOF	0.151(0.036)	0.356(0.069)	0.699(0.070)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
SRB_DOF	0.222(0.053)	0.514(0.105)	0.695(0.071)	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
CZE_DOF & CHE_DOF	*n.a*	*n.a*	*n.a*	0.441(0.026)	0.186(0.029)	0.296(0.036)	0.378(0.026)	0.243(0.105)	0.088(0.030) 0.092(0.036)	0.390(0.045) 0.394(0.045)

R^2 of models was calculated as the ratio of the sum of the main and the interaction variance relative to the sum of the residual, the main and the interaction variance. n.a., not applicable.

higher for CZE than for CHE. For the MxE model the proportions of main effect, specific environmental effect and residual variance were similar for the two locations (Supplementary Figure S3). In the SingleEnv analysis for GBR, NOR and SRB the proportion of the variance explained by the genetic markers was similar for the three locations, and somewhat higher than the corresponding ones from the SingleEnv analysis of the CZE and CHE data (Table 3).

3.4 Predictive ability for DMY

PA of the joint analysis of the annual DMY for two years at each location is shown in Figure 1. Analysis of variance of the DMY data showed that location_year, model and CV method, and their interactions had a significant effect on PA (Supplementary Table S5). The CV2 cross validation method resulted in higher PAs than CV1, and PAs from CHE and GBR were consistently high, while NOR and SRB gave lower and more variable PAs. The PA values varied between 0.232 in NOR_DMY2 AcrossEnv model and CV2 to 0.923 for GBR.1, MxE model and CV2 (Supplementary Table S5). Overall, the MxE model gave the highest PAs and AcrossEnv the lowest. The CV2 cross validation method generally resulted in higher PAs than CV1 (0.653 vs 0.638).

Both the three-way joint analysis (Figure 2A) and the pairwise joint analysis of CHE_DMY1 and CZE_DMY1 (Figure 2B) show that the PAs for CZE_DMY1 are much lower than those for CHE, irrespective of year, model and cross validation method (Supplementary Tables S6, S7). The MxE model resulted in the highest PA values, and the AcrossEnv model the lowest PA for CHE, while the SingleEnv gave the highest and the MxE interaction model, the lowest PA for CZE (Supplementary Table S6). The pairwise comparison between CHE_DMY1 and CZE_DMY1 in year 1 was similar, with the PAs for CHE varying between 0.774 and 0.879, while those for CZE_DMY1 were no higher than 0.331 (Figure 2B; Supplementary Table S7). Overall, the SingleEnv model resulted in the highest PA values and the AcrossEnv model gave the lowest PAs. There was no significant effect of the CV method used.

3.5 Predictive ability for CP content

The PAs for the CP content were generally much lower than for DMY, except for NOR_CP2 (Figure 3; Supplementary Table S2), for which a higher proportion of the variance was explained by the marker main effect (Table 3; Supplementary Figure S4). PA from NOR_CP2 varied between 0.733 and 0.836, but were low, even zero or negative for the other Location_cut combinations. Overall, the MxE model performed best, while the SingleEnv model gave the lowest PAs (Supplementary Table S8). The CV2 method was slightly superior to CV1. The multi-environment analysis of all four environments in CHE and CZE showed that CHE_CP1 had the highest PA (Supplementary Figure S5), while CZE_CP1 and CZE_CP2 both resulted in PA values around zero. The MxE model performed best, and the AcrossEnv model was poorest. Overall, the

10.3389/fpls.2024.1407609

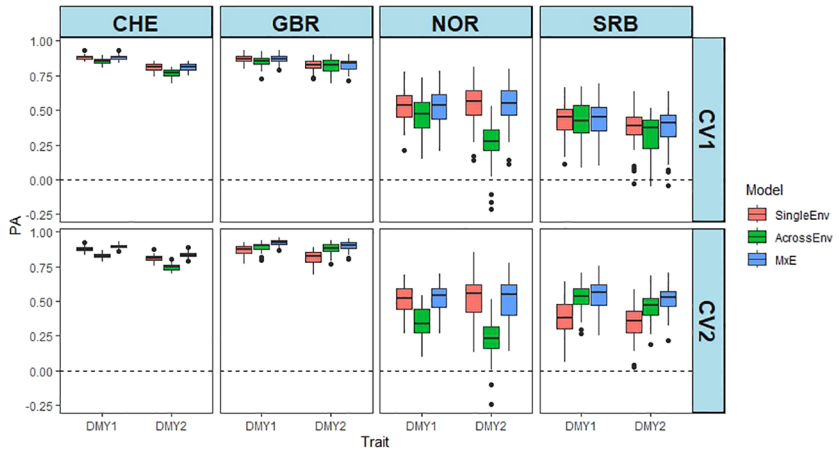


FIGURE 1
Predictive ability (PA) of DMY in year 1 and 2 (e.g. CHE_DMY1, CHE_DMY2 etc.) at four locations. PA is the Pearson correlation between predicted and the scaled and normalised BLUE values of DMY using SingleEnv, AcrossEnv and MxE models and CV1 and CV2 cross validation methods. The environmental comparisons are between year 1 and year 2 at each location separately. The boxplots represent 50 training-test partitions.

CV2 method resulted in higher PA values than CV1 (Supplementary Table S9). The PAs from CHE_CP1 had low to moderate PAs, while the PAs for CZE were around zero for both CV methods (Supplementary Table S9). The SingleEnv and the MxE interaction models were better than AcrossEnv. The mean PA of the CV2 method was slightly higher than CV1.

3.6 Predictive ability for DOF

DOF was recorded in year 1 only, so the data from GBR, NOR and SRB were analysed by the SingleEnv model. Figure 4A shows that the PA values were high for all three locations, varying between 0.762 in NOR and 0.838 in GBR (Supplementary Table S10). For CHE and CZE, a pairwise comparison between the two locations was carried out, using all three models and both cross validation methods. In both locations the PA was high (0.671 for CHE, and 0.690 for CZE) (Figure 4B; Supplementary Table S11). The MxE model again performed better than SingleEnv and AcrossEnv. There were no significant differences between CV methods. The high PA values are in line with the high heritability values recorded for DOF (Supplementary Table 2) (Nay et al., 2023).

3.7 Prediction bias

Possible biases of predictions can be revealed by regressing the observed phenotypes on the GEBV. The results of this analysis are

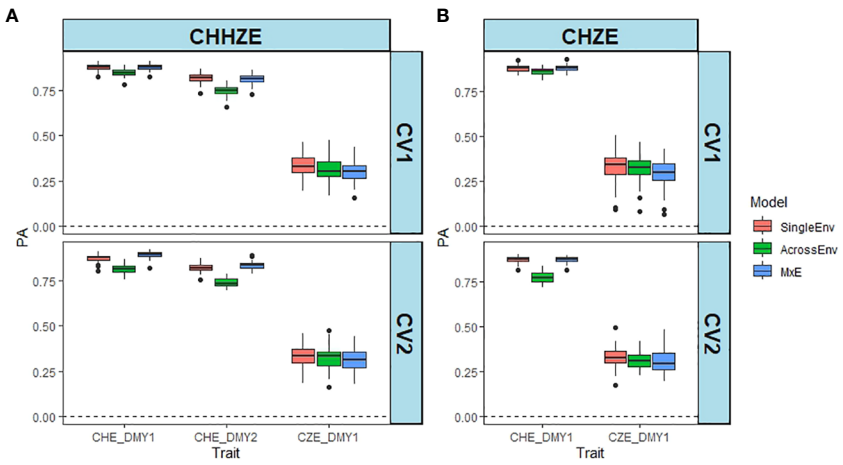


FIGURE 2
Comparison between predictive ability (PA) of DMY estimated in CHE and CZE. PA is the Pearson correlation between predicted and the scaled and normalised BLUE of DMY using SingleEnv, AcrossEnv and MxE models and CV1 and CV2 cross validation methods. **(A)** Comparison across three environments (CHHZE): CHE_DMY1, CHE_DMY2 and CZE_DMY1. **(B)** Pairwise comparison (CHZE) between CHE_DMY1 and CZE_DMY1. The boxplots represent 50 training-test partitions.

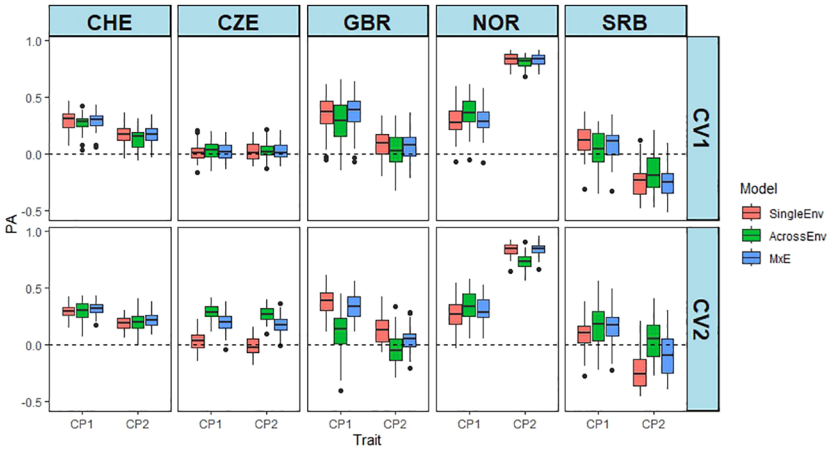


FIGURE 3
Predictive abilities (PA) of CP content from the joint analyses of cut 1 and cut 2 of year 1 at each location separately. PA is the Pearson correlation between predicted and the scaled and normalised BLUE values of CP content using SingleEnv, AcrossEnv and MxE models and CV1 and CV2 cross validation methods.

shown in Figure 5. We used the MxE interaction model and the CV2 cross validation method on the data from CHE and CZE. A slope of 1 indicates no bias. The larger the deviation from 1, the larger the bias. For CP content the bias was largest (3.13), while for DMY (1.07) and DOF (1.15), the slope values were near 1, indicating low bias (Supplementary Table S12). This is consistent with PA values, which were generally high for DMY and flowering, and low for CP overall (Figures 1–4).

3.8 Effect of marker numbers and training set size on predictive ability

The effect of marker numbers on the PA values was investigated by sampling subsets of the full marker set of 20,156 SNPs (20K), down to 100 SNP markers (0.1K). The PA values are based on DMY data from year 1 and 2 in Switzerland (CHE_DMY1 and CHE_DMY2). The CV2 method of cross validation was used, as it was performing best throughout this work. Remarkably, the effect of lowering the marker numbers was not dramatic until less than 1000 markers were used (Figure 6A; Supplementary Table S13). Overall, PA was reduced from 0.83 at 10K, 15K and 20K markers to 0.74 at 100 markers (0.1K), still high for such a small number of markers. At all marker numbers, the MxE model performed best (0.85), and AcrossEnv performed worst (0.78) (Supplementary Table S13).

The effect of the percentages of all accessions used as the test set was also analysed. The default split used throughout this work was 70% as training set and 30% as test set. Figure 6B summarises the results obtained by increasing the test set percentage from 30% to

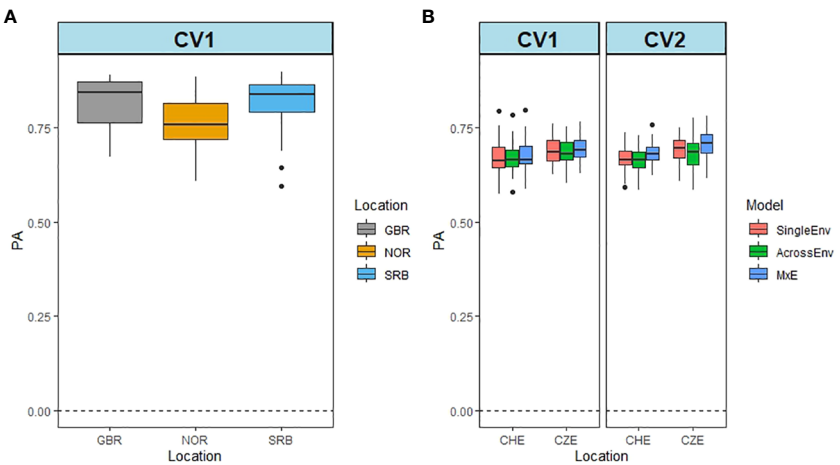


FIGURE 4
Predictive ability of flowering time (DOF) from Britain (GBR), Norway (NOR) and Serbia (SRB) using the SingleEnv analysis and the CV1 cross validation method **(A)**. Prediction ability from CHE and CZE using SingleEnv, AcrossEnv and MxE models and CV1 and CV2 cross validation methods **(B)**.

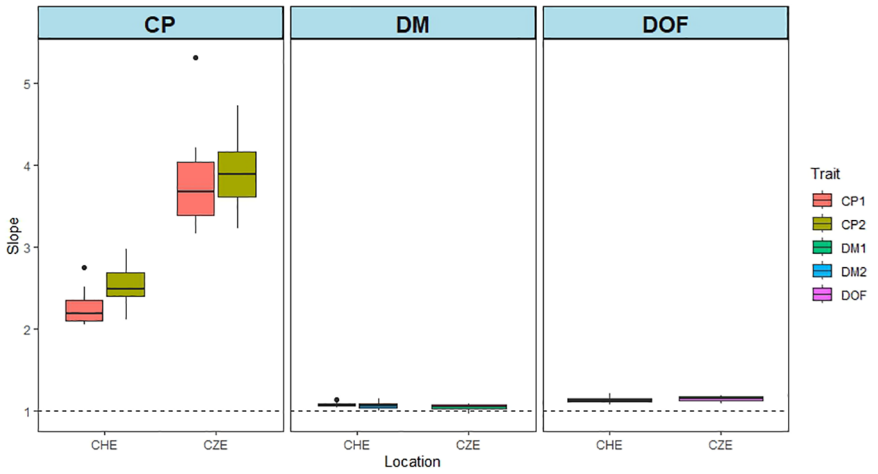


FIGURE 5
Bias of predictions for data from CHE and CZE. The graphs show the slopes of the regression of GEBVs on phenotypic values. The stippled line indicates a slope value of 1. The GEBVs are based on the joint analysis of all the environments for each of the three traits: CP content, DMY and DOF. The MxE interaction model and the CV2 cross validation method were used. Fifty iterations of training-test set combinations were performed. CP1 and CP2 represents CP content of cut 1 and cut 2 in year 1, respectively. DMY1 and DMY2 represent annual DMY in year 1 and year 2, respectively.

90%, and therefore decreasing the training set percentage from 70% to 10%. There was a significant effect of test set percentage with the smallest (30%) having the highest PA of 0.83 and the largest (90%) having the lowest PA of 0.75 (Supplementary Table S14). Again, the MxE interaction model gave the highest PA values, and AcrossEnv the lowest. Visual inspection of the data (Figure 6B) suggests that the decrease in PA takes effect when the test set percentage is larger than 50%.

The effect of removing accessions from specific regions or specific breeders' material was more pronounced for the latter (Figure 7). There was also a larger reduction in PAs for DOF than for DMY. Concerning the regions removing Northern European accessions resulted in the largest reduction in PA compared to the random control for both traits, but removing Swiss accessions also had a sizeable effect (Figures 7A, C). For DMY, removing the Lantmännen breeders' material had the smallest effect on PA, at least in year 1, while it reduced PA to zero or negative values for DOF (Figures 7B, D).

