

Advances in conservation and utilization of plant genetic resources

Edited by

Svein Øivind Solberg, Maarten Van Zonneveld
and Axel Diederichsen

Published in

Frontiers in Plant Science



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 978-2-8325-5334-3
DOI 10.3389/978-2-8325-5334-3

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

Advances in conservation and utilization of plant genetic resources

Topic editors

Svein Øivind Solberg — Inland Norway University of Applied Sciences, Norway

Maarten Van Zonneveld — World Vegetable Center, Taiwan

Axel Diederichsen — Agriculture and Agri-Food Canada (AAFC), Canada

Citation

Solberg, S. Ø., Van Zonneveld, M., Diederichsen, A., eds. (2024). *Advances in conservation and utilization of plant genetic resources*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5334-3

Table of contents

- 06 Editorial: Advances in conservation and utilization of plant genetic resources
Svein Øivind Solberg, Maarten van Zonneveld and Axel Diederichsen
- 10 *Brassica* biodiversity conservation: prevailing constraints and future avenues for sustainable distribution of plant genetic resources
Parthiban Subramanian, Seong-Hoon Kim and Bum-Soo Hahn
- 23 Development of cassava core collections based on morphological and agronomic traits and SNPS markers
Caroline Cardoso dos Santos, Luciano Rogerio Braatz de Andrade, Cátia Dias do Carmo and Eder Jorge de Oliveira
- 39 Analysis of gaps in rapeseed (*Brassica napus* L.) collections in European genebanks
Stephan Weise, Roel Hoekstra, Kim Jana Kutschan, Markus Oppermann, Rob van Treuren and Ulrike Lohwasser
- 55 Genomic characterization and gene bank curation of *Aegilops*: the wild relatives of wheat
Laxman Adhikari, John Raupp, Shuangye Wu, Dal-Hoe Koo, Bernd Friebe and Jesse Poland
- 70 Twenty-four years lucerne (*Medicago sativa* L.) breeder seed production in India: a retrospective study
Subhash Chand, Ajoy Kumar Roy, Tejveer Singh, Rajiv Kumar Agrawal, Vijay Kumar Yadav, Sanjay Kumar, Devendra Ram Malaviya, Amaresh Chandra and Devendra Kumar Yadava
- 81 Complete mitochondrial genome of the endangered *Prunus pedunculata* (Prunoideae, Rosaceae) in China: characterization and phylogenetic analysis
Qian Liu, Zinian Wu, Chunyu Tian, Yanting Yang, Lemeng Liu, Yumei Feng and Zhiyong Li
- 97 DNA barcoding and comparative RNA-Seq analysis provide new insights into leaf formation using a novel resource of high-yielding *Epimedium koreanum*
Jiaxin Yang, Siqing Fan, Min Guo, Zhaoqi Xie, Qiqing Cheng, Puxin Gao and Chunsong Cheng
- 111 New insights in the Spanish gene pool of olive (*Olea europaea* L.) preserved *ex situ* and *in situ* based on high-throughput molecular markers
Francisco Jesús Gómez-Gálvez, Antònia Ninot, Juan Cano Rodríguez, Sergio Paz Compañ, Javier Ugarte Andrevia, Javier Alfonso García Rubio, Isis Pinilla Aragón, Javier Viñuales-Andreu, José Casanova-Gascón, Zlatko Šatović, Ignacio Jesús Lorite, Raúl De la Rosa-Navarro and Angjelina Belaj

- 123 **Genomic prediction reveals unexplored variation in grain protein and lysine content across a vast winter wheat genebank collection**
Marcel O. Berkner, Stephan Weise, Jochen C. Reif and Albert W. Schulthess
- 135 **Identifying genetically redundant accessions in the world's largest cassava collection**
Monica Carvajal-Yepes, Jessica A. Ospina, Ericson Aranzales, Monica Velez-Tobon, Miguel Correa Abondano, Norma Constanza Manrique-Carpintero and Peter Wenzl
- 154 **Towards a practical threat assessment methodology for crop landraces**
Maria João Almeida, Ana Maria Barata, Stef De Haan, Bal Krishna Joshi, Joana Magos Brehm, Mariana Yazbek and Nigel Maxted
- 167 **An international breeding project using a wild potato relative *Solanum commersonii* resulted in two new frost-tolerant native potato cultivars for the Andes and the Altiplano**
Jesus H. Arcos-Pineda, Alfonso H. del Rio, John B. Bamberg, Sandra E. Vega-Semorile, Jiwan P. Palta, Alberto Salas, Rene Gomez, William Roca and David Ellis
- 178 **Low-coverage whole genome sequencing of diverse *Dioscorea bulbifera* accessions for plastome resource development, polymorphic nuclear SSR identification, and phylogenetic analyses**
Ruisen Lu, Ke Hu, Xiaoqin Sun and Min Chen
- 191 **Phylogeographic analysis reveals extensive genetic variation of native grass *Elymus nutans* (Poaceae) on the Qinghai-Tibetan plateau**
Jin Li, Xinda Li, Changbing Zhang, Qingping Zhou and Shiyong Chen
- 202 **Unlocking the genetic diversity and population structure of the newly introduced two-row spring European Heritage Barley collection (ExHIBiT)**
Villő Bernád, Nadia Al-Tamimi, Patrick Langan, Gary Gillespie, Timothy Dempsey, Joey Henchy, Mary Harty, Luke Ramsay, Kelly Houston, Malcolm Macaulay, Paul D. Shaw, Sebastian Raubach, Kevin P. McDonnell, Joanne Russell, Robbie Waugh, Mortaza Khodaeiaminjan and Sónia Negrão
- 219 **Genetic diversity, population structure, and taxonomic confirmation in annual medic (*Medicago* spp.) collections from Crimea, Ukraine**
Dongyan Zhao, Manoj Sapkota, Meng Lin, Craig Beil, Moira Sheehan, Stephanie Greene and Brian M. Irish
- 236 **Advancing utilization of diverse global carrot (*Daucus carota* L.) germplasm with flowering habit trait ontology**
Jenyne Loarca, Michael Liou, Julie C. Dawson and Philipp W. Simon