3.9 Test of predictive ability on new phenotypic data

So far, the prediction abilities presented here are based on cross validations derived from training and test sets where both phenotypic and genotypic data from the same accessions are used

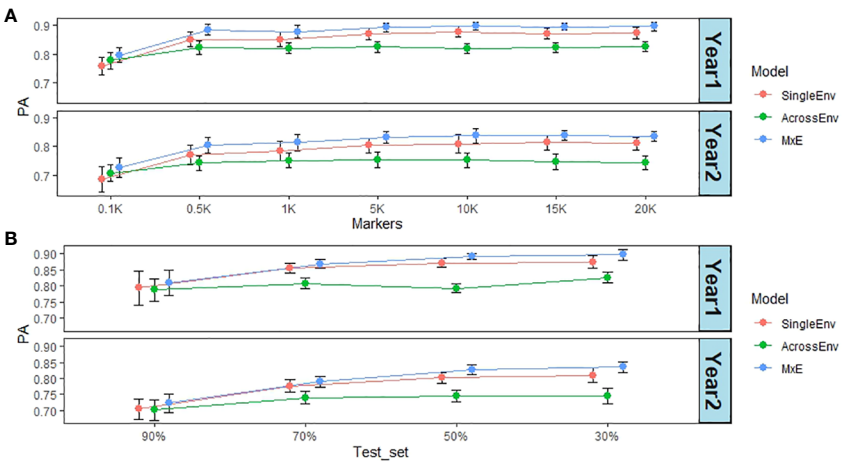


FIGURE 6
(A) Effect of marker numbers (in thousands) on prediction ability (PA). The default training set/test set ratio of 70%/30% was used; **(B)** Effect of test set percentage on PA. The full set of 20,156 markers were used. For both analyses DMY from year 1 and 2 in CHE were used as phenotypes, together with the CV2 cross validation method.

10.3389/fpls.2024.1407609

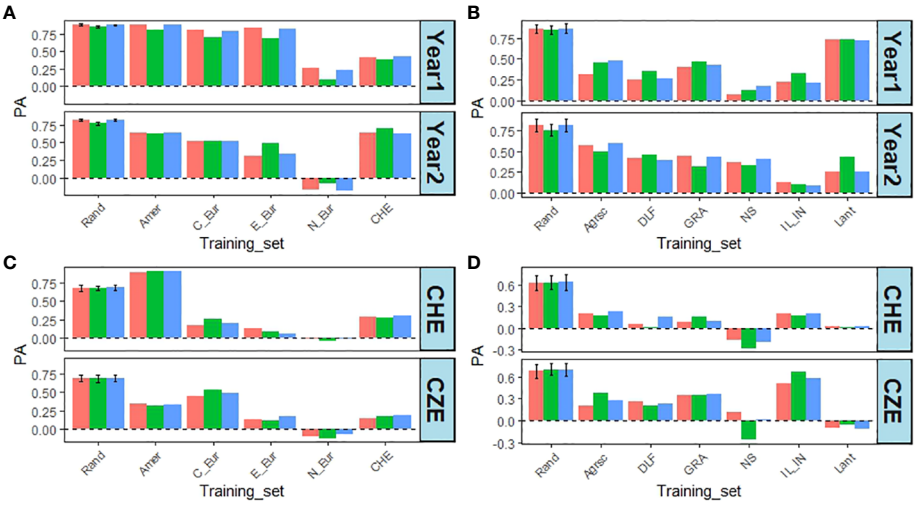


FIGURE 7
Effect of leaving one region or one breeders' material out from the training set for DMY, from CHE in year 1 and 2, and DOF from CHE and CZE in year 1. **(A)** PA for DMY when leaving one region out. **(B)** PA for DMY when leaving out one breeders' material. **(C)** PA for DOF from CHE or CZE, when leaving one region out. **(D)** PA for DOF from CHE or CZE when leaving out one breeders' material. Rand – Mean and standard deviation of 10 random removals of 142 accessions (regions) or 61 accessions (breeders' material) using the CV1 method; Amer – Americas; C_eur – Central Europe; E_eur – Eastern Europe; N_eur – Northern Europe; CHE – Switzerland; Agrsc – Agroscope; DLF – DLF; GRA – Graminor; NS – IFVCNS; IL_IN – ILVO_INRAe; Lant – Lantmännen. Bar colours: red = SingleEnv; green = AcrossEnv; blue = MxE.

by the models. To test the predictions on new phenotypic data not included in any of the analyses, we used DMY and CP content from field trials carried out at three locations in Sweden (BJT, KLB, SVA; Supplementary Figure S1). A total of 42 accessions from the EUCLEG panel were used. The DMY and CP content data are shown in Supplementary Table S3 and in Supplementary Figures S6 and S7. The broad sense heritability for each of the traits revealed that DMY traits had high heritability, while CP content had low to medium heritability (Supplementary Table S4). This is broadly in agreement with the results from the other five locations. The PAs for the four traits at the three locations are shown in Figure 8. There are four sets of PAs per location in Sweden, and they represent the origin of the training set. Each PA value is the correlation between the phenotypic value of the trait in one of the Swedish locations and the GEBV of the corresponding trait from one of the four other countries. Using NOR data as training sets gave rise to uniformly low PAs, which were either negative or close to zero. Data with GBR training sets gave low or negative values for CP content, but



FIGURE 8
Effect of training sets from CHE (392), GBR (99), NOR (107) and SRB (100) on prediction ability (PA) at three locations in Sweden (BJT = Bjertorp; KLB = Kölbäck; SVA = Svalöv) for DMY year 1 and 2, (DM1 and DM2) and CP in cut 1 and cut 2 in year 1 (CP.1 and CP.2).

moderate to high PA values for DMY, particularly in year 1. The SRB training sets gave rise to low or poor values throughout with only CP1 at SVA being high. The CHE training set was the largest (392 accessions) and gave low to moderate PA values for CP content, except for SVA, and high prediction ability for DMY in year 1, but not year 2. These results confirm our earlier results with DMY giving higher PA than CP content, and they are overall in line with the contrasting heritability for the two traits.

4 Discussion

In our work on red clover presented here the MxE model outperformed the SingleEnv and AcrossEnv models. The MxE model was initially developed and tested using wheat yield data by Lopez-Cruz et al. (2015). They showed that phenotypic correlation between pairs of environments was directly associated with the proportion of genomic variance explained by the main effect of the markers. They also demonstrated that the MxE model fitted the data better than the AcrossEnv model. In the vast majority of cases prediction accuracy was highest when the MxE model was used.

Similar results were obtained in an analysis of agronomic traits in sesame (Sabag et al., 2023). However, only two environments were compared (two consecutive years), and the phenotypic correlation for all nine traits tested were all rather high between years (0.50 to 0.96, depending on the trait). While the trait with the lowest correlation also had the lowest PA, it also had the lowest heritability, so it is difficult to disentangle the effects of the two factors. Nevertheless, it seems that the results described here for red clover follow the same trends.

4.1 Variance components

The proportion of total variance explained by the main marker effects was a good indicator of correlation between environments. This was the case for DMY and DOF, where the pairwise correlation between phenotypic values associated with the main effect variance proportion (Supplementary Figure S3; Supplementary Table S2). This is similar to what was found by others (Lopez-Cruz et al., 2015; Crossa et al., 2016; Bandeira E Sousa et al., 2017; Sabag et al., 2023). For DMY in year 1 in CZE however, the variance due to the main marker effect was small, while the environment-specific effect from the MxE model was very large (Table 1; Supplementary Figure S3). The model thus explains a large proportion of the phenotypic variance, but most of it is environment specific. It may be in this environment that many markers with environment-specific effects will be found. Crossa et al. (2016) also observed that in complex traits with lower heritability, such as grain yield and grain density, a larger fraction of the total genetic variance was due to environment-specific effects. This contrasted with the heading date and thousand grain weight, traits for which heritability was high and the environment-specific proportion of the genetic variance was low. This is consistent with the low heritability observed for DMY in year 1 in CZE (Supplementary Table S2).

The variance component analyses showed that the MxE model for the red clover data gave a better fit to the data than the AcrossEnv model (Tables 1–3), similar to what was found in wheat (Lopez-Cruz et al., 2015; Crossa et al., 2016) and maize (Bandeira E Sousa et al., 2017). In sesame, the AcrossEnv model was not included in the analyses, but the MxE model was found to enhance the PAs relative to the SingleEnv model (Sabag et al., 2023).

4.2 Predictive ability

Overall, the PAs were high for DMY and DOF, and low for CP content. This follows the heritability for the traits (Figures 1–4; Supplementary Table S2). The high phenotypic correlation between year 1 and year 2 of DMY in GBR (0.87) and CHE (0.71), respectively, may explain why the PA for those traits were uniformly high in both years (0.83 – 0.91) for the MxE model and CV2 cross validation method.

For CP content an exception was observed for the PA for NOR in cut 2, which was 0.83 for the SingleEnv and MxE models (Supplementary Table S8), and the heritability for this specific CP trait was also high (Supplementary Table S2). It is unclear whether this aberrant result is connected to the notion that most of Northern European accessions tend to have been adapted to one cut per season. The phenotypic correlation between NOR_CP1 and NOR_CP2 was 0.35, and this is probably why the PAs for cut 1 (0.26 – 0.37) were significantly lower than for cut 2 (Supplementary Table S2).

The MxE model performed better than the AcrossEnv model when pairs of environments were analysed jointly. This was the case for both DMY and CP content (Supplementary Tables S5, S8). The superiority of the MxE model was more pronounced when the CV2 cross validation method was used, and for pairs of environments with high phenotypic correlation (Figures 1–3; Supplementary Table S2). This is because the genetic covariance between environments is forced to be positive, as it is a product of the variance of the main effects (Lopez-Cruz et al., 2015). This also makes sense, because the CV2 method allows for borrowing information about accessions tested in one environment, but not in another. If the correlation is high, the information borrowed will be more accurate than if the correlation is low (Lopez-Cruz et al., 2015; Crossa et al., 2016; Sabag et al., 2023). In contrast, when CV1 is used for cross validation, the same accessions are missing from environments being analysed jointly. The SingleEnv analysis would then be expected to perform similarly to the AcrossEnv and MxE models. This was the case in red clover, at least for DMY (Figure 1), when correlations between pairs of traits were moderate to high. It should also be noted that for the SingleEnv analyses, the distinction between CV1 and CV2 is meaningless.

Overall, PAs for DMY were high (Figures 1, 4; Supplementary Tables S5-S7). Particularly, CHE and GBR had high PA for DMY (0.80–0,88) and for DOF (0.67 – 0.84) (Figure 4; Supplementary Tables S10-S11). To the best of our knowledge, no GP work has been published yet for red clover to compare directly with our study. A PA of 0.30 for biomass yield was reported for white clover (Griffiths et al., 2022). Despite such a comparatively low accuracy, selection based on the GEBVs resulted in higher gain than phenotypic selection. In alfalfa, Annicchiarico et al. (2015) reported PA values of 0.32 – 0.35 for

biomass yield in two populations, while Jia et al. (2018) found PAs of 0.13 for biomass yield in a diverse alfalfa germplasm collection. In breeding populations, PAs were found to vary between 0.21 – 0.66 depending on breeding cycle and location (Li et al., 2015). Recently, Pégard et al. (2023) reported a mean PA of 0.65 for DMY from two locations, France and Serbia, and 0.41 for DOF in a diverse panel of alfalfa accessions, where the marker data also were obtained as allele frequencies derived from sequencing of pooled plant material representing each accession. The PA values for DMY reported in the present work thus compare well with those reported for alfalfa and white clover.

Prediction abilities for DMY in the perennial ryegrass forage crop were 0.013 – 0.275 (Grinberg et al., 2016), 0.07 – 0.43 (Faville et al., 2018), and 0.28 – 0.59 (Pembleton et al., 2018). Despite low PAs on DMY, Faville et al. (2020) found that genomic prediction could still be used to improve the trait. It should be noted that these values were obtained using breeding populations, i.e. likely narrower germplasm, than in the present work. The germplasm used here was a diverse panel with significant population structure (Nay et al., 2023), which can inflate PA (Guo et al., 2014; Werner et al., 2020). Figure 7 demonstrates how this affected PAs of DMY and DOF. By having all accessions representing e.g. a single breeder in the test set and not in the training set, the PA values were mostly affected negatively. This is likely because no accessions from such a relatively narrow set were represented in the training population, whereas for the random CV, those accessions were present in both sets (Werner et al., 2020). There were variations in the magnitude of the effect depending on the trait, year (DMY) or location (DOF). Some of this may be explained by differences in the genetic architecture of the traits. The Northern European accessions tend to have been adapted to one cut per season, and late flowering, and may thus be genetically distinct from the other accessions, which are early flowering and adapted to a multi-cut management regime, hence the low PA for DMY (Figure 7A, C). There was a more modest decline for the two Northern European breeding materials, Graminor, and Lantmännen. This could possibly be explained by the presence of one of the two in either the training set and the test set (Figures 7B, D).

The PA values were remarkably resilient to lowering the number of SNP markers used to obtain the genomic relationship matrix (GRM) (Figure 6; Supplementary Table S13). While the PA values for marker numbers at 1K and below were significantly reduced, the PA value for 100 markers were still 0.74. The most likely explanation is that the genomic relationship does not need many markers to be reasonably stable. The GRM was obtained according to VanRaden (2008) for the GBLUP model. It has been reported that a GRM matrix, derived from the Euclidian distance between individuals based on the markers, can better capture non-additive marker interactions (Cuevas et al., 2016; Bandeira E Sousa et al., 2017). This remains to be tested in red clover. There are also other ways of modelling the GxE effect to better predict the performance of accessions across environments. Incorporating a factor analytic (FA) structure accommodates different trial designs and unbalanced data, allows for heterogeneity of variance across environments, and can help to explain GxE interaction in terms of a few latent factors (common factors) affecting the performance of genotypes across environments (Smith et al., 2001; Smith and Cullis 2018). This should be explored in future work.

4.3 Validating the training sets

Given the geographical proximity of field trial sites, it was surprising to find that the NOR training set predicted the phenotypes obtained in Swedish field trials so poorly. The CHE training set was the largest (392 accessions), and it contained 41 of the 42 EUCLEG accessions tested in Sweden. To ascertain whether this overlap was important, three training sets were compared with

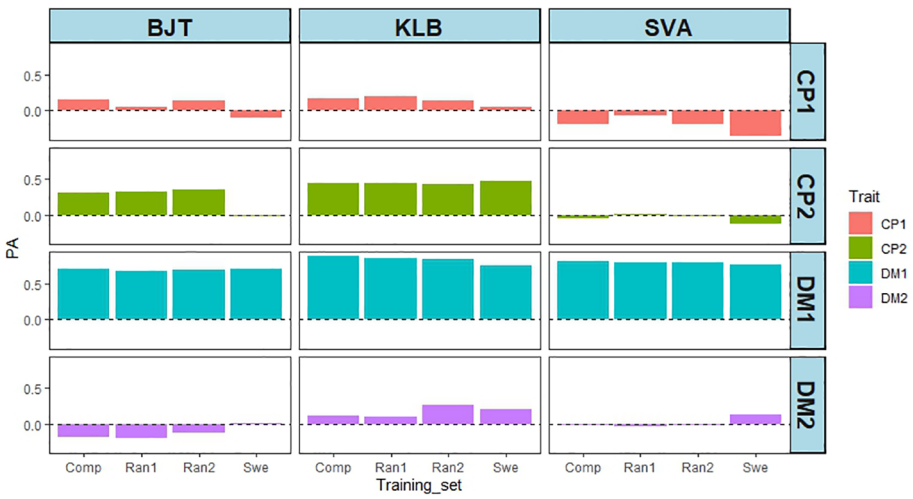


FIGURE 9
Effect of varying the training sets from CHE on PA from three locations in Sweden (BJT = Bjertorp; KLB = Kölbäck; SVA = Svalöv) for DMY in year 1 and 2, (DMY1 and DMY2) and CP content in cut 1 and cut 2 in year 1 (CP1 and CP2). Training sets were: Comp – Complete set of 392 accessions; Swe – Training set without the 42 accessions tested in Sweden; Ran1 and Ran2 – Training sets with 42 accessions removed randomly.

10.3389/fpls.2024.1407609

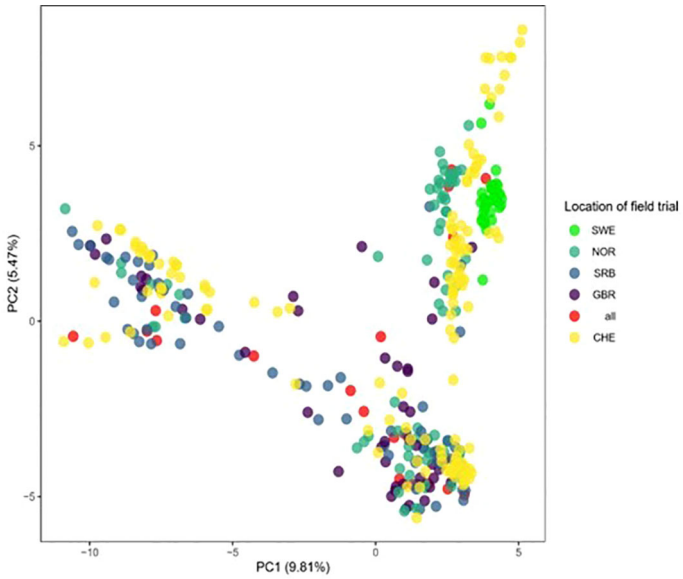


FIGURE 10
Principal component analysis (PCA) revealing the genetic relationship between the 392 accessions used in this work. CHE: Accessions that were exclusively assessed in Switzerland; GBR: Accessions exclusively assessed in GBR and CHE; NOR: Accessions exclusively tested in NOR and CHE; SRB: Accessions exclusively tested in SRB and CHE; all: Accessions tested in CHE, GBR, NOR and SRB; SWE: Accessions tested exclusively in Sweden and CHE.

the complete (Comp, 392 accessions) set: a training set in which the 42 accessions tested in all three locations in Sweden were omitted, and two training sets in which 42 accessions were removed randomly (Ran1 and Ran2). Figure 9 shows that reducing the size of the training set by 42 did not appear by itself to have any effect on PA, but removing the 42 accessions tested in Sweden did lower the PA for CP1, and particularly for CP2 in BJT (0.32 - -0.01). For DMY in year 1 there was also a slight decrease of PA in KLB from 0.84 to 0.75, and in SVA from 0.80 to 0.77. A similar analysis was not feasible for the GBR, NOR and SRB training sets, as they were much smaller and had varying numbers of accessions overlapping with the 42 from Sweden (3 from GBR, 5 from SRB, and 22 from NOR).

The genetic relationship between training set and test set is important for accuracy of genomic predictions (Liu et al., 2015). We used the principal component analysis (PCA) carried out by Nay et al. (2023) to illustrate the genetic relationship between the training sets with accessions from CHE, GBR, NOR and SRB with the test set from Sweden (Figure 10). There is a sizeable number of accessions tested in NOR that are closely related with the accessions tested in Sweden (in addition to the 22 accessions that are in common with those tested in Sweden). In contrast, the GBR and SRB accessions are more dispersed among the 392 accessions. Together with the results shown in Figure 9, it suggests that genetic relationship *per se* does not explain why the NOR training set predicted the phenotypes obtained in the Swedish field trials so poorly. What might explain it is the effect of leaving one breeders' material out. Leaving out the Lantmännen breeders' material does not reduce the PA by much relative to the random control (0.85 to 0.77) in year 1 (Figure 7B), which could be explained by the presence of related accessions in both training and test set. However, that does not explain why the PAs from year 2 reduced to 0.31 on average. Factors related to the environment and the genetics of the trait must be playing a role.

Overall, the present work has demonstrated the importance of correlation between phenotypic traits for PA. We therefore investigated the correlation between the phenotypic data from the populations from CHE, GBR, NOR and SRB with the test sets from Sweden. This could only be ascertained fully using the CHE population as it encompassed 41 of the 42 accessions tested in Sweden. Table 4 shows that DMY in year 1 was the only trait for which there was a significant positive correlation between the phenotypic values in Sweden and the corresponding phenotypic values from CHE. This was also the only trait for which the PA was very high. This would appear to underline the importance of phenotypic correlation for PA.

TABLE 4 Pearson correlation between phenotypes from the three Swedish locations (BJT, Bjertorp; KLB, Kölbäck; SVA, Svalöv) and the phenotypic data from the corresponding CHE traits.

	BJT		KLB		SVA	
Trait	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
DMY1	0.577***		0.824***		0.698***	
DMY2		-0.447		-0.652		-0.749
	Cut 1	Cut 2	Cut 1	Cut 2	Cut 1	Cut 2
CP1	0.131		0.129		-0.035	
CP2		-0.333		-0.243		0.052

Correlations with statistically significant ($P< 0.001$) positive values are indicated with ***.

5 Conclusions

This work is the first evaluation of GP in red clover. It shows that PAs were high for DMY and DOF, but mostly low for CP content. The results probably reflect differences in heritability, prediction bias, and correlation between traits. A lower number of markers in the models resulted in lower PAs, but only when they dropped below 1000 markers. Similarly, increasing the test set size at the expense of the training set size also reduced PA, but only when the training set size dropped to 10%. Such high PA values may be caused by the population structure present in the diverse red clover panel used here, because genetically related accessions are present in both training and test sets. Another important factor enhancing PAs seems to be a positive correlation between phenotypic traits in the training and test sets.

The prediction models incorporated GxE by capturing MxE interaction effects, which overall enhanced PA. This has perspectives for identifying markers with effects that are stable across environments, and those that have environment-specific effects.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA842231.

Author contributions

LS: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing, Funding acquisition. MMN: Formal Analysis, Investigation, Writing – review & editing. CG: Conceptualization, Formal Analysis, Investigation, Methodology, Writing – review & editing. LAF: Formal Analysis, Investigation, Writing – review & editing, Conceptualization, Data curation, Validation. MP: Formal Analysis, Investigation, Methodology, Writing – review & editing. LO: Data curation, Investigation, Methodology, Writing – review & editing. HA: Investigation, Writing – review & editing, Data curation, Resources. JR: Investigation, Writing – review & editing. LJ: Investigation, Writing – review & editing. AP: Conceptualization, Data curation, Resources, Writing – review & editing. TR: Writing – review & editing, Formal Analysis, Investigation, Methodology. DL: Funding acquisition, Investigation, Writing – review & editing. CH: Funding acquisition, Resources, Supervision, Writing – review & editing. RK: Data curation, Formal Analysis, Funding acquisition, Investigation, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project has received funding from the European Union's Horizon 2020 Program for Research and Innovation under grant agreement no. 727312, project EUCLEG. CJH and LS were also supported by BBSRC grants BBS/E/IB/230001 and BBS/E/W/0012843.

Acknowledgments

We thank the EUCLEG consortium for the valuable training and exchange sessions and all institutions providing red clover seeds, including: AgResearch (NZL), Agricultural Res. Ltd. (CZE), Agroscope (CHE), Boreal (FIN), DLF Seeds (CZE), DSV (DEU), Graminor (NOR), HBLFA (AUT), Hokkaido Ag. Res. (JPN), IBERS (GBR), IFVCNS (SRB), ILVO (BEL), INRAE (FRA), Lantmännen (SWE), NordGen (SWE), PGG Wrightson (NZL), RAGT2n (FRA), and USDA (USA). In addition, we would like to thank the technical staff at Aberystwyth University, Agroscope, DLF Seeds, Graminor, Institute for Forage Crops Kruševac and Lantmännen for their management of the field trials.

Conflict of interest

Author HA was employed by the company Graminor Breeding Ltd and author LJ was employed by the company DLF Seeds.

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 author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1407609/full#supplementary-material

References

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16, 1–13. doi: 10.1186/s12864–015-2212-y

Annicchiarico, P., Nazzicari, N., Wei, Y., Pecetti, L., and Brummer, E. C. (2017). Genotyping-by-sequencing and its exploitation for forage and cool-season grain legume breeding. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00679

Arojju, S. K., Cao, M., Zulfi Jahufer, M. Z., Barrett, B. A., and Faville, M. J. (2020). Genomic predictive ability for foliar nutritive traits in perennial ryegrass. *G3.: Genes|Genomes|Genetics.* 10, 695–708. doi: 10.1534/g3.119.400880

Arojju, S. K., Conaghan, P., Barth, S., Milbourne, D., Casler, M. D., Hodkinson, T. R., et al. (2018). Genomic prediction of crown rust resistance in Lolium perenne. *BMC Genet.* 19, 35. doi: 10.1186/s12863–018-0613-z

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3, e3376. doi: 10.1371/journal.pone.0003376

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

Barre, P., Asp, T., Byrne, S., Casler, M., Faville, M., Rognli, O. A., et al. (2022). Genomic prediction of complex traits in forage plants species: perennial grasses case. *Methods Mol. Biol.* 2467, 521–541. doi: 10.1007/978–1-0716–2205-6_19

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., and Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Sci. (Oxford).* 242, 23–36. doi: 10.1016/j.plantsci.2015.08.021

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

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PloS One* 8, e57438. doi: 10.1371/journal.pone.0057438

Byrne, S. L., Conaghan, P., Barth, S., Arojju, S. K., Casler, M., Michel, T., et al. (2017). Using variable importance measures to identify a small set of SNPs to predict heading date in perennial ryegrass. *Sci. Rep.* 7, 3566. doi: 10.1038/s41598–017-03232–8

Campbell, M. T., Hu, H., Yeats, T. H., Brzozowski, L. J., Caffe-Treml, M., Gutiérrez, L., et al. (2021). Improving genomic prediction for seed quality traits in oat (Avena sativa L.) using trait-specific relationship matrices. *Front. Genet.* 12. doi: 10.3389/fgene.2021.643733

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., et al. (2018). Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (Lolium perenne L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00369

Crossa, J., De Los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., and Pérez-Rodríguez, P. (2016). Extending the marker × Environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Sci.* 56, 2193–2209. doi: 10.2135/cropsci2015.04.0260

Cuevas, J., Crossa, J., Soberanis, V., Perez-Elizalde, S., Perez-Rodriguez, P., de los Campos, G., et al. (2016). Genomic prediction of Genotype x Environment interaction Kernel regression model. *Plant Genome* 9. doi: 10.3835/plantgenome2016.03.0024

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., et al. (2020). Hybrid breeding of rice via genomic selection. *Plant Biotechnol. J.* 18, 57–67. doi: 10.1111/pbi.13170

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

Ergon, Å., Milvang, Ø.W., Skøt, L., and Ruttink, T. (2022). Identification of loci controlling timing of stem elongation in red clover using genotyping by sequencing of pooled phenotypic extremes. *Mol. Genet. Genomics* 297, 1587–1600. doi: 10.1007/s00438–022-01942-x

Ergon, Å., Skøt, L., Sæther, V. E., and Rognli, O. A. (2019). Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (*Trifolium pratense* L.). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00718

Faville, M. J., Cao, M., Schmidt, J., Ryan, D. L., Ganesh, S., Jahufer, M. Z. Z., et al. (2020). Divergent genomic selection for herbage accumulation and days-to-heading in perennial ryegrass. *Agronomy* 10, 340. doi: 10.3390/agronomy10030340

Faville, M. J., Ganesh, S., Cao, M., Jahufer, M. Z. Z., Bilton, T. P., Easton, H. S., et al. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor. Appl. Genet.* 131, 703–720. doi: 10.1007/s00122–017-3030–1

Fè, D., Ashraf, B. H., Pedersen, M. G., Janss, L., Byrne, S., Roulund, N., et al. (2016). Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *Plant Genome* 9, 3. doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B., Pedersen, M., et al. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics* 16, 1–15. doi: 10.1186/s12864–015-2163–3

Forster, J. W., Hand, M. L., Cogan, N. O. I., Hayes, B. J., Spangenberg, G. C., and Smith, K. F. (2014). Resources and strategies for implementation of genomic selection in breeding of forage species. *Crop Pasture Sci.* 65, 1238–1247. doi: 10.1071/CP13361

Frame, J., Charlton, J. F. L., and Laidlaw, A. S. (1998). *Temperate forage Legumes* (Wallingford: CAB International), 327, ISBN: .

Frey, L. A., Vleugels, T., Ruttink, T., Schubiger, F. X., Pégard, M., Skøt, L., et al. (2022). Phenotypic variation and quantitative trait loci for resistance to southern anthracnose and clover rot in red clover. *Theor. Appl. Genet.* 135, 4337–4349. doi: 10.1007/s00122–022-04223–8

Griffiths, A., Ehoche, G., Arojju, S., Larking, A., Jauregui, R., Cousins, G., et al. (2022). Developing genomic selection for dry matter yield in white clover. *J. New Z. Grasslands.* 83, 83–90. doi: 10.33584/jnzg.2021.83.3502

Grinberg, N. F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K. P., Kelly, R., et al. (2016). Implementation of genomic prediction in Lolium perenne (L.) breeding populations. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00133

Guo, 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

Hayes, B. J., Cogan, N. O. I., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C., et al. (2013a). Prospects for genomic selection in forage plant species. *Plant Breed.* 132, 133–143. doi: 10.1111/pbr.12037

Hayes, B. J., Lewin, H. A., and Goddard, M. E. (2013b). The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends Genet.* 29, 206–214. doi: 10.1016/j.tig.2012.11.009

Heslot, N., Akdemir, D., Sorrells, M., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480. doi: 10.1007/s00122–013-2231–5

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

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., et al. (2018). Genomic prediction for 25 agronomic and quality traits in alfalfa (Medicago sativa). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01220

Juliana, P., Singh, R. P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., et al. (2020). Genomic selection for grain yield in the CIMMYT wheat breeding program—Status and perspectives. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.564183

Keep, T., Sampoux, J. P., Blanco-Pastor, J. L., Dehmer, K. J., Hegarty, M. J., Ledauphin, T., et al. (2020). High-throughput genome-wide genotyping to optimize the use of natural genetic resources in the grassland species perennial ryegrass (Lolium perenne l.). *G3* 10, 3347–3364. doi: 10.1534/g3.120.401491

Lee, M. R. F. (2014). Forage polyphenol oxidase and ruminant livestock nutrition. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00694

Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., et al. (2015). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8, 90. doi: 10.3835/plantgenome2014.12.0090

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, 46. doi: 10.3835/plantgenome2015.06.0046

Lin, Z., Wang, J., Cogan, N. O. I., Pembleton, L. W., Badenhorst, P., Forster, J. W., et al. (2017). Optimizing resource allocation in a genomic breeding program for perennial ryegrass to balance genetic gain, cost, and inbreeding. *Crop Sci.* 57, 243–252. doi: 10.2135/cropsci2016.07.0577

Liu, H., Zhou, H., Wu, Y., Li, X., Zhao, J., Zuo, T., et al. (2015). The impact of genetic relationship and linkage disequilibrium on genomic selection. *PloS One* 10, e0132379. doi: 10.1371/journal.pone.0132379

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.: Genes|Genomes|Genetics.* 5, 569–582. doi: 10.1534/g3.114.016097

Lüscher, A., Mueller-Harvey, I., Soussana, J. F., Rees, R. M., and Peyraud, J. L. (2014). Potential of legume-based grassland–livestock systems in Europe: a review. *Grass. Forage. Sci.* 69, 206–228. doi: 10.1111/gfs.12124

McKenna, P., Cannon, N., Conway, J., and Dooley, J. (2018). The use of red clover (*Trifolium pratense*) in soil fertility-building: A Review. *Field Crops Res.* 221, 38–49. doi: 10.1016/j.fcr.2018.02.006

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. doi: 10.1093/genetics/157.4.1819

Nay, M. M., Grieder, C., Frey, L. A., Amdahl, H., Radovic, J., Jaluvka, L., et al. (2023). Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1128823

Pégard, M., Barre, P., Delaunay, S., Surault, F., Karagić, D., Milić, D., et al. (2023). Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1196134

Pembleton, L. W., Inch, C., Baillie, R. C., Drayton, M. C., Thakur, P., Ogaji, Y. O., et al. (2018). Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. *Theor. Appl. Genet.* 131, 1891–1902. doi: 10.1007/s00122–018-3121–7

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

Puglisi, D., Delbono, S., Visioni, A., Ozkan, H., Kara, İ., Casas, A. M., et al. (2021). Genomic prediction of grain yield in a barley MAGIC population modeling genotype per environment interaction. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.664148

R Core Team (2023). *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing).