- 254 **Evaluation of shoot-growth variation in diverse carrot (*Daucus carota* L.) germplasm for genetic improvement of stand establishment**
Jenyne Loarca, Michael Liou, Julie C. Dawson and Philipp W. Simon
- 271 **Differences in *Albizia odoratissima* genetic diversity between Hainan Island and mainland populations in China**
Qi An, Yuanheng Feng, Zhangqi Yang, La Hu, Dongshan Wu and Guifang Gong
- 283 **Sampling strategies for genotyping common bean (*Phaseolus vulgaris* L.) Genebank accessions with DArTseq: a comparison of single plants, multiple plants, and DNA pools**
Miguel Correa Abondano, Jessica Alejandra Ospina, Peter Wenzl and Monica Carvajal-Yepes



OPEN ACCESS

EDITED AND REVIEWED BY
Diego Rubiales,
Spanish National Research Council (CSIC),
Spain

*CORRESPONDENCE
Svein Øivind Solberg
✉ svein.solberg@inn.no

RECEIVED 22 July 2024
ACCEPTED 29 July 2024
PUBLISHED 07 August 2024

CITATION
Solberg SØ, van Zonneveld M and
Diederichsen A (2024) Editorial:
Advances in conservation and utilization of
plant genetic resources.
Front. Plant Sci. 15:1468904.
doi: 10.3389/fpls.2024.1468904

COPYRIGHT
© 2024 Solberg, van Zonneveld and
Diederichsen. 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.

Editorial: Advances in conservation and utilization of plant genetic resources

Svein Øivind Solberg^{1*}, Maarten van Zonneveld²
and Axel Diederichsen³

¹Department of Agriculture, Inland Norway University of Applied Sciences, Elverum, Norway, ²World Vegetable Center, Shanhua, Taiwan, ³Plant Gene Resources of Canada, Agriculture and Agri-Food Canada (AAFC), Saskatoon, SK, Canada

KEYWORDS

conservation, crop wild relatives, diversity, evaluation, food security, genebank, plant genetic resources, seed bank

Editorial on the Research Topic

[Advances in conservation and utilization of plant genetic resources](#)

Background

We made a broad call for perspective papers, systematic reviews, meta-analyses, or traditional research papers on topics related to conservation and use of plant genetic resources. We invited to present advances in characterization and evaluation, strategies to improve gene bank operations and collaboration, new tools for managing and sharing information, or novel knowledge of conservation gaps. We especially encouraged contributions on underutilized crops and crops wild relatives. We received 58 positive responses, whereof 20 finalized their submissions and passed the review process with a total of 125 authors included in the accepted papers. Our motivation was to assemble contributions underlining that genetic resources are essential for crop improvement, which is achieved via plant breeding and release of new varieties that farmers can access and use. We see that future crops need to produce high and stable yields but also be of high nutritional quality. They must adapt to shifting climates and support a sustainable agricultural or horticultural production avoiding negative environmental impacts. In addition, we need to contribute to efforts to avoid the loss of biodiversity, including the genetic resources of crop plants and crop wild relatives, which are in our focus. We are aware that there is a large body of research and literature on this, because plant breeders in the late 19th century already imitated research and conservation activities which serve the same purpose and they recognized the loss of biodiversity, albeit not using this term. However, the global biodiversity crisis has become much bigger and globally recognized and any additional step, any additional insight addressing this topic is worth to be shared. At the same time more and more genomic tools and other technologies are available to add to our understanding of diversity and can support conservation and target promising germplasm for further research and breeding. We hope this special edition will be a small additional step in a positive direction, part of the evolutionary process of coping with the grand challenges we all face.

Conservation

Six papers focused on conservation, which included conservation methods and priorities as well as core collections. Of these, one is a review paper and five are research articles.

Subramanian et al.'s review on *Brassica* biodiversity conservation discusses issues related to distribution of accessions, conservation methods, technical hurdles and future avenues for research. Despite that there are more than 80,000 *Brassica* accessions conserved across 81 countries the authors see that the diversity of the conserved taxa is limited in most countries, which they see may lead to biodiversity loss in the longer run. In addition, they see practical challenges as taxonomic issues in the conservation system.

The research paper from Weise et al. focuses on the rapeseed (*Brassica napus* L.) gene pool that is conserved in European genebanks. They highlight that most of these species are underrepresented in current collections and that many of the natural distribution areas are not covered as it stands now. By using niche modelling they further illustrate how climate change may affect the species' distribution ranges. The authors suggest to further develop the conservation strategies for the rapeseed gene pool and propose a list of priority species that should be targeted for collecting missions.

Carvajal-Yepes et al. made a work to identify genetically redundant accessions in the world's largest cassava (*Manihot esculenta*) collection with 5,302 accessions maintained at the International Center for Tropical Agriculture (CIAT). An empirical distance threshold methodology was applied with two types of molecular markers (SNP and SilicoDArT). The results showed 2,776 (SNP) and 2,785 (SilicoDArT) accessions were part of accession clusters. By comparing passport and historical characterization data clusters of genetically redundant accessions the authors provided a roadmap for genebank curators to assess redundancy within collections and/or identifying subsets of genetically distinct accessions.

Almeida et al. presented a methodology for landrace threat assessment which can assist in setting priorities for conservation. The methodology of this work was in line with the IUCN Red Listing judgement for wild species and involves the collation of time series information on population range and trends. Unlike for wild species the information used here involved farmers and market actors. The authors conclude that the archived information can be compared to a standardized set of threat criteria with a set threshold level and the methodology can be applied to any crop and geographical scale.

Dos Santos et al. developed a core collection of the cassava (*Manihot esculenta* Crantz) germplasm bank in Brazil based on morphological traits, agronomic traits, and genotypic data from 20k single-nucleotide polymorphisms (SNPs). Out of the 1,486 accessions in the germplasm bank a consolidated core with 204 accessions was suggested, which is 14% of the complete collection. This core collection showed less genetic variation but retained over 97% of the allelic richness compared to the complete collection. The authors see that the core approach provides a robust and representative resource for further research and breeding in cassava.

Bernad et al. examined the genetic diversity and population structure of the newly introduced two-row spring European Heritage Barley collection. This is a small collection which consists of 363 spring-barley accessions mainly from Northern Europe and include both landraces (~14%), old cultivars (~18%), elite cultivars (~67%) and accessions with unknown breeding history (~1%). The authors used 26,585 informative SNPs based on 50k iSelect array data and the results showed that the collection could be subdivided into three main clusters. The clusters were primarily based on the accession's year of release. Furthermore, power analysis identified a core with 230 genotypically and phenotypically diverse accessions, which shows that the collection represents a high diversity and can be a resource for research and breeding.

Characterization and evaluation

Twelve research articles from recent characterization and evaluation projects are presented, which included a range of crops, methods and aims and showing how molecular markers and genomic tools can help to identify promising germplasm for further phenotyping and evaluation.

Gomez-Galvez et al. examined more than 500 olive (*Olea europaea* L.) genotypes from different regions in Spain based on 96 EST-SNP markers and identified 173 new genotypes. Based on this work the number of distinct Spanish genotypes documented in the World Olive Germplasm Bank of IFAPA, Córdoba increased from 269 to 427 accessions. In addition, a new diversity hot spot was identified in the northern regions of La Rioja and Aragon. According to the authors this adds to the great diversity already described in Spanish olive germplasm. The authors highlight the risk of genetic erosion given the expansion of modern olive cultivation with only a few cultivars and argue that conserving a broad range of genotypes will be crucial to meet the future challenges of olive cultivation. To further enhance the conservation and use of these accessions, the World Olive Germplasm Bank of Córdoba was recognized in June 2024 to become an international collection following Article 15 of the International Treaty on Plant Genetic Resources for Food and Agriculture.

Adhikari et al. conducted a genomic characterization using more than 45k SNPs of the 1,041 *Aegilops* accessions preserved in the Wheat Genetics Resource Center (WGRC) collection. The *Aegilops* genus contains a range of crop wild relative species which are regarded as critical for future wheat crop improvements. The authors present phylogenetic tree and principal component analyses that showed some species overlap but also pathways of species evolution and diversification. The high genetic diversity identified among the species indicate their importance as genetic resources for future wheat breeding. For genebank curation the study found 49 misclassified and 28 sets of redundant accessions in the WGRC collection.

Berkner et al. worked on genomic prediction to identify winter wheat accessions that can be used to breed varieties with a combined high protein and high lysine content. Their point of departure was a large-scale screening that was conducted back in

the 1970s at the Leibniz Institute of Plant Genetics and Crop Plant Research in Germany. Additionally, they used a genomic dataset generated in 2022. The datasets were curated, four genomic prediction approaches were compared, and the best model was used to predict the traits of interests. Out of the 7,651 accessions included in the predictions, five accessions were highlighted as combining outstanding high protein content with high lysine content.

Chao et al. examined the genetic diversity and population structure of annual medicks (*Medicago* spp.) from the Crimean Peninsula of Ukraine collected in 2008. The collecting mission was done to fill gaps at National Plant Germplasm System in USA and 102 accessions from 10 species were collected. The authors present the results from characterization work, which included 24 phenotypic descriptors and a 3k SNP marker set developed for lucerne (alfalfa). The results showed a high reproducibility between single and pooled biological replicate leaf samples, which indicates that sampling individual plants for these mostly self-pollinating species is sufficient. According to the authors the phenotypic descriptors and the applied SNP marker set was useful in assessing the population structure.

Liu et al. conducted a complete mitochondrial genome characterization and phylogenetic analysis of the endangered species *Prunus pedunculata* in China. The authors highlight that the results provide a basis for understanding the evolution of the genetic background and genetic breeding of *Prunus*.

Abondano et al. compared single plants, multiple plants, and DNA pools sampling strategies for DArTseq genotyping common bean (*Phaseolus vulgaris* L.) landraces from the Alliance Biodiversity and CIAT gene bank. They concluded that pooling tissue from 25 individual plants per accession was a viable approach for characterizing germplasm compared to genotyping individual plants separately by balancing genotyping effort and costs. The results add valuable insights for characterization of collections and in marker-trait association studies.

Loarca et al. evaluated shoot-growth variation in a diversity panel of 695 accessions of carrot (*Daucus carota* L.) from the United States Department of Agriculture National Plant Germplasm System. They found phenotypic variability for seedling emergence and early-season canopy coverage, which is indicating quantitative inheritance and potential for improvement through plant breeding. Accessions with high emergence and vigorous canopy growth are of immediate use to breeders targeting stand establishment, weed-tolerance, or weed-suppressant carrots, which is of advantage to the organic carrot production. In a second paper Loarca et al. evaluated flowering habit trait of the same accessions. They found a high broad-sense heritability for biennial flowering habit which indicates a strong genetic component of this trait.