Sabag, I., Bi, Y., Peleg, Z., and Morota, G. (2023). Multi-environment analysis enhances genomic prediction accuracy of agronomic traits in sesame. *Front. Genet.* 14. doi: 10.3389/fgene.2023.1108416

Schefers, J. M., and Weigel, K. A. (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. *Anim. Front.* 2, 4–9. doi: 10.2527/af.2011–0032

Smith, A., and Cullis, B. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214, 143. doi: 10.1007/s10681–018-2220–5

Smith, A., Cullis, B., and Thompson, R. (2001). Analyzing Variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006–341X.2001.01138.x

Sukumaran, S., Jarquin, D., Crossa, J., and Reynolds, M. (2018). Genomic-enabled prediction accuracies increased by modeling genotype × Environment interaction in durum wheat. *Plant Genome* 11, 170112. doi: 10.3835/plantgenome2017.12.0112

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

Walsh, B., and Lynch, M. (2018). *Evolution and selection of quantitative traits Vol. 1* (Oxford, UK: Oxford University Press).

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., et al. (2020). How population structure impacts genomic selection accuracy in cross-validation: Implications for practical breeding. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.592977

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: the USDA experience. *Annu. Rev. Anim. Biosci.* 5, 309–327. doi: 10.1146/annurev-animal-021815–111422

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., et al. (2021). Genomic selection: A breakthrough technology in rice breeding. *Crop J.* 9, 669–677. doi: 10.1016/j.cj.2021.03.008

Zanotto, S., Ruttink, T., Pégard, M., Skøt, L., Grieder, C., Kölliker, R., et al. (2023). A genome-wide association study of freezing tolerance in red clover (*Trifolium pratense* L.) germplasm of European origin. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1189662

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H., Longin, F., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi: 10.1007/s00122–011-1745-y

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 April 2025
DOI 10.3389/fpls.2025.1422162

Check for updates

OPEN ACCESS

EDITED BY
Francis Chuks Ogbonnaya,
Grains Research and Development Corporation, Australia

REVIEWED BY
Mariola Staniak,
Institute of Soil Science and Plant Cultivation, Poland
Åshild Ergon,
Norwegian University of Life Sciences, Norway

*CORRESPONDENCE
Tomáš Vymyslický
vymyslicky@vupt.cz

RECEIVED 23 April 2024
ACCEPTED 22 January 2025
PUBLISHED 07 April 2025

CITATION
Vymyslický T, Trněný O, Rietman H, Balko C, Đorđević V, Ranđelović P and Dybová M (2025) Phenotypic characterization of soybean genetic resources at multiple locations: breeding implications for enhancing environmental resilience, yield and protein content.
Front. Plant Sci. 16:1422162.
doi: 10.3389/fpls.2025.1422162

COPYRIGHT
© 2025 Vymyslický, Trněný, Rietman, Balko, Đorđević, Ranđelović and Dybová. 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.

Phenotypic characterization of soybean genetic resources at multiple locations: breeding implications for enhancing environmental resilience, yield and protein content

Tomáš Vymyslický[1]*, Oldřich Trněný[1], Hendrik Rietman[2], Christiane Balko[3], Vuk Đorđević[4], Predrag Ranđelović[4] and Magdaléna Dybová[1]

[1]Agricultural Research, Ltd., Troubsko, Czechia, [2]Agro Seed Research bv, Opglabbeek, Belgium, [3]Julius Kühn-Institut, Federal Research Centre for Cultivated Plants, Institute for Breeding Research on Cultivated Plants, Sanitz, Germany, [4]Institute for Field and Vegetable Crops, Novi Sad, Serbia

Introduction: Soybean is an important legume crop and a leading source of dietary protein and oil in animal feed, as well as an important food for human consumption. The objective of our research was to study soybean genetic resources in context of future protein self-sufficiency both in human and animal nutrition.

Methods: Collection of 360 different accessions from various regions worldwide was evaluated across four European locations during two consecutive years in phenotyping trials. The five most important traits of soybean – plant emergence, plant length, protein content, seed yield, and R8 stage – were carefully analysed, revealing significant variability.

Results: Ten exceptionally stable genotypes were identified based on their protein content and yield, presenting promising candidates for breeding programs.

Discussion: Our findings underscore the importance of integrating genotype-environment interaction analyses into breeding initiatives, considering the observed variability in phenotypic traits across diverse environments and genotypes.

KEYWORDS

***Glycine max*, field trials, morphological and phenological traits, weather data, genotype-environment interactions**

Introduction

Plant Genetic Resources for Food and Agriculture are defined as diversity of genetic material contained in traditional varieties and modern cultivars, as well as crop wild relatives and other wild plant species that can be used now or in the future for food and agriculture (Halewood et al., 2018). Preservation of biodiversity and natural variation within species has become a global concern. The diversity within a species allows it to survive and adapt to new environments, new pests, and changing climates (Iriondo et al., 2008). Genetic resources are essential for maintaining and enhancing the efficiency and the resilience of production systems, as well as for healthy diets and the delivery of ecosystem services, such as pollination and pest and disease regulation (Halewood et al., 2018). Genetic resources are essential sources of genes for breeding. In an environmentally dynamic world, with constantly increasing population and limited resources, we need to conserve genetic diversity for our own food and environmental security (Iriondo et al., 2008). Agricultural production is dependent on genetic resources domesticated elsewhere and subsequently developed in other countries and regions. The erosion of these resources poses a severe threat to the world's food security in the long term (Ebert, 2020). The importance of genetic resources has also been increasing recently due to ongoing climate change. Crop breeders and their growers are increasingly forced to face climatic extremes precisely with the use of promising genetic resources in breeding processes (Muluneh, 2021).

Modern cultivated soybean was domesticated from the annual wild soybean (*G. soja* Sieb. & Zucc.) in East Asia 6000–9000 years ago (Kim et al., 2012). The distribution of wild soybean is limited to East Asia – China, Japan, Korea and part of Russia (Jeong et al., 2019). Within this region, China is considered to be the domestication centre of soybean. China has the earliest written records of soybean cultivation (Kim et al., 2012). Soybean has been found in unearthed artefacts across dynasties and provinces (Qiu and Chang, 2010). The crop was introduced to Europe in the 18th century and then to the United States (Hymowitz and Harlan, 1983). The five countries with the highest soybean production in the world are Brazil, United States, Argentina, China and India (FAOSTAT, 2023).

Soybean (*Glycine max* (L.) Merr.) is an important legume crop and a leading source of dietary protein and oil in animal feed, as well as a staple food for human consumption (Hartman et al., 2011). Soy protein is one of the major components of the livestock diet and it is increasingly important in the human diet. Soybean is a rich source of high-quality proteins containing all the essential amino acids found in animal proteins (Gorissen and Witard, 2018). Soybean seeds contain approximately 35 - 40% of proteins, 20% of lipids, 9% of dietary fibre and 8% of moisture (He and Chen, 2013). The content of starch in the seeds is about 6%, total soluble sugars is about 9% and sucrose about 7% (Sharma et al., 2014). Soybean and its products are rich sources of minor non-nutrient components with potential health benefits (Cicero et al., 2017), which are often referred to in the literature as phytochemicals (biologically active proteins and peptides, lectins, lunasin). Soybean products are important sources of isoflavones, phytosterols, phytic acid and saponins (Barnes, 2010). Because of high content of polyunsaturated fats, fibre, vitamins and minerals and low content of saturated fat, soybean products have many beneficial health and therapeutic effects (Zuo et al., 2023; Chen et al., 2019; Fan et al., 2022; Kao and Chen, 2006; Liu et al., 2005). Through fermentation by *Bacillus subtilis* or *Aspergillus oryzeae*, the nutritional value of fermented soybean products is enhanced. Fermentation improves the digestibility, and increases the soy protein and isoflavone profiles compared to non-fermented soy foods (Jayachandran and Xu, 2019).

The biggest commercial interest in soy is the protein and oil. In soybean, the protein content is negatively correlated with the oil content and the yield (Ohara and Shimamoto, 2002). Soybean protein is the only plant protein source, containing all the essential amino acids necessary for humans. Soybean protein does not contain cholesterol, and the contents of methionine and branched-chain amino acids are low compared with animal proteins (Gorissen and Witard, 2018).

Genetic resources play very important role in improving and breeding of all agricultural crops. In soybean, genetic resources play an important role in increasing the resistance and tolerance to abiotic and biotic stressors. The main aim of current breeding efforts is promoting the sustainable production of soybean worldwide. Resistant or tolerant genetic resources could be used as input materials in the forthcoming selection of soybean varieties with high protein content and high yield (Guo et al., 2022).

Soybean is currently the most important protein crop in Europe (Guilpart et al., 2022). EU soybean production is 2.7 million tons, while non-EU European countries produced another 8.4 million tons (EUROStat, 2021; FAOStat, 2021). European production covers only a small part of soybean demands. The European Union imported an annual average of 14 million tons of soybeans and 18 million tons of soybean cake (EUROStat, 2021). Most of the import is from America (USA, Argentina, Brasil). The high demand for soybean protein in Europe is therefore an important reason for expanding soybean cultivation to central and northern growing areas where it is a minor crop. European self-sufficiency in soybean would require 9 – 12% of its arable land to be sown to this crop (Guilpart et al., 2022). High yielding and resistant varieties of soybean, with stable yields in different environmental conditions and stable protein contents, are the basis for the future protein self-sufficiency of Europe (Gorissen and Witard, 2018).

Within the EUCLEG project "Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency" (www.eucleg.eu) promising genetic resources from different regions in the world were gathered and subjected to detailed phenotyping under the conditions of four different localities in Europe.

In order to provide a solid foundation for the utilization of soybean genetic resources in breeding programs, we aimed at (i) identification of promising soybean accessions for future breeding, (ii) characterizing the phenotypic diversity of agriculturally relevant traits at multiple locations and (iii) getting an insight into the extent of accession × location interaction for future adaptive breeding efforts. Based on the evaluation, the most promising materials were proposed for the further use in breeding process and discussed in the context of global climate change.

10.3389/fpls.2025.1422162

Materials and methods

Soybean genetic resources

A set of soybean genetic resources was gathered from the EUCLEG consortium members, breeders, research institutes and gene banks. Accessions were delivered with specific material transfer agreements from each supplier to a central coordinator, who further distributed seeds under the same agreements to the researchers performing the experiments. Altogether 360 accessions of different origin were evaluated in the trials. Most accessions originated from Europe (242 accessions), North America (44 accessions) and Asia (30 accessions). Some accessions had unknown origin (44 accessions). The most represented countries were Belgium (50 accessions), Serbia (38 accessions), Germany and Canada (33 accessions) and China (21 accessions).

Experimental sites

Four different localities of soybean trials representing different agroclimatic regions of Europe were chosen: Troubsko – Czech Republic (CZE), Novi Sad – Serbia (SRB), Groß Lüsewitz – Germany (GER) and Kessenich – Belgium (BEL). Localities Troubsko and Novi Sad belong to Pannonian region, Kessenich belongs to North Maritime region and the locality Groß Lüsewitz belongs to Continental zone (Ceglar et al., 2019).

The same trials were conducted both in 2018 and in 2019. The plot size used in the trials was 5 m^2. Sowing density was 75 seeds/m^2 for maturity group MG 000, 65 seeds/m^2 for the MG 00, 55 seeds/m^2 for the MG 0 and 45 seeds/m^2 for the MG I/II. Information about maturity groups was obtained from the breeders or sales representatives of the variety. Plants for the evaluation were taken from the inner part of the plot. Irrigation was only applied at the Kessenich location in 2018. The amount of water used for irrigation was included in the rainfall totals shown in Table 1. Herbicide treatment was applied if necessary. Fertilisation was applied in each year before sowing using the following quantities: N – P – K 45 – 45 – 45 kg.ha^{-1}.

Experimental design

A partially replicated plot design was used for all the localities, see Table 1. GER and CZE locality had 100 soybean accessions and BEL and SRB locality had 360 accessions, in one, two or four replications. Border plots were used in order to avoid the marginal effect.

For the inoculation the product NITRAZON was used (Farma Žiro, Ltd. Czech Republic). In all locations except Novi Sad, soybean seeds were inoculated before sowing. In Novi Sad the inoculation was not necessary, due to the long-term cultivation of soybeans and its representation in crop rotations. Therefore, there was a sufficient supply of *Rhizobium* bacteria in the soil at this location.

We used different seeding densities for each maturity group, because varieties with a later ripening period have a longer growing season and the plants reach larger size. A detailed description of this collection is provided in Supplementary Table 1.

Evaluated morphological and phenological data

The traits listed in Table 2 were evaluated. Phenological stages were evaluated using the BBCH scale for soybean (Munger et al., 1997). The protein content, was uniformly assessed for all genotypes at the EV-ILVO institute in Belgium employing NIRS (Near Infra Red Spectrometry), using the methodology described by Zhu et al., 2018.

TABLE 1 Layout of field trials at the different locations.

Parameter	Trial location							
	BEL	BEL	GER	GER	CZE	CZE	SRB	SRB
	2018	2019	2018	2019	2018	2019	2018	2019
Number of accessions	360	360	100	100	100	100	360	360
– with 1 plot	320	320	76	76	76	76	320	320
– with 2 plots	32	32	16	16	16	16	32	32
– with 4 plots	8	8	8	8	8	8	8	8
Number of plots	432	432	140	140	140	140	432	432
Plot size	1 x 5 m	1 x 5 m	1,5 x 3,35 m	1,5 x 3,35 m	1.25 x 4 m	1.25 x 4 m	2 x 4 m	2 x 4 m
Distance between rows	0.25 m	0.25 m	0.25 m	0.25 m	0.25 m	0.25 m	0.5 m	0.5 m
Sowing date	7. 5.	10. 5.	3. 5.	7. 5.	3. 5.	16. 4.	19. 4.	7. 4.
Inoculum	Yes	Yes	Yes	Yes	Yes	Yes	No	No

Data analyses

To analyze the five traits considered to be the most important (plant emergence, plant length, date of reaching the R8 stage, seed yield and protein content), we employed a combination of statistical approaches. First, we built a linear mixed model, from which we performed variance component analysis and derived BLUPs (Best Linear Unbiased Predictions). Next, we conducted principal component analysis (PCA) to explore the structure phenotypic expression of genotypes, followed by redundancy analysis (RDA) to assess the influence of genetic structure on trait variation. Finally, we used Bayesian regression to analyze genotype-environment interactions (GxE).