Li et al. conducted phylogeographic analysis of the native grass *Elymus nutans* using microsatellite markers and covering 361 individual plants across 35 populations from the Qinghai-Tibetan plateau. The species has pastoral and environmental importance, and the study unveiled a notable degree of genetic diversity. Correlations were established between external environmental

factors and effective alleles potentially linked to glutathione S-transferases T1 or hypothetical proteins, which are affecting environmental adaptation.

An et al. analyzed the genetic diversity and structure of a perennial evergreen tree *Albizia odoratissima* using 16 simple sequence repeat markers and covering 280 individuals across 10 populations from Hainan Island and mainland China. The genetic diversity of Hainan population was lower than that of the mainland population. Furthermore there were significant differences in the genetic structure between Hainan and mainland populations.

Lu et al. examined the diversity of an herbaceous climber *Dioscorea bulbifera* native to Africa and Asia and locally used as vegetable and medicine. The study included accessions from mainland China and Taiwan that were analyzed using SSR marker and phylogenetic analyses. They showed structural features across accessions and three distinct clades indicating potential genetic divergence among populations from different geographic regions in China and Taiwan.

Yang et al. present a characterization work on an unusual type of horny goat weed (*Epimedium koreanum* Nakai) discovered in the Jilin Province in China. Horny goat weed is a well-known traditional Chinese medicinal herb that is collected from natural habitats. The newly discovered type had much higher number of leaflets than commonly found (27 compared to 9). By DNA barcoding this novel type was identified as *E. koreanum*. Parallel RNA-seq analysis showed 1171 differentially expressed genes compared to wild type. Due to a decreasing natural population cultivation could be an alternative source for utilization and this high leaf-yielding *Epimedium* plant could be potentially used in breeding or cultivation.

Breeding and seed systems

Two research articles are presented, which includes one article on international breeding collaboration to achieve frost tolerance in potato, and one article overviewing breeding seeds as part of the official seed system in India.

Arcos-Pineda et al. report the results from an international breeding project using a wild potato relative *Solanum commersonii* that resulted in two new frost-tolerant native potato cultivars for the Andes and the Altiplano. The project was a collaboration between partners from USA and Peru as well as the International Potato Center (CIP). After 8 years of breeding the two new cultivars were released. The project shows that international collaboration and the use of valuable genetic diversity can produce results of importance for food security.

Chand et al. provide a retrospective overview on the seed production system of lucerne (*Medicago sativa* L.) in India. Out of 14 lucerne varieties released and notified over the past 24 years, only nine entered the seed chain. The varietal replacement rate was found to be moderate, and the authors present a holistic overview and a way forward to develop more varieties and improved production of certified seeds in the country.

Summary

The provided contributions cover a good mix of topics related to conservation and utilization of plant genetic resources from across the globe, but we miss reports from Africa. We know that more and more research is being done on neglected and underutilized species in this region. On the other hand, we note active work on these matters especially in China. We further note a strong molecular emphasis but that many challenges of managing genetic resources be in genebanks or in *in situ* situations are the same as they were before.

Author contributions

SS: Writing – original draft, Writing – review & editing. MZ: Writing – original draft, Writing – review & editing. AD: Writing – original draft, Writing – review & editing.

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.



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Evenstad Campus, Norway

REVIEWED BY

Ana María Fita,
Universitat Politècnica de València, Spain
Surinder Banga,
Punjab Agricultural University, India

*CORRESPONDENCE

Bum-Soo Hahn
✉ bshahn@korea.kr

[†]These authors have contributed equally to this work

RECEIVED 10 May 2023

ACCEPTED 06 July 2023

PUBLISHED 27 July 2023

CITATION

Subramanian P, Kim S-H and Hahn B-S
(2023) *Brassica* biodiversity conservation:
prevailing constraints and future avenues
for sustainable distribution of plant
genetic resources.
Front. Plant Sci. 14:1220134.
doi: 10.3389/fpls.2023.1220134

COPYRIGHT

© 2023 Subramanian, Kim and Hahn. 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.

Brassica biodiversity conservation: prevailing constraints and future avenues for sustainable distribution of plant genetic resources

Parthiban Subramanian[†], Seong-Hoon Kim[†]
and Bum-Soo Hahn*

National Agrobiodiversity Center, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju-si, Jeollabuk-do, Republic of Korea

The past decade has seen an observable loss of plant biodiversity which can be attributed to changing climate conditions, destroying ecosystems to create farmlands and continuous selective breeding for limited traits. This loss of biodiversity poses a significant bottleneck to plant biologists across the globe working on sustainable solutions to address the current barriers of agricultural productivity. Plant genetic resources centers or genebanks that conserve plant germplasm can majorly contribute towards addressing this problem. Second only to soybean, *Brassica* remains the largest oil-seed crop and is cultivated across 124 countries, and FAO estimates for a combined gross production values of broccoli, cabbages, cauliflower, mustard and rape seeds stands at a staggering 67.5 billion US dollars during the year 2020. With such a global status, wide variety of uses and more recently, growing importance in the health food sector, the conservation of diverse genetic resources of *Brassica* appeals for higher priority. Here we review the current status of *Brassica* conservation across plant genebanks. At present, at least 81,752 accessions of *Brassica* are recorded to be conserved in 148 holding institutes spread across only 81 countries. Several aspects that need to be addressed to improve proper conservation of the *Brassica* diversity as well as dissemination of germplasm are discussed. Primarily, the number of accessions conserved across countries and the diversity of *Brassica* taxa most countries has been highly limited which may lead to biodiversity loss in the longer run. Moreover, several practical challenges in *Brassica* germplasm conservation especially with respect to taxonomic authorities have been discussed. The current review identifies and highlights areas for progress in *Brassica* conservation, which include but are not limited to, distribution of conserved *Brassica* biodiversity, challenges faced by conservation biologists, conservation methods, technical hurdles and future avenues for research in diverse *Brassica* species.

KEYWORDS

plant genetic resources, genebank, *Brassica* biodiversity, taxonomy, *Brassica* conservation, *Brassica* species

1 Introduction

Brassica is one of the highly diverse and largest genera of plants, cultivated and consumed all over the world in its different forms (Fahey, 2016; El-Esawi, 2017). Members of the genus *Brassica* include, but are not limited to morphologically diverse crops including bok choy, broccoli, Brussel sprouts, cabbage, cauliflower, canola, Chinese cabbage, kale, kohlrabi, mustard, rapeseed, and turnips. These plants, given their enormity in diversity and distribution, have equally substantial uses to mankind both directly as health foods, in cuisines, source of oil, source of therapeutics and indirectly as energy crops, in pest control as well as biofumigants (Dixon, 2006; Cornblatt et al., 2007; Szczygłowska et al., 2011; Fahey, 2016; Hagos et al., 2020; Wijaya et al., 2020). Therefore, it is important to study the diverse *Brassica* species to tap their genetic potential for making the best use of them in the future. To accelerate worldwide research on *Brassica*, conservation and dissemination of *Brassica* genetic resources becomes essential with the responsibility falling on plant genetic resource centers or plant genebanks to ensure adequate supply of germplasm. Also, given the narrowing gene pool in *Brassica* crops as a result of continuous breeding for limited selective traits and overall loss of biodiversity due to changing climate conditions, genebanks offer a great choice for conservation of its plant genetic resources including cultivars, crop wild relatives (CWR) and landraces. The International Treaty on Plant Genetic Resources for Food and Agriculture (Plant Treaty) was framed in 2001 for sustainable use of the plant genetic resources for food and agriculture (PGRFA) across the world (Panis et al., 2020; Pathirana and Carimi, 2022). Currently, the Consortium of International Agricultural Research Centers (CGIAR) maintains a network among 15 research centers and with a presence in 89 countries conserves more than 770,000 accessions of plants (CGIAR, 2022). On the other hand, the WIEWS (World Information and Early Warning System) of FAO connects 22,708 member institutes spread across 114 countries that conserve 5.7 million accessions of plant species (WIEWS, 2022). Both of these institutions strive conservation and adequate supply of the plant genetic resources (PGRs) to the global community.

With regards to conservation of PGRs, continuous research is being carried out to 1. Optimize the storage conditions for different crop germplasm as several factors influence the quality of conserved germplasm during long-term storage; 2. Improve the data availability and networking by optimizing the current gene bank information management systems (Kameswara Rao et al., 2017; Weise et al., 2020). This has led to adequate resources being published for proper documentation as well as conservation of PGRs in genebanks (Reed et al., 2004; Food and Agriculture Organization of the United Nations, 2013; Kameswara Rao et al., 2017; Langridge and Waugh, 2019; Volk et al., 2019). In the present agricultural scenario, the responsibility of conserving the genetic diversity of specific plant species and distribution to breeders as well as plant researchers fall majorly on the genebanks. In this present review, we try to summarize the current status of *Brassica* conservation at genebanks in terms of availability and diversity,

the challenges faced by the genebanks in *Brassica* germplasm, hurdles in field collections, and future need for conservation of *Brassica* biodiversity.

2 Current status of *Brassica* conservation

2.1 Data collection

To collect updated information on *Brassica* genetic resources, the following webpages were accessed: FAO-WIEWS, ITIS (Integrated Taxonomic Information System), POWO (Plants of the world online), Genesys, EURISCO, GRIN (Germplasm Resources Information Network), GBIS-IPK (Genebank Information System of the IPK - Institut Für Pflanzengenetik Und Kulturpflanzenforschung Gatersleben), UKVGB (United Kingdom Vegetable Gene bank), RDA (Rural Development Administration, S. Korea), NARO (National Agriculture and Food Research Organization), VIR (N.I. Vavilov All-Russian Scientific Research Institute of Plant Industry), WVC (World Vegetable Center), and CGN (Centre for Genetic Resources, the Netherlands) were accessed. For literature review the keywords used were “*Brassica* germplasm”, “*Brassica* conservation”, “Plant genetic resources”, “*Brassica* biodiversity” at Google scholar website (www.scholar.google.com). Only publications that discussed *Brassica* physiology or germplasm conservation strategies and/or policies were chosen. Preference was given to publications on policies and germplasms status updates published after the year 2017.

2.2 Plant genetic resource centers and *Brassica* germplasm collections

While WIEWS remains to be the global platform for information on conserved plant germplasm, two online data repositories, Genesys, managed by the Crop Trust and the European Search Catalogue for Plant Genetic Resources (EURISCO) also cater as major online directories for obtaining ordering information on specific accessions. Genesys database provides information on about 4.1 million accessions conserved in 450 genebanks all over the world of which approximately less than 1% (0.96%, 41,487 accessions as on 1st December, 2022) belong to *Brassica* (Genesys, 2017) (Figure 1). EURISCO stores data of about 2 million accessions of plants and their wild relatives, preserved in 400 institutes located in 43 member countries of the European Cooperative Programme for Plant Genetic Resources (ECPGR) (Weise et al., 2017). EURISCO stores only 26,321 accessions of *Brassica* (as on 1st December, 2022) which constitute to only 1.3% of the total accessions in the database (Figure 1). *Brassica* also does not appear among the list of top ten crops with at least 100,000 accessions conserved worldwide (Börner and Khlestkina, 2019). However, European nations are major contributors to the overall production of Broccoli, cauliflower and rapeseed/canola in the world with a combined output of over 7.6