We evaluated traits results using a linear mixed models to account for both fixed and random sources of variation. The results for the selected traits were used as the predicted variables to set up the models. Based on these models, we assessed variance components to investigate the influence of genotype on these five plant traits in deferent trials. The mixed model analysis was conducted using the R package "lme4" (Bates et al., 2015). The seed yield data were standardized to z-scores by subtracting the mean of each variable and dividing by its standard deviation, which transforms the data to have a mean of 0 and a standard deviation of 1 to enhance the stability of model parameter estimation. Each trial assignment was used as an independent fixed effect in the model. Random effects were included to account for genetic variability of genotypes and to control for the influence of row and column positions within each trial. Geographic location, season, and their interaction were not considered due to the significant variation in environmental conditions across both years. Statistical significance of random and fixed effects was assessed using backward elimination procedure implemented in lmerTest R package (Kuznetsova et al., 2017). The step function testing random effects with likelihood ratio tests and fixed effects with F-tests using Satterthwaite's method for degrees of freedom. Weather characteristics were initially included as fixed effects but were removed from the final model due to non-significance during the model selection procedure. Convergence of model estimates was verified using diagnostic plots, including residuals versus fitted values and quantile-quantile plots, to ensure the assumptions of normality and homoscedasticity were met. To estimate the marginal means of the trials with standard errors (SE) and 95% confidence intervals (CI), we used the *emmeans* package in R (Lenth, 2024). Generalized heritability (Cullis et al., 2006) was computed based on genotypic variance and BLUP conditional variance, appropriate for linear mixed models in partially replicated and unbalanced plot designs.

Formula 1: Formula of used linear mixed models to predict and analyse selected traits

$$y_{ijkl} = \beta_0 + \beta_i Trial_i + u_j + v_{ik} + w_{il} + \varepsilon_{ijkl}$$

In this model, y_{ijkl} represents the outcome for a given trait, with β_0 as the intercept. The fixed effect $\beta_i \mathrm{Trial_i}$ captures differences associated with the ith trial. Random effects include u_j representing genetic variation due to genotype j, v_{ik} which accounts for variability across the kth row within the ith trial, and w_{il} representing the effect of the lth column within the ith trial. Finally, ε_{ijkl} is the residual error term.

BLUPs were derived from the mixed models to estimate the genetic performance of each genotype, accounting for other effects and sources of variability included in the models. Combined correlation plots of BLUPs (Best linear unbiased prediction) were generated utilizing the Performance Analytics R package. BLUPs are used in linear mixed models for the estimation of random effects. Pearson correlation coefficients and the statistical significance tests among variables were computed using the cor.test function from the stats package. Principal component analysis (PCA) was conducted using prcomp function from R Stats 4.0.3 package on unit variance scaled BLUPs in R. PCAs was plotted using fviz_pca package.

Redundancy Analysis (RDA) was conducted to quantify the influence of genetic groups on phenotypic traits. The genetic assignment based on genotyping using the 355K SoySNP microarray (Wang et al., 2016) from Saleem et al. (2021) was used to categorize analysed genotypes into five distinct genetic groups and an admixed group. RDA was conducted using the Vegan 2.5-7 R package, utilizing Z-score scaled BLUPs of R8, protein content, plant length, plant emergence, and seed yield traits. Genetic cluster assignments were employed as constrained variables. The biplot function was used to generate the plot, with scaling for the genetic groups (scaling = 2).

The stability of genotypic performance across our multi-environment trials was assessed utilizing a Bayesian Finlay-Wilkinson regression framework, as implemented in the FW package for R (Lian and de los Campos, 2016). This approach allows for the nuanced characterization of genotype-environment interactions by estimating model parameters that reflect genotypic performance in response to environmental conditions. The analysis integrated a covariance matrix to adjust for genotypic similarities among accessions, using a marker-derived kinship matrix, and considered environmental variability among trials using a covariance matrix of scaled climatic data (Table 1). Bayesian estimations of environmental and genotype model parameters were conducted using the GibbsA method, specified in the FW package with the following settings: 1,000,000 iterations for thorough exploration of the parameter space, a burn-in period of 100,000 to ensure model stabilization, and a thinning interval of 10,000 to reduce autocorrelation in the sampled parameters.

Results

Weather data

Weather data from all the experimental sites were collected and are presented in Table 3. From the presented data the clear difference between the two with continental climate (Novi Sad and Troubsko) and two with oceanic climate (Kessenich and

TABLE 2 Overview of evaluated morphological and phenological traits.

No.	Observation	Score	Remarks
1	V-stage at emergence	None – VE – VC	BBCH scale for soybean (Munger et al., 1997).
2	Plant emergence	1-9	1 low, 9 high
3	Plant vigour at emergence	1-5	1 low, 5 high
4	V2 stage	date	Date when 50% of the plants of a plot are in V2. BBCH scale for soybean (Munger et al., 1997).
5	Plant height at V2	cm	BBCH scale for soybean (Munger et al., 1997).
6	R1 stage	date	Days from sowing. BBCH scale for soybean (Munger et al., 1997).
7	R2 stage	date	Days from sowing. BBCH scale for soybean (Munger et al., 1997).
8	Diseases, pests	1-9	1 absence, 9 strong attack
9	Abiotic stress	1-9	1 no stress, 9 severe stress
10	R8 stage	date	Days from sowing. BBCH scale for soybean (Munger et al., 1997).
11	Lodging at R8	1-5	1 no lodging, 5 abundant lodging. BBCH scale for soybean (Munger et al., 1997).
12	Seed yield	$kg.ha^{-1}$	Expressed at 14% moisture content. Seed yield was calculated in $kg.ha^{-1}$ from each experimental plot.
13	Moisture content	%	Drying for 72 hours at 70°C
14	Protein content	%/dry matter	Analysed by NIRS
15	Seed weight	g	4 x 50 seeds (expressed at 14% moisture content)
16	Plant length	cm	5 plants/plot
17	Height first pod	cm	5 plants/plot
18	Mottled seeds	%	From the dried sample
19	Node number on the main stem of the plant	counting	5 plants per plot
20	Number of branches with pods	counting	5 plants per plot
21	Distribution of pods on the plant	1-5	5 plants per plot, 1 - pods only in the basal part, 5 - pods distributed evenly on the plant
22	Seed number	counting	5 plants per plot (after threshing)
23	Seed weight per plant	g	5 plants per plot (after threshing) calculated

Groß Lüsewitz) can be seen. The difference is clearly visible in the temperatures, average temperature of growing season (1.4.-31.10.) varied from 14,2°C (Groß Lüsewitz, 2019) to 20,1°C (Novi Sad, 2018). The sum of effective temperatures above 10°C varied from 2048 (Groß Lüsewitz, 2019) to 3490 (Novi Sad, 2018). The day of the last frost from 1. 4. varied from 0 (more localities) to 44 days (Groß Lüsewitz, 2019). The day of the first frost before 1. 11. varied from 0 (more localities) to 26 (Groß Lüsewitz, 2019). Based on temperatures, Novi Sad was the warmest location, while Groß Lüsewitz was the coldest. Precipitation was distributed rather randomly, with significant drought recorded in 2018 in the localities Troubsko and Kessenich. If we focus on precipitation, two parameters are shown in the Table 2 – sum of precipitation in the period 1. 4. – 31. 10. (soybean growing season) and sum of precipitation in the period 1. 4. – 31. 5. (emergence period). While in the whole growing period the driest was the locality Kessenich (188 mm), in the emergence period the driest was the locality Troubsko (47 mm). The driest locality was Kessenich in 2018, while the wettest location was Novi Sad in 2019. In 2018, only the Kessenich site received irrigation. However, temperatures and precipitation need to be considered together, because higher temperatures mean higher evaporation.

Evaluated traits and their variability

The summarising analysis of plant emergence, plant length, protein content, R8 stage, and seed yield as delineated in Table 4 offers insights into the variability of these traits across different locations and growing seasons. Original measured values are in Supplementary Table 2. The trial conducted in Kessenich during the 2019 growing season showcased the most favourable outcomes. This particular trial is indicative of outcomes approaching ideal cultivation conditions, thereby elucidating the genotypes' yield

10.3389/fpls.2025.1422162

TABLE 3 The most important weather data of all the experimental sites.

Country	BEL		CZE		GER		SRB	
Locality	Kessenich		Troubsko		Groß Lüsewitz		Novi Sad	
Coordinates	51°08′N, 5°48′E		49°10′N, 16°30′E		54°07′N, 12°32′E		45°20′N, 19°51′E	
Year	2018	2019	2018	2019	2018	2019	2018	2019
Average temperature in the growing season (1. 4. - 31. 10.)	16.8°C	15.6°C	17.6°C	15.8°C	15.3°C	14.2°C	20.1°C	19°C
Sum of effective temperatures above 10°C during the growing season (1. 4. - 31. 10.)	2741	2364	3066	2621	2379	2048	3490	3281
Average temperature in the summer full growing season (1. 7. - 31. 8.)	20.6°C	19.3°C	22.2°C	20.6°C	19.2°C	18.1°C	23.2°C	23.9°C
Day of last frost after 1. 4.	0	13	6	37	33	44	0	0
Day of first frost before 1. 11.	0	0	0	24	13	26	0	0
Sum of precipitation in the growing season (1. 4. - 31. 10.)	188 mm	294 mm	238 mm	380 mm	263 mm	417 mm	226 mm	439 mm
Sum of precipitation in the emergence period (1. 4. - 31. 5.)	72 mm	55 mm	47 mm	95 mm	75 mm	57 mm	113 mm	202 mm
Sum of precipitation in the summer full growing season (1. 7. - 31. 8.)	51 mm	64 mm	55 mm	116 mm	80 mm	135 mm	78 mm	100 mm

potential in the absence of abiotic stressors. Conversely, the 2018 Troubsko trial exhibited the least favourable results, a consequence of several adverse factors, most notably a deficiency in moisture during the initial stages of plant emergence followed by an excessively dry season. Despite these unfavourable conditions, the results from the Troubsko trial are invaluable for identifying genotypes with exceptional resilience, capable of contributing positively under less-than-ideal circumstances. A notable disparity was observed between the highest yielding trial and the least productive trial. The intermediate trials have conspicuously demonstrated the genotypes' potential under the standard growth conditions prevalent at the respective locations.

A significant variation in protein content was observed across genotypes, years, and locations. The most notable values were recorded at Troubsko and Novi Sad, with figures reaching up to 50%. The average protein content values derived from the trials ranged between 38% and 45%. Trials conducted within the same locality yielded similar mean values. For detailed results, see Table 4.

The correlation analysis of R8 stage, protein content, and seed yield (Figure 1) across different trials elucidates the general relationship among these trials. Higher correlation values suggest a diminished influence of genotype-environment interactions on trait values. Conversely, lower correlations between trials indicate variability in genotype responses due to environmental influences within each trial. The R8 stage trait exhibited the most consistent positive correlation across trials, except for those from the Troubsko 2018 trial, showing a predominantly strong correlation. For plant length and seed yield, medium to strong and consistent correlations

TABLE 4 Summary of mean values and standard deviations (SD) of raw measured trait values across all trials.

		Plant emergence class	Plant length [cm]	Seed protein content [%]	R8 phase	Seed yield [kg.ha^{-1}]
Groß Lüsewitz	2018	6.3 (1.5)	58.2 (17.2)	40.1 (1.8)	118.4 (25.2)	2760 (855)
GER	2019	6.3 (1.2)	71.2 (16.1)	40.7 (2.1)	133.4 (17.1)	2902 (594)
Novi Sad	2018	7.2 (1.4)	63.7 (20.9)	42 (1.7)	112.2 (14.2)	2414 (833)
SRB	2019	8.9 (0.7)	66.1 (19.2)	42.7 (2.1)	132.2 (12.8)	2244 (800)
Kessenich	2018	7.7 (2)	54.3 (11.5)	37.9 (3.1)	121.1 (14.5)	2761 (851)
BEL	2019	9.5 (1.1)	98.7 (20.9)	37.9 (2.2)	137 (16.6)	4204 (1311)
Troubsko	2018	5.7 (0.8)	42.9 (6.4)	44.5 (1.6)	166.8 (12.9)	385 (235)
CZE	2019	8 (0.7)	64.8 (12.4)	41.3 (2.6)	167.2 (16.7)	1320 (612)

SD are given in parentheses. R8 are days from sowing to the maturity stage.

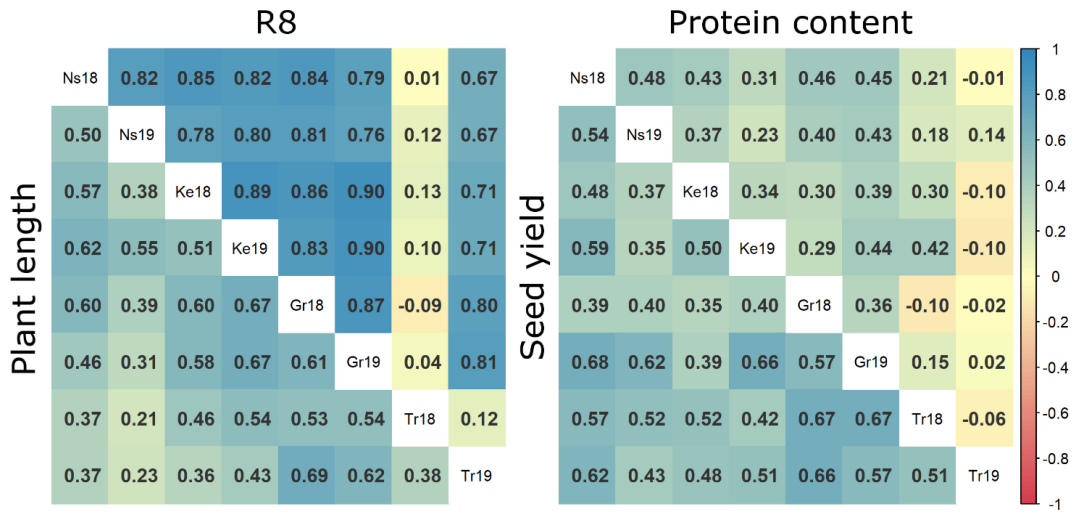


FIGURE 1
Comparative heatmaps illustrating Pearson correlation coefficients for measured data of plant length, R8, seed yield and protein content across various crop trials. Ns18 = Novi Sad 2018, Ns19 = Novi Sad 2019, Ke18 = Kessenich 2018, Ke19 = Kessenich 2019, Gr18 = Groß Lüsewitz 2018, Gr19 = Groß Lüsewitz 2019, Tr18 = Troubsko 2018, Tr19 = Troubsko 2019.

were observed across all trials, including those under extreme conditions, without any exceptions. In contrast, protein content outcomes displayed the weakest correlation among trials. While the majority of trials exhibited weak to moderate positive correlations, the results from the Troubsko 2019 trial showed no discernible correlation with those of other trials for protein content.