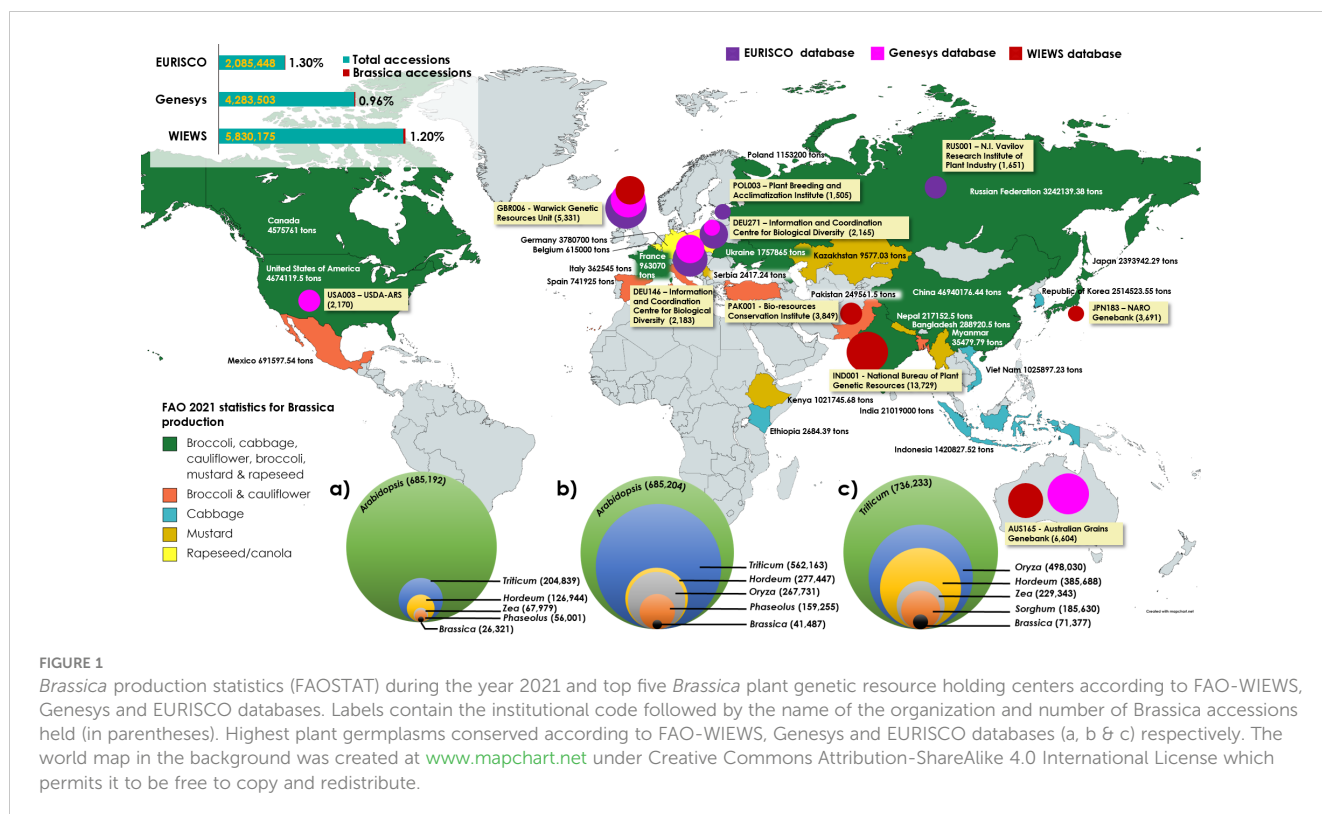


FIGURE 1

Brassica production statistics (FAOSTAT) during the year 2021 and top five Brassica plant genetic resource holding centers according to FAO-WIEWS, Genesys and EURISCO databases. Labels contain the institutional code followed by the name of the organization and number of Brassica accessions held (in parentheses). Highest plant germplasms conserved according to FAO-WIEWS, Genesys and EURISCO databases (a, b & c) respectively. The world map in the background was created at www.mapchart.net under Creative Commons Attribution-ShareAlike 4.0 International License which permits it to be free to copy and redistribute.

million tons of agricultural crop production in 2021 (Figure 1) (FAOSTAT, 1997). There is an evident underrepresentation of Brassica accessions among the conserved germplasm despite their nature of being such a diverse and anthropologically important genus of crop plants (Kellingray et al., 2017; Wijaya et al., 2020; Salehi et al., 2021).

A comparison of overall Brassica production in 2021 and major plant genetic resources centers involved in conserving Brassica biodiversity across the world indicate that Asia, the United States and Europe remain to be major producers of Brassica crops (Figure 1). The WIEWS official data from the SDG 2.5.1.a indicates that a total of 71,378 accessions of Brassica are currently maintained *ex situ* all over the world (Table S1). Primary observation may indicate that the highest number of accessions being conserved in genebanks in Europe and Asia, but in terms of species diversity, after Europe, North America (USDA-ARS) conserves diverse species of Brassica germplasm (Table S1). Further analysis indicate that less than 10 accessions are conserved in genebanks from Africa (Niger, Nigeria, Eswatini, South Africa, Libya), Middle East (Cyprus, Tajikistan, Jordan, Lebanon), Eastern Europe (Georgia, Montenegro, Armenia) and South America (Argentina, Mexico) (WIEWS, 2022). However, these parts of the world also contribute significantly to the world Brassica production and fall among the top 5 producers of crops such as Broccoli, cauliflower, cabbage and mustard seeds (Figure 1). Improving the available Brassica species biodiversity and dissemination could considerably advance further breeding and research in these parts of the world.

Currently, for ordering Brassica germplasm, the Germplasm Resources Information Network (GRIN) and Gene bank

Information System (GBIS) from the US (maintained by United States Department of Agriculture - Agricultural Research Service (USDA-ARS) and Leibniz Institute of Plant Genetics and Crop Plant Research serve as channels where one can directly request Brassica germplasm for research purposes (Postman et al., 2010; Oppermann et al., 2015). In addition to these widely used genebanks, several other facilities are also available to source Brassica germplasm locally and internationally albeit the diversity being very limited (Liu et al., 2020). The current status of Brassica germplasm in major genebanks are given in the table below (Table 1). A Principal Component Analysis (PCA) of the available Brassica germplasm information indicated that country such as India conserves very high number of accessions (accession) whereas Korea, Germany, Great Britain conserves high diversity of Brassica species (unique). At the same time, high number of accessions conserved in Korea and Japan possess data on their biological status (status) (Figure 2A). Comparison of species indicates that *B. juncea*, remains the highly conserved Brassica taxon followed by *B. napus*, *B. rapa* and unsegregated *B. spp* contribute to the highest number of accessions stored in genebanks (Figure 2B). A majority of these Brassica germplasm are maintained as *ex situ* collections by germplasm collections and materials are preserved in short, medium and long term storage and available as seeds, plant material, and/or DNA (Walters et al., 2004; Acker et al., 2017; Lusty et al., 2021). A search on Genesys and EURISCO webpages indicate that a majority of Brassica germplasm are stored as long term storage, followed by seed collections and mid-term storage (Figure 2C). The major plant material conserved in all these three strategies are of seed type and seeds are the easiest way for exchange or dissemination of Brassica germplasm.

TABLE 1 Major genebanks and germplasm information (as accessed on 5th October, 2022).

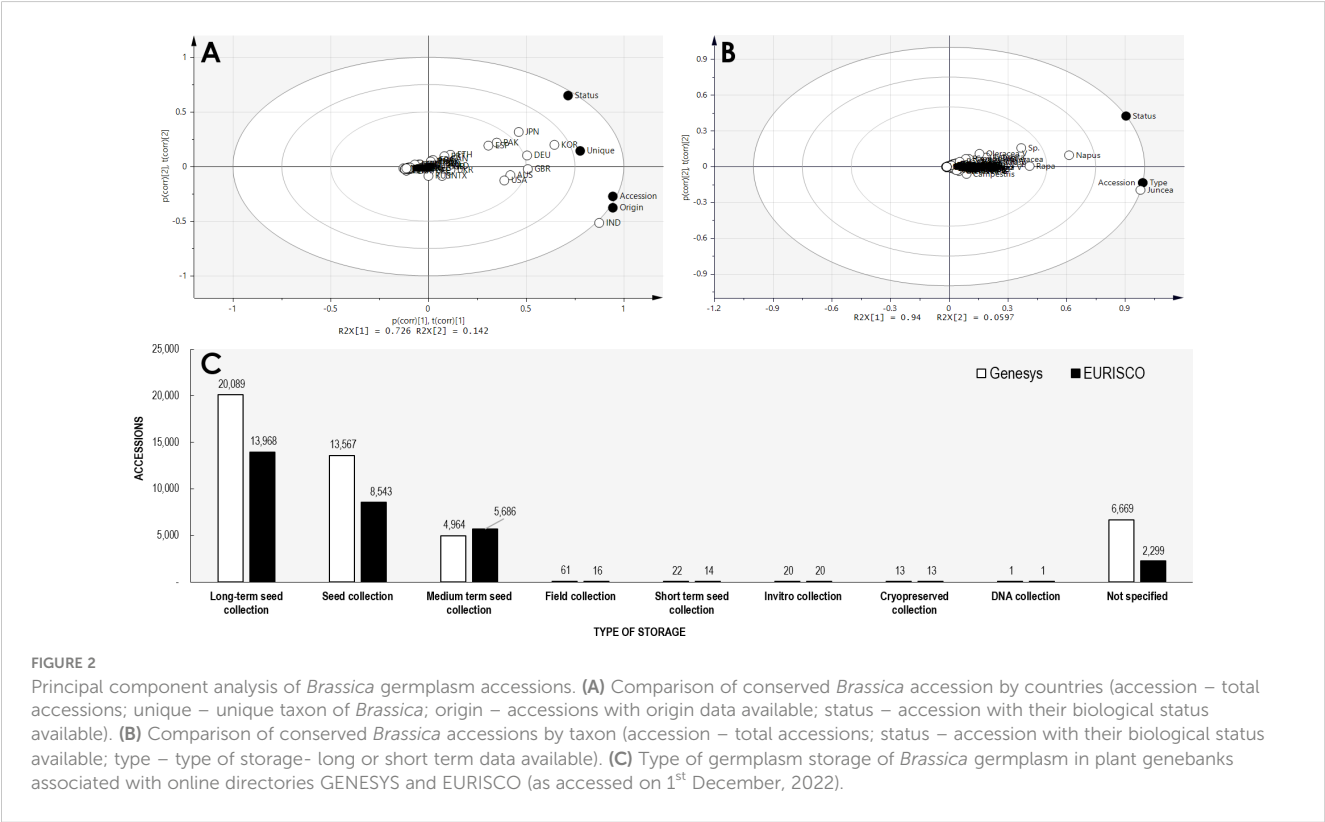
Country	INDIA	AUSTRALIA	KOREA	UNITED KINGDOM	RUSSIA	GERMANY	PAKISTAN	USA	JAPAN
Resource	NBPGR	AGG	RDA	UKVGB	VIR	GBIS-IPK	BCI	GRIN	NARO
<i>Brassica</i> accessions	13,729	6,604	6,399	5,314	4,881	4,351	3,849	3,167	1,833
Total accessions	420,324	164,044	272,351	12,158	243,829	129,748	23,722	42,743	225,811
Percentage <i>Brassica</i> germplasm	3.27%	4.03%	2.40%	43.70%	2%	3.40%	16.23%	7.40%	0.80%
Species diversity	13	26	20	23	15	31	9	24	7
Subtaxa diversity*	42	43	58	29	41	77	12	63	25
Major taxa conserved	<i>B. juncea</i> , <i>B.campestris</i> , <i>B. rapa</i>	<i>B. rapa</i> , <i>B.napus</i> , <i>B. juncea</i>	<i>B. napus</i> ; <i>B. juncea</i> ; <i>B. rapa</i>	<i>B. oleracea</i> ; <i>B. napus</i> ; <i>B. rapa</i>	<i>B. juncea</i> ; <i>B. oleracea</i> ; <i>B. rapa</i>	<i>B. oleracea</i> ; <i>B. napus</i> ; <i>B. rapa</i>	<i>B.juncea</i> , <i>B. campestris</i> , <i>B. rapa</i>	<i>B. oleracea</i> ; <i>B. rapa</i> ; <i>B. napus</i>	<i>B. napus</i> ; <i>B. rapa</i> ; <i>B. oleracea</i>