Linear mixed models

To analyze the individual factors and variance components influencing the results of the multi-trial experiment, we fitted a linear mixed regression model. From this model, we estimated the effects of the trials and calculated the Best Linear Unbiased Predictor (BLUP) (Supplementary Table 3) for four key soybean traits: seed yield, R8 stage, protein content, and plant height. Summary characteristics of the fitted models are provided in Supplementary Table 4.

The linear mixed regression models for plant length, protein content, R8 stage, and seed yield revealed statistically significant effects ($\alpha < 0.01$) attributed to all fixed and random effects included in the models. The maximum likelihood estimation process successfully converged for each trait's mixed linear model, indicating that was found stable and optimal estimates for both fixed effects and variance components. Variance components showed substantial contributions from different sources, including genotype, position in trial and residuals. Furthermore, the genotype variance component was higher than the variance components of random effects originating from plot rows and columns. The fixed effects on the traits showed notable variation across trials, reflecting primarily the differences in environmental conditions. All models exhibit residuals with a mean close to zero and normal distribution, indicating that the models effectively captured the variation.

Environmental conditions significantly modulated crop performance across trial sites. Statistical testing of fixed effects revealed a significant trial-dependent effect on all four traits. The Troubsko 2018 trial, characterized by extreme dry conditions, demonstrated the most negative impact on plant length and seed yield while simultaneously exhibiting the highest protein content. Conversely, the Kessenich trials (2018 and 2019) represented optimal growth conditions, showing the most substantial positive effects on seed yield, with Kessenich 2019 additionally displaying the most significant increase in plant length. Maturity phases (R8) varied across locations, ranging from 112 to 163 days, with Czech Republic sites consistently presenting the longest maturity period at 163 days.

Generalized heritability (Cullis et al., 2006) was highest for the R8 stage (92.1%), followed by protein content (84.7%) and plant length (84.5%). The lowest heritability 70.3% was observed for seed yield.

Correlations of BLUP´s are presented in Figure 2. The most significant correlation was found between the plant length and the date of R8 (maturity) phase observation. The second strongest correlation was between the plant length and the seed yield. The third strongest positive correlation was between the seed yield and the date of R8 (maturity) phase observation. A notable inverse relationship was observed between protein content and other analyzed traits. The results of the BLUPs for protein content and seed yield are presented in Figure 3. The ten most promising genotypes are marked in the graph by the red colour and the accession name.

Multivariate analysis

We conducted a multidimensional reduction of BLUPs for plant length, R8 stage, protein content, and seed yield using PCA

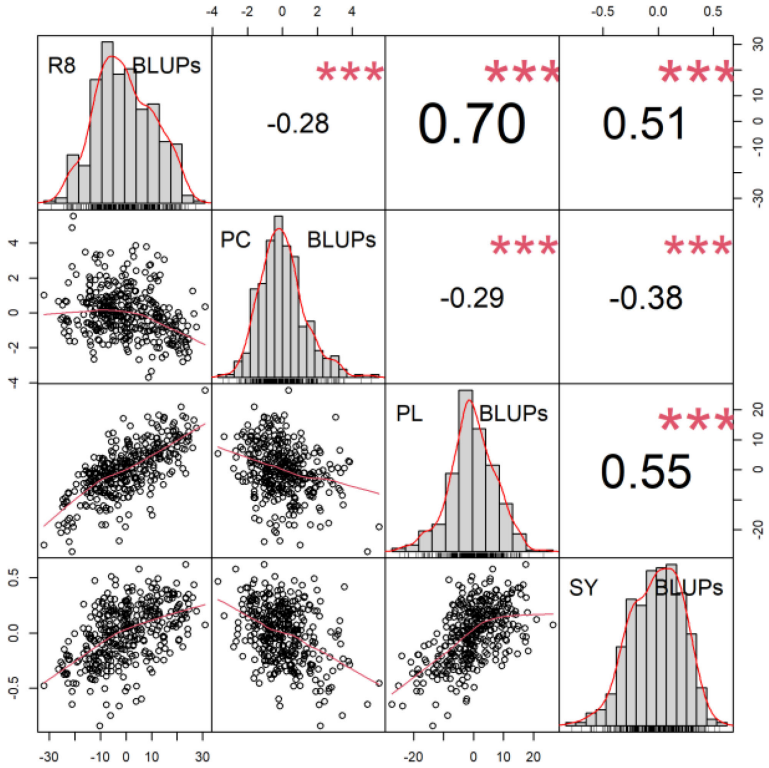


FIGURE 2
Correlogram for BLUPs of R8, protein content (PC), plant length (PL) and seed yield (SY). Scatter plots of relationship between parameters are described below diagonal. The correlation coefficient and the results of the cor.test are displayed above the diagonal. The stars mark the significance level of the test for association between paired samples using Pearson's product moment correlation coefficient α *** < 0.01.

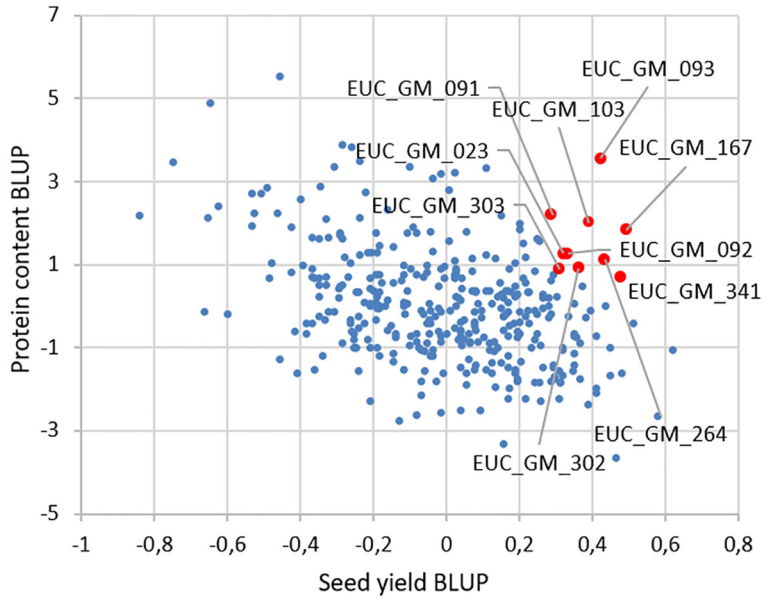


FIGURE 3
BLUP values for protein content and seed yield from all the soybean trials. By the red dots are marked the top ten soybean genotypes concerning the protein content and the seed yield.

analysis. The results were then compared with the assignment of five genetic groups (Saleem et al., 2021) and regional origin of accession to assess their distribution across the derived principal components. PCA plots (Figure 4) illustrate the spread and overlap of data points within the first two principal components, explaining 59.7% and 20.5% of the variance, respectively.

PCA plots, annotated with accession assignments, show an overview of the relationships between the profiles of key traits of accessions and their genetic groupings or origins. he distribution of accessions according to their phenotypic profiles of key traits partially reflects patterns in genetic groups and origins, but many of them are tightly overlapped. Genetic groups and origins are interconnected because the positions of some groups reflect their origins.

Redundancy Analysis (RDA) quantifies potential trait associations and genetic diversity within the soybean population. RDA assesses the relationship between the predicted breeding values (BLUPs) of plant length, protein content, seed yield and R8 stage traits and five genetic groups (G1 to G5). The analysis aimed to discern how genetic groupings might explain variation in trait BLUPs. The partitioning of variance in our RDA model revealed that the genetic groups accounted for 40% of the variance in trait BLUPs, according to the adjusted R-squared value. This indicates a considerable effect of genetic background on the expression of the traits we evaluated. Examining the eigenvalues and their contribution to the total variance, we found that the first RDA axis (RDA1) explained a substantial proportion of the variance at 35% with an eigenvalue of 1.4. The subsequent axes (RDA2 to RDA4) explained smaller fractions of the variance, specifically 8.94% for RDA2, 5.9% for RDA3, and 0.9% for RDA4. The analysis highlighted the significance of the association between trait BLUPs and genetic groups. The traits scores provided insight into how specific trait BLUPs respond to the genetic grouping. The strongest response on RDA1 shows trait R8 stage followed with seed yield and plant length. Conversely, the BLUP for trait protein content had the lowest association with RDA1, which represents 88% of genetic grouping influence on traits. In summary, the RDA has shown that genetic background, as represented by genetic groups G1 to G5, is a significant predictor of trait variation. This relationship is crucial for our understanding of the genetic architecture of these traits and could have implications for selective breeding programs. Graphic inspection of the results (Supplementary Figure 1) suggests that membership to genetic group G1, in particular, have a strong positive influence on the traits R8 stage, plant length and seed yield.

Genotype by environment interaction

The stability of genotypic performance across diverse environments was evaluated using the Finlay-Wilkinson regression method. This approach measures genotype responses to environmental variability. Trait results for the R8 stage, plant length, protein content, and seed yield of each genotype were modelled against an environmental score, which denotes the average trait values for all genotypes within each environment. The regression line's slope for each genotype acts as a stability indicator. Environmental scores were calculated for each trial, and Bayesian methods were employed to derive the slope values for each genotype. Our analysis disclosed a spectrum of stability responses across the genotypes for different traits (Figure 5a). Protein content displayed the most substantial stability. However, protein content also exhibited various responses to environmental changes, allowing us to distinguish between stable and responsive genotypes. In contrast, seed yield is exclusively associated with a positive response to an improving environment (Figure 5b) as a result of extreme environmental influences across trials, in which case, when the slope is far above 1, we have to compare genotypes relative to each other, with more stable seed yield genotypes being exhibited by individuals with lower slope values. Genotypes with a slope value near 1 were response stable. A slope of 1 indicates that the genotype's trait values increase proportionally with improvements in environmental conditions. Those genotypes are considered to have a stable response to environmental changes and

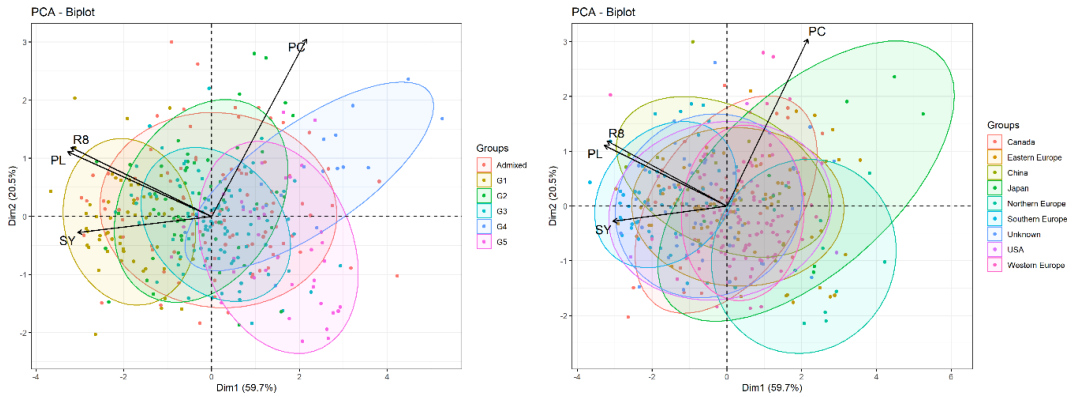


FIGURE 4
Comparative PCA plots of BLUPs of R8 stage, plant length, protein content and seed yield and distribution by genetic groups and the origin of accessions. The left plot categorizes observations into six groups (G1-G5, Admixed) and the right plot by geographic origin. Arrows show loadings of variables and ellipses show distribution of the accession origin.

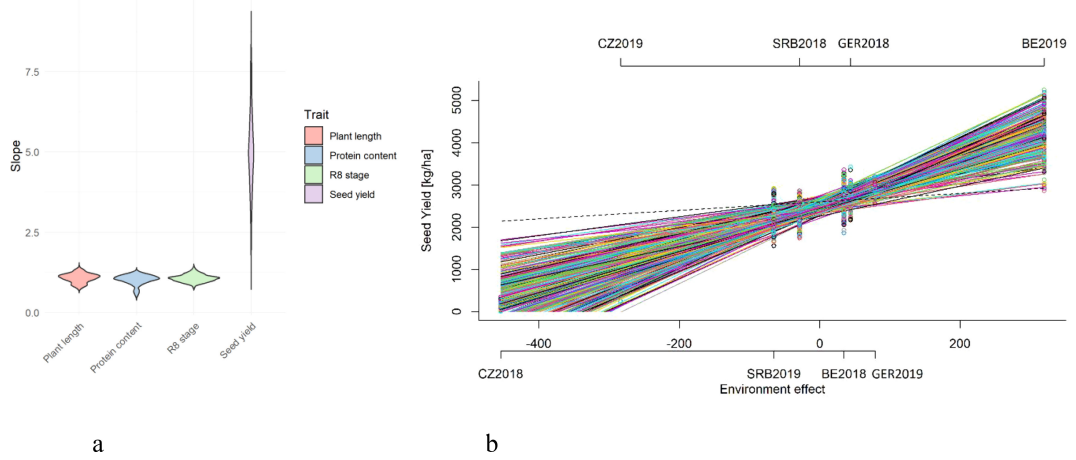


FIGURE 5
Distribution of Finlay-Wilkinson regression slope values for selected traits **(a)**. Genotypic response to estimated environmental variation in seed yield across eight distinct trials **(b)**. Each line represents the performance of genotype and the slope of line represents stability of an accession across different trial environments. Dashed line corresponds to a slope equal to b=1.

are generally stable across environments. Conversely, genotypes with the values of the slope significantly greater than 1 showed positive response. These genotypes are more responsive to environments. Such genotypes perform exceptionally well in good conditions but may suffer more in poorer conditions. Although all genotypes have a slope for seed yield greater than 1 (Figure 5b), for other traits (Figure 5a), genotypes with slope less than 1 indicate that the genotype is less responsive to changes in environmental conditions. These genotypes may perform assessed trait relatively well under poor conditions but do not show much higher values with better conditions. Genotypes with the lowest slope should be chosen because of their stability in adverse environments. The variation in response to the environment may indicate strategic genotype resource utilization.

The Finlay-Wilkinson regression slopes for all evaluated genotypes across the four traits are detailed in Supplementary Table 5, illustrating the range of genotypic reactions to environmental conditions. These results imply that breeders should carefully select genotypes based on the intended cultivation regime and location, which is consequential for choosing stable genotypes for parent selection. In summary, the Finlay-Wilkinson regression has yielded valuable insights into the stability and adaptability of genotypes within our trials, guiding selection criteria for future breeding programs that aim to develop cultivars tailored for broad adaptation or specific favourable environmental responses.

Discussion

Altogether, 360 accessions from different regions of the world were evaluated in phenotyping trials. We found a significant variability in the observed agronomic traits. We were interested not only in the variability of the selected important characters, but especially in their stability in different conditions. The stability of performance of individual genotypes grown in different environmental conditions is important for soybean growers and breeders (Rani and Kumar, 2022).

The geographical origin of individual genotypes played a very important role in our dataset. The most represented were accessions from Europe, North America and China. China itself is a very important region from the point of view of genetic resources, because it represents the domestication centre of soybean (Kim et al., 2012). It is not surprising that cultivated soybean prefers warm and humid climatic conditions during the vegetation period, typical for the region where soybean was firstly domesticated. The best soybean producing areas are located in warm areas with sufficient rainfall during soybean vegetation period (Gonçalves et al., 2021).

Selecting the best varieties should be considered with other existing limitations such as irrigation and field management, soil, growing season and its climatic conditions. Differences in crop management and agricultural practices also contribute to variations in yield and biomass (Araji et al., 2018). Soybean grows optimally at temperatures between 20 and 30°C (de Avila et al., 2013). High temperatures during flowering can cause reduction in seed number and seed weight. If the high temperatures are associated with a drought, the losses of grain production are even higher. On the other hand, the regions with temperatures below 10°C are not suitable for soybean cultivation (de Avila et al., 2013).

Concerning the climatic conditions, there are several critical moments, influencing the successful soybean cultivation: 1) sufficient rainfall during the emergence period, 2) absence of ground frosts during the emergence period, 3) combination of warm weather with occasional rainfall during the growing season and 4) warm and dry weather during the ripening period. Additional irrigation in case of drought significantly helps to increase the soybean yields. The importance of irrigation was mentioned by Karges et al. (2022). In their trial irrigation increased soybean yields by 41% on average. In the year with

sufficient precipitation, no additional irrigation is necessary (Karges et al., 2022). We observed differences between different maturity groups and growth types of soybeans in sensitivity to stress conditions. Early and short-growing cultivars are considered to be the most tolerant under mild and severe water stress (Araji et al., 2018).

Climatic conditions limit the expansion of new cultivation areas of soybean. The limiting factors are favourable temperatures and enough precipitation, mainly at the time of germination and flowering (Gawęda et al., 2020; Mandić et al., 2017). Soybean requires a sufficient number of warm days to mature. Kühling et al. (2018) reported a site in northwest Germany as sufficient for soybean cultivation. Rainfall during the later phase of the season, however, can have negative effects on the maturation process. Since soybean will be in the field until autumn in the northern parts of Europe, the risk of higher rainfall rates increases towards the harvest period. Soybean pods are fragile and repeated cycles of drying and wetting increase pod shattering and loss of seeds (Đorđević et al., 2021). Żarski et al. (2019) studied climatic risks for soybean cultivation in central Poland. Their analysis led to determination of the following unfavourable climatic conditions for soybean cultivation: shortening of the active growth period, a delay of the date on which the soil warms up to 8°C at a depth of 5 cm, occurrences of meteorological and agricultural droughts and of late spring ground frosts.

Soybean requires a soil temperature of 8 – 12°C for germination, with lower temperatures reducing plant density and yield (Yamaguchi et al., 2014). It is also sensitive to cold at flowering (Balko et al., 2014) with temperatures below 8°C associated with poor fertilization of ovules and subsequent blossom dropping (Yamaguchi et al., 2014). On the other hand, heat stress during the flowering stage can decrease yield significantly (Araji et al., 2018).

Future changes in temperature, rainfall, and CO_2 concentration will influence soybean growth and the final grain yield. Soybean will achieve an optimal threshold temperature in the future, leading to yield increases in the 2030s in temperate climate areas (Araji et al., 2018). Thanks to climate change, soybean cultivation is moving to suboptimal conditions, increasingly to colder conditions, areas in high latitudes or in high altitudes. Those areas are for example southern Scandinavia or northern Canada. On the other hand, the plants have to deal with even greater fluctuations in the weather during the growing season.

Identification of early maturity genotypes enables the soybean growers and breeders to provide suitable genotypes for marginal areas, especially those with colder climate. Based on the results of our experiments, we can recommend for cultivation in marginal areas those varieties that matured successfully in both experimental years at locations in Germany and Belgium. For successful soybean cultivation in marginal areas, it is important to choose not only early but also photoperiod-insensitive cultivars. These cultivars seem to be more strongly influenced by temperature, with higher temperatures resulting in earlier flowering (Kurasch et al., 2017).

The earlier the soybean genotype is, the more it is positively influenced by the length of the day. Late genotypes are not so much affected by the length of the day and grow well even under the condition of shorter days (Yang et al., 2019). Later breeding efforts for long-day suitability gradually added three more groups at the lower end (0000–00; Jia et al., 2014). Even though this day length × temperature range covered by the MGs already allows a broad adaptation across Europe, the early MGs – especially 0000 – currently produce significantly lower yields (Ortel et al., 2020) and are, therefore, not very attractive. Season length, as a combination of both factors, is, therefore, still a major constraint for growing soybean in northern latitudes (Yang et al., 2019).

Soybean yield and seed protein content are still the main breeding goals (Berschneider, 2016). We identified ten most promising soybean genotypes through the analysis of their seed yield and protein content using BLUPs. Genotypes were ranked based on their performance in producing high seed yield and high protein content, providing valuable insights into potential candidates for further breeding or cultivation efforts.

Identification of the most stable genotype even under variable environmental conditions is very important for the future cultivation and breeding of soybean genotypes. Current breeding efforts have concentrated on developing more drought-tolerant varieties as well as on weed-suppressing traits (higher and more branched plants with more leaves), which are especially important for organic production systems (Klaiss et al., 2020). Nendel et al. (2023) studied soybean potential in Europe under the conditions of climatic change. Their projections suggest a substantial increase in potential soybean production area and productivity in Central Europe, while southern European production would become increasingly dependent on supplementary irrigation. While wet conditions at harvest and incidental cold spells are the current key challenges for extending soybean production, the models and climate data analysis anticipate that drought and heat will become the dominant limitations in the future. Breeding for heat-tolerant and water-efficient genotypes is needed to further improve soybean adaptation to changing climatic conditions (Klaiss et al., 2020).

The observed variability in phenotypic traits across diverse environments and genotypes underscores the imperative need for integrating genotype-environment (G × E) interaction analyses within breeding programs. The results of Finlay Wilkinson regression recognize genotypes that have stable trait response across environments (slope near 1) as well as those that show good performance in marginal conditions (slope different from 1). The differences in the environmental conditions of the trials, as a result of the extreme dry years and the additional irrigation, were so different that the slope of the genotypes for the seed yield trait has values far above of 1. Even so, it is possible to select from relative comparisons between each other those genotypes that are stable in their response to extremes of conditions and those that are solely dependent on good conditions. This approach is crucial for elucidating the differential responses of genotypes to varying environmental conditions, thereby facilitating the development of crops that exhibit enhanced resilience, adaptability, and consistent high-yield performance across a spectrum of climatic and agronomic scenarios.

Our investigation revealed significant fluctuations in seed yield attributable to varying environmental conditions, highlighting the pronounced effect of climatic factors such as temperature and

precipitation on yield outcomes. The observed yield variability not only reflects the genetic diversity among soybean genotypes but also delineates a spectrum of resilience and yield potential under diverse conditions. Importantly, certain genotypes exhibited notable resistance and stable high-yield performance irrespective of environmental stressors. This phenomenon of yield fluctuation and the consequent G × E interactions, which highlight the differential genotype responses to environmental conditions, have been substantiated by studies from Alghamdi (2004), Krisnawati and Adie (2018), Oliveira et al. (2006), and Ngalamu et al. (2013). These studies underscore the significance of selecting genotypes that maintain stable performance across varied environments, contributing significantly to the predictability and reliability of crop yields. Additionally, Elmerich et al. (2023) identified specific eco-climatic factors influencing yield in early maturity soybeans, further emphasizing the importance of stability in performance.

The variation in plant length across different environments and genotypes underscores a complex interplay between soybean genetics and environmental cues, significantly impacting plant structure and consequently, plant density, harvesting efficiency. The importance of stable performance in terms of plant architecture, facilitated by the resilience of specific genotypes against environmental fluctuations, is critical for optimizing agricultural output. The contribution of environmental factors to G × E interaction, particularly in defining plant traits, was explored by Kang et al. (1989), who highlighted the critical role of weather variables and rainfall in influencing plant architecture. Furthermore, Yang et al. (2020) extended this understanding by identifying significant G × E effects on traits related to plant height and pinpointing specific quantitative trait loci that contribute to plant height across various environmental settings, while in soybean the results of mapping studies can be further specified to the level of causal mutations using synthetic phenotype association (Škrabišová et al., 2022).

Plant length is closely related to stem determination. Stem determinacy is quite variable among evaluated genotypes (Borra-Serrano et al., 2020). An important aspect for breeders and breeders is also the difference in sensitivity between determinate and indeterminate soybean genotypes. Indeterminate varieties are least affected by day length. Determinate varieties perform less well at high latitudes (Kato et al., 2019) including a large part of Northern and Western Europe (Schori et al., 2003).

Additionally, although there is observed variability in the timing of maturity (R8 stage), the response of different genotypes to environmental conditions demonstrates remarkable stability. This common stability facilitates the adaptation of genotypes to the photoperiod and temperature conditions characteristic of new geographic regions. The research conducted by Persa et al. (2022); Kantolic and Slafer (2005) and Elmerich et al. (2023) collectively highlights the significant influence of photoperiod on the development and yield of soybeans.

Regarding the nutritional quality of soybean seeds, specifically protein content, results indicate minimal response variability among genotypes across different environments, suggesting an inherent stability in nutritional quality despite environmental challenges. This stability offers promising prospects for enhancing the nutritional value of soybean crops through the identification and selection of genotypes that consistently exhibit high protein content alongside stable yield performance. Studies by Perić et al. (2021); Natarajan et al. (2016); Carrera et al. (2014) and Sethi et al. (2012) further confirm the feasibility of identifying genotypes with superior protein content, despite environmental variations.

At the end of this article, it is good to mention the main practical outputs of our study for breeders and soybean growers. On the basis of the above-described and then discussed results, it is possible to select, on the one hand, genotypes that have both a high yield and a high protein content, regardless of the growing conditions (Berschneider, 2016). These genotypes are interesting not only for breeders, as input materials for the breeding process, but also for soybean growers (Singh and Shivakumar, 2010). The second important group consists of mainly early genotypes, which are interesting from the point of view of cultivation and breeding for marginal conditions (Singh and Shivakumar, 2010). In our experiments, the riskiest factors from the point of view of soybean cultivation appears to be a period of long and intense drought, especially in the spring period, when soybeans go through a period of intensive growth and development (Klaiss et al., 2020; Saleem et al., 2021). For this reason, for the effective cultivation of soybeans, it is necessary to have an irrigation system available, without irrigation it can lead to a very significant reduction in yield (Karges et al., 2022).

Conclusions

Within the EUCLEG project, 360 accessions from different regions of the world were evaluated in phenotyping trials using the same methodology. The five most important soybean traits were selected for detailed analyses: plant emergence, plant length, protein content, seed yield and R8 stage. We found a significant variability in the observed agronomic traits. From the point of view of soybean breeding and cultivation, however, the most important characteristics are seed yield and protein content. The ten most promising genotypes from the point of view of protein content and yield were selected and could be used in breeding programmes. Our results and observed variability in phenotypic traits across diverse environments and genotypes underscore the need for integrating genotype-environment interaction analyses within breeding programs.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

TV: Conceptualization, Data curation, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. OT: Data curation,

Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. HR: Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – review & editing. CB: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – review & editing. VD: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – review & editing. PR: Data curation, Formal analysis, Investigation, Visualization, Writing – review & editing. MD: Resources, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project has received funding from the European Union's Horizon 2020 Program for Research & Innovation under grant agreement no. 727312 (project: "EUCLEG – Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency").

Acknowledgments

We would especially like to thank the staff of the four partners where the experiments were carried out - Agricultural Research Ltd., Institute for Field and Vegetable Crops, Julius Kühn-Institute and Agro Seed Research bv. Of course, we would also like to thank the entire team of the EUCLEG project, who, thanks to their very good and close cooperation, helped to obtain high-quality data, process it and prepare the outputs of this project. Many thanks to all those who provided seeds of different soybean genotypes that were evaluated in the framework of the EUCLEG project.

Conflict of interest

Authors TV, OT, and MD were employed by the company Agricultural Research, Ltd. Author HR was employed by the company Agro Seed Research bv.

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 AE declared a past co-authorship with the author MP to the handling editor.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1422162/full#supplementary-material

References

Alghamdi, S. S. (2004). Yield stability of some soybean genotypes across diverse environments. *Pak. J. Biol. Sci* 7, 2109–2114. doi: 10.3923/PJBS.2004.2109.2114

Araji, H. A., Wayayok, A., Bavani, A. R. M., Amiri, E., Abdullah, A. F., Daneshian, J., et al. (2018). Impacts of climate change on soybean production under different treatments of field experiments considering the uncertainty of general circulation models. *Agric. Water Manage* 205, 63–71. doi: 10.1016/j.agwat.2018.04.023

Balko, C., Hahn, V., and Ordon, F. (2014). Chilling tolerance in soybeans (*Glycine max*) - a prerequisite for soybean cultivation in Germany. *J. fur Kult* 66, 378–388. doi: 10.5073/JfK.2014.11.02

Barnes, S. (2010). The biochemistry, chemistry and physiology of the isoflavones in soybeans and their food products. *Lymphat. Res. Biol* 8, 89–98. doi: 10.1089/lrb.2009.0030

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Berschneider, J. (2016). *Chances and limitations of european soybean production: market potential analysis [master's thesis]* (Stuttgart (Germany: Universität Hohenheim).

Borra-Serrano, I., De Swaef, T., Quataert, P., Aper, J., Saleem, A., Saeys, W., et al. (2020). Closing the phenotyping gap: high resolution UAV time series for soybean growth analysis provides objective data from field trials. *Remote Sens* 12, 1644. doi: 10.3390/rs12101644

Carrera, C., Dardanelli, J., and Soldini, D. O. (2014). Genotypic and environmental variation in seed nutraceutical and industrial composition of non-transgenic soybean (*Glycine max*) genotypes. *Crop Pasture Sci* 65, 1311–1322. doi: 10.1071/CP14114

Ceglar, A., Zampieri, M., Toreti, A., and Dentener, F. (2019). Observed northward migration of agro-climate zones in Europe will further accelerate under climate change. *Earth's Future* 7, 1088–1101. doi: 10.1029/2019EF001178

Chen, L. R., Ko, N. Y., and Chen, K. H. (2019). Isoflavone supplements for menopausal women: a systematic review. *Nutrients* 11, 2649. doi: 10.3390/nu11112649

Cicero, A. F. G., Fogacci, F., and Colletti, A. (2017). Potential role of bioactive peptides in prevention and treatment of chronic diseases: a narrative review. *Br. J. Pharmacol* 174, 1378–1394. doi: 10.1111/bph.13608

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *J. Agr. Biol. Envir. St* 11, 381–393. doi: 10.1198/108571106x154443

de Avila, A. M. H., Farias, J. B., Pinto, H. S., and Pilau, F. G. (2013). "Climatic restrictions for maximizing soybean yields," in *A comprehensive survey of international soybean research - genetics, physiology, agronomy and nitrogen relationships* (IntechOpen Limited, London), 367–375.

Đorđević, V., Balešević-Tubić, S., Miladinović, J., and Vasiljević, M. (2021).Harvesting soybean. In: *Legumes translated practice note 23. Institute of field and vegetable crops novi sad*. Available online at: https://www.legumestranslated.eu/ (Accessed October 3, 2023).

Ebert, A. W. (2020). The role of vegetable genetic resources in nutrition security and vegetable breeding. *Plants* 9, 736. doi: 10.3390/plants9060736

Elmerich, C., Boulch, G., Faucon, M., Lakhal, L., and Lange, B. (2023). Identification of eco-climatic factors driving yields and genotype by environment interactions for yield in early maturity soybean using crop simulation. *Agronomy* 13, 322. doi: 10.3390/agronomy13020322

EUROStat (2021). Available online at: https://ec.europa.eu/eurostat/de/. (Accessed May 25, 2023)

Fan, Y., Wang, M., Li, Z., Jiang, H., Shi, J., Shi, X., et al. (2022). Intake of soy, soy isoflavones and soy protein and risk of cancer incidence and mortality. *Front. Nutr* 9. doi: 10.3389/fnut.2022.847421

FAOStat (2021). Available online at: http://www.fao.org/faostat/en/data. (Accessed May 25, 2023)

FAOSTAT (2023). *Food and agriculture organization of the united nations* (FAOSTAT). Available online at: www.faostat.org (Accessed May 25, 2023).

Gawęda, D., Nowak, A., Haliniarz, M., and Woźniak, A. (2020). Yield and economic effectiveness of soybean grown under different cropping systems. *Int. J. Plant Prod* 14, 475–485. doi: 10.1007/s42106-020-00098-1

Gonçalves, S. L., Farias, J. R. B., and Sibaldelli, R. N. R. (2021). Soybean production and yield in the context of global climatic changes. *CABI Rev* 16, 11. doi: 10.1079/PAVSNNR202116011

Gorissen, S. H., and Witard, O. C. (2018). Characterising the muscle anabolic potential of dairy, meat and plant-based protein sources in older adults. *Proc. Nutr. Soc* 77, 20–31. doi: 10.1017/S002966511700194X

Guilpart, N., Iizumi, T., and Makowski, D. (2022). Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. *Nat. Food* 3, 255–265. doi: 10.1038/s43016-022-00481-3

Guo, B., Sun, L., Jiang, S., Ren, H., Sun, R., Wei, Z., et al. (2022). Soybean genetic resources contributing to sustainable protein production. *Theor. Appl. Genet* 135, 4095–4121. doi: 10.1007/s00122-022-04222-9

Halewood, M., Chiurugwi, T., Sackville-Hamilton, R., Kurtz, B., Marden, E., Welch, E., et al. (2018). Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution. *New Phytol* 217, 1407–1419. doi: 10.1111/nph.14993

Hartman, G. L., West, E. D., and Herman, T. K. (2011). Crops that feed the World 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests. *Food Secur* 3, 5–17. doi: 10.1007/s12571-010-0108-x

He, F.-J., and Chen, J.-Q. (2013). Consumption of soybean, soy foods, soy isoflavones and breast cancer incidence: Differences between Chinese women and women in Western countries and possible mechanisms. *Food Sci. Hum. Wellness* 2, 146–161. doi: 10.1016/j.fshw.2013.08.002

Hymowitz, T., and Harlan, J. R. (1983). Introduction of soybean to north america by samuel bowen in 1765. *Economic Bot* 37, 371–379. Available at: http://www.jstor.org/stable/4254529. (Accessed May 15, 2023)

Iriondo, J. M., Maxted, N., and Dulloo, M. E. (2008). *Conserving plant genetic diversity in protected areas: Population management of crop wild relatives* (Oxford: CAB International, Wallingsford), 212.

Jayachandran, M., and Xu, B. (2019). An insight into the health benefits of fermented soy products. *Food Chem* 271, 362–371. doi: 10.1016/j.foodchem.2018.07.158

Jeong, S. C., Moon, J. K., Park, S. K., Kim, M. S., Lee, K., Lee, S. R., et al. (2019). Genetic diversity patterns and domestication origin of soybean. *Theor. Appl. Genet* 132, 1179–1193. doi: 10.1007/s00122-018-3271-7

Jia, H., Jiang, B., Wu, C., Lu, W., Hou, W., Sun, S., et al. (2014). Maturity group classification and maturity locus genotyping of early-maturing soybean varieties from high-latitude cold regions. *PloS One* 9, e94139. doi: 10.1371/journal.pone.0094139

Kang, M., Harville, B. G., and Gorman, D. (1989). Contribution of weather variables to genotype × environment interaction in soybean. *Field Crops Res* 21, 297–300. doi: 10.1016/0378-4290(89)90011-7

Kantolic, A. G., and Slafer, G. A. (2005). Reproductive development and yield components in indeterminate soybean as affected by post-flowering photoperiod. *Field Crops Res* 93, 212–222. doi: 10.1016/j.fcr.2004.10.001

Kao, T.-H., and Chen, B.-H. (2006). Functional components in soybean cake and their effects on antioxidant activity. *J. Agric. Food Chem* 54, 7544–7555. doi: 10.1021/jf061586x

Karges, K., Bellingrath-Kimura, S. D., Watson, C. A., Stoddard, F. L., Halwani, M., and Reckling, M. (2022). Agro-economic prospects for expanding soybean production beyond its current northerly limit in Europe. *Eur. J. Agron* 133, 126415. doi: 10.1016/j.eja.2021.126415

Kato, S., Sayama, T., Taguchi-Shiobara, F., Kikuchi, A., Ishimoto, M., and Cober, E. (2019). Effect of change from a determinate to a semi-determinate growth habit on the yield and lodging resistance of soybeans in the northeast region of Japan. *Breed. Sci* 69, 151–159. doi: 10.1270/jsbbs.18112

Kim, M. Y., Van, K., Kang, Y. J., Kim, K. H., and Lee, S.-H. (2012). Tracing soybean domestication history: from nucleotide to genome. *Breed. Sci* 61, 445–452. doi: 10.1270/jsbbs.61.445

Klaiss, M., Schmid, N., Betrix, C.-A., Baux, A., Charles, R., and Messmer, M. M. (2020). Organic soybean production in Switzerland. *OCL* 27, 64. doi: 10.1051/ocl/2020059

Krisnawati, A., and Adie, M. M. (2018). Yield stability of soybean genotypes in tropical environments based on genotype and genotype-by-environment biplot. *J. Agron. Indonesia* 46, 231–239. doi: 10.24831/jai.v46i3.18333

Kühling, I., Hüsing, B., Bome, N., and Trautz, D. (2018). Soybeans in high latitudes: effects of *Bradyrhizobium* inoculation in Northwest Germany and southern West Siberia. *Org. Agric* 8, 159–171. doi: 10.1007/s13165-017-0181-y

Kurasch, A., Hahn, V., Leiser, W., Vollmann, J., Schori, A., Betrix, C.-A., et al. (2017). Identification of mega-environments in Europe and effect of allelic variation at maturity E loci on adaptation of European soybean. *Plant Cell Environ* 40, 765–778. doi: 10.1111/pce.12896

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Software* 82, 1–26. doi: 10.18637/jss.v082.i13

Lenth, R. (2024). emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version 1.10.5*. Available online at: https://rvlenth.github.io/emmeans/. (Accessed March 27, 2024)

Lian, L., and de los Campos, G. (2016). FW: an R package for finlay–wilkinson regression that incorporates genomic/pedigree information and covariance structures between environments. *G3 Genes Genom. Genet* 6, 589–597. doi: 10.1534/g3.115.026328

Liu, J., Chang, S. K. C., and Wiesenborn, D. (2005). Antioxidant properties of soybean isoflavone extract and tofu *in vitro* and in *vivo*. *J. Agric. Food Chem* 53, 2333–2340. doi: 10.1021/jf048552e

Mandić, V., Bijelić, Z., Krnjaja, V. S., Simić, A., Ružić-Muslić, D., Dragičević, V., et al. (2017). The rainfall use efficiency and soybean grain yield under rainfed conditions in Vojvodina. *Bio. Anim. Husb* 33, 475–486. doi: 10.2298/BAH1704475M

Muluneh, M. G. (2021). Impact of climate change on biodiversity and food security: a global perspective–a review article. *Agric. Food Secur* 10, 36. doi: 10.1186/s40066-021-00318-5

Munger, P., Bleiholder, H., Hack, H., Hess, M., Stauss, R., Van den Boom, T., et al. (1997). Phenological growth stages of the soybean plant (*Glycine max* L. Merr.): codification and description according to the BBCH scale. *J. Agron. Crop Sci* 179, 209–217. doi: 10.1111/j.1439-037X.1997.tb00519.x

Natarajan, S., Khan, F., Song, Q., Lakshman, S., Cregan, P., Scott, R., et al. (2016). Characterization of soybean storage and allergen proteins affected by environmental and genetic factors. *J. Agric. Food Chem* 64, 1433–1445. doi: 10.1021/acs.jafc.5b05172

Nendel, C., Reckling, M., Schulz, S., Berg-Mohnicke, M., Constantin, J., Debaeke, P., et al. (2023). Future area expansion outweighs increasing drought risk for soybean in Europe. *Glob. Change Biol* 29, 1340–1358. doi: 10.1111/gcb.16562

Ngalamu, T., Ashraf, M., and Meseka, S. (2013). Soybean (*Glycine max* L) genotype and environment interaction effect on yield and other related traits. *Am. J. Exp. Agric* 3, 977–987. doi: 10.9734/AJEA/2013/5069

Ohara, M., and Shimamoto, Y. (2002). Importance of genetic characterization and conservation of plant genetic resources: The breeding system and genetic diversity of wild soybean (*Glycine soja*). *Plant Species Biol* 17, 51–58. doi: 10.1046/j.1442-1984.2002.00073.x

Oliveira, A., Duarte, J., Chaves, L., and Couto, M. A. (2006). Environmental and genotypic factors associated with genotype by environment interactions in soybean. *Crop Breed. Appl. Biot* 6, 79–86. doi: 10.12702/1984-7033.V06N01A11

Ortel, C. C., Roberts, T. L., Hoegenauer, K. A., Purcell, L. C., Slaton, N. A., and Gbur, E. E. (2020). Soybean maturity group and planting date influence grain yield and nitrogen dynamics. *Agrosyst. Geosci. Environ* 3, e20077. doi: 10.1002/agg2.20077

Perić, V., Srebrić, M., Nikolić, A., Ristić, D., Mladenović-Drinić, S., and Dumanović, Z. (2021). Stability of yield and seed composition in early maturing soybean genotypes assessed by AMMI analysis. *Genetika* 53, 323–338. doi: 10.2298/GENSR2101323P

Persa, R., Graef, G., Specht, J., Rios, E., Messina, C., and Jarquín, D. (2022). Enhancing genomic prediction models for forecasting days to maturity in soybean genotypes using site-specific and cumulative photoperiod data. *Agriculture*. 12 (4). doi: 10.3390/agriculture12040545

Qiu, L. J., and Chang, R. Z. (2010). "The origin and history of soybean," in *The soybean: botany, production and uses. 1–23*. Ed. S. C. Guriqbal (CABI, Wallingford).

Rani, A., and Kumar, V. (2022). "Soybean breeding," in *Fundamentals of field crop breeding* (Springer Nature Singapore, Singapore), 907–944.

Saleem, A., Muylle, H., Aper, J., Ruttink, T., Wang, J., Yu, D., et al. (2021). A genome-wide genetic diversity scan reveals multiple signatures of selection in a european soybean collection compared to chinese collections of wild and cultivated soybean accessions. *Front. Plant Sci* 12. doi: 10.3389/fpls.2021.631767

Schori, A., Charles, R., and Peter, D. (2003). Soja: sélection, agronomie et production en Suisse. *Revue suisse d'Agriculture*. 35 (2), 69–67. Available at: https://ira.agroscope.ch/fr-CH/publication/3354 (Accessed May 20, 2024).

Sethi, J., Tyagi, S., and Sameer, S. (2012). Assessment of Genotype x Environment Interactions and Stability Analysis in Soybean [*Glycine max* (L.) Merrill. *SKUAST J. Res*. 13, 34–39. Available at: https://www.indianjournals.com/ijor.aspx?target=ijor:skuastjr&volume=14&issue=1and2&article=005&type=pdf (Accessed May 20, 2024).

Sharma, S., Kaur, M., Goyal, R., and Gill, B. S. (2014). Physical characteristics and nutritional composition of some new soybean (*Glycine max* (L.) Merrill) genotypes. *J. Food Sci. Tech* 51, 551–557. doi: 10.1007/s13197-011-0517-7

Singh, G., and Shivakumar, B. G. (2010). "The role of soybean in agriculture," in *The soybean: Botany, production and uses* (CAB International, Oxfordshire, UK), 24–47.

Škrabišová, M., Dietz, N., Zeng, S., Chan, Y. O., Wang, J., Liu, Y., et al. (2022). A novel Synthetic phenotype association study approach reveals the landscape of association for genomic variants and phenotypes. *J. Adv. Res* 42, 117–133. doi: 10.1016/j.jare.2022.04.004

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep* 6, 20728. doi: 10.1038/srep20728

Yamaguchi, N., Yamazaki, H., Ohnishi, S., Suzuki, C., Hagihara, S., Miyoshi, T., et al. (2014). Method for selection of soybeans tolerant to seed cracking under chilling temperatures. *Breed. Sci* 64, 103–108. doi: 10.1270/jsbbs.64.103

Yang, Q., Lin, G., Lv, H., Wang, C., Yang, Y., and Liao, H. (2020). Environmental and genetic regulation of plant height in soybean. *BMC Plant Biol*, 21 (1). doi: 10.1186/s12870-021-02836-7

Yang, W. Y., Wu, T. T., Zhang, X. Y., Song, W. W., Xu, C. L., Sun, S., et al. (2019). Critical photoperiod measurement of soybean genotypes in different maturity groups. *Crop Sci* 59, 2055–2061. doi: 10.2135/cropsci2019.03.0170

Żarski, J., Kuśmierek-Tomaszewska, R., Dudek, S., Kropkowski, M., and Kledzik, R. (2019). Identifying climatic risk to soybean cultivation in the transitional type of moderate climate in Central Poland. *J. Cent. Eur. Agric* 20, 143–156. doi: 10.5513/JCEA01/20.1.2210

Zhu, Z., Chen, S., Wu, X., Xing, C., and Yuan, J. (2018). Determination of soybean routine quality parameters using near-infrared spectroscopy. *Food Sci. Nutr* 6, 1109–1118. doi: 10.1002/fsn3.652

Zuo, X., Zhao, R., Wu, M., Wan, Q., and Li, T. (2023). Soy consumption and the risk of type 2 diabetes and cardiovascular diseases: A systematic review and meta-analysis. *Nutrients* 15, 1358. doi: 10.3390/nu15061358

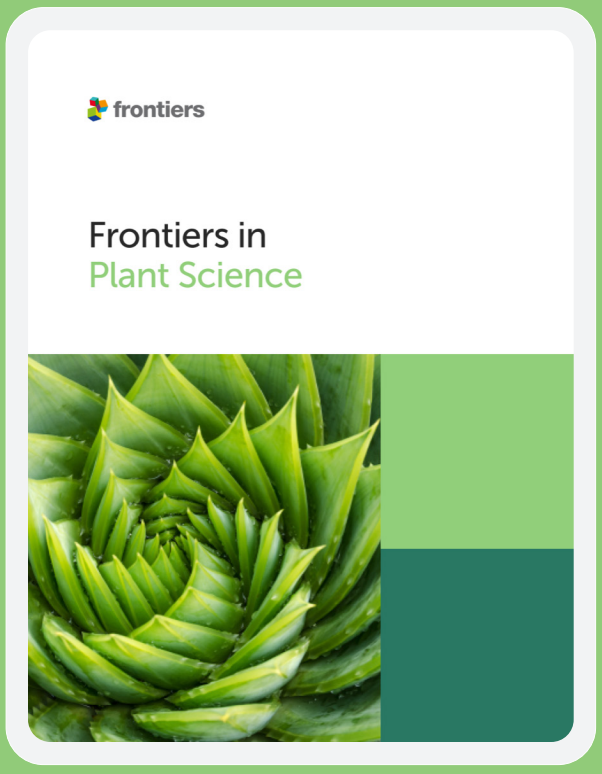
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

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