NBPGR, National Bureau of Plant Genetic Resources; AGG, Australian Grains Genebank; RDA, Rural Development Administration; UKVGB, United Kingdom Vegetable Gene bank; VIR, N.I. Vavilov All-Russian Scientific Research Institute of Plant Industry; GBIS-IPK, Gene bank Information System of the IPK Gatersleben; BCI, Bio-resources Conservation Institute; GRIN, Germplasm Resources Information Network; NARO, National Agriculture and Food Research Organization. *Number of unique species, subspecies or varieties available.

2.3 Worldwide assessment of conserved Brassica germplasm

Currently, comparison of the three online directories yields a total of 81,752 *Brassica* accessions stored in 81 countries across the world in addition to 3 regional and 3 international research centers (data summarized from WIEWS, Genesys & EURISCO).

Additionally, few genebanks such as The National Crop Genebank of China (NCGC) and National Agrobiodiversity Center-RDA, South Korea also conserve *Brassica* germplasm but are limited for sharing of resources due to governmental policies. Historically, the six most common species of *Brassica* explained by the U triangle are well documented and are widely conserved all across the world (U, 1935) (Figure S1). It can be observed that



genebanks that conserve high number of accessions of species such as *Brassica napus*, *B. oleracea* and *B. carinata* are located in vicinity to their geographical origins (Figure S2). Across Asia, green leafy vegetable crops of *B. rapa* (subsp. *chinensis* and *pekinensis*), oil producing rape *B. napus* and mustards of *B. juncea* with local origins are highly conserved in genebanks of India, Japan, Pakistan, Taiwan and South Korea (Figure S2). Co-ordination and dissemination of *Brassica* germplasm have resulted in countries such as Australia, Canada and the United States of America (USA) which are geographically far from the natural areas of origin of *Brassica* species to accumulate in both number and diversity (Table 1; Figure 3).

A country wise comparison of the WIEWS and Genesys databases for conserved *Brassica* accessions indicate that Australia, India, Germany, United Kingdom (UK), USA and Taiwan conserve the largest number of *Brassica* germplasm (Figure 3A). In terms of species diversity, the most common *Brassica* crop seeds conserved are *B. oleracea* (WIEWS: 17,389 accessions; Genesys 14,798 accessions); *B. rapa* (WIEWS: 15,132 accessions; Genesys 8,991 accessions) and *B. juncea* (WIEWS: 14,810 accessions; Genesys 4,698 accessions). However other *Brassica* species such as *B. tyrhena*, *B. taurica*, *B. bioniana*, *B. atlantica*, *B. nivalis*, and *B. aucheri* etc. are not highly conserved (<10 accessions). Moreover, there also exists disproportion in the biodiversity conserved at the plant genebanks. Among 146 plant

genebanks recorded to conserve *Brassica* >38% genebanks (56) have a *Brassica* biodiversity of less than 10 taxa (including subspecies and varieties) and largely conserve *B. oleracea*, *B. rapa* and *B. juncea* genetic resources. Data from FAO-WIEWS indicate that Australia, Germany, Spain, UK and USA conserve more than 25 representative species of *Brassica* including wild types (Figure 3B). Although the FAO promotes sharing of plant genetic resources for food and agriculture (PGRFA) through its sustainable developmental goals (SDGs), much work has to be done to ensure that the currently available biodiversity of *Brassica* species is not lost. Crop wild types are excellent sources for breeding for development of disease tolerant and abiotic stress tolerant plants that would be suitable for fresh challenges posed by the changing climate conditions.

3 Conservation methods for *Brassica* germplasm

The Seed Information Database (SID) maintained by Millennium Seed Bank of the Royal Botanic Gardens, Kew, UK designate all the 48 taxa of *Brassica* (species and subspecies) stored by them to be orthodox seeds. The orthodox nature of seeds from *Brassica* has in fact been well established (Genesys, 2017; Solberg et al., 2020). Therefore, long-term storage at -18 to -20°C for 40-60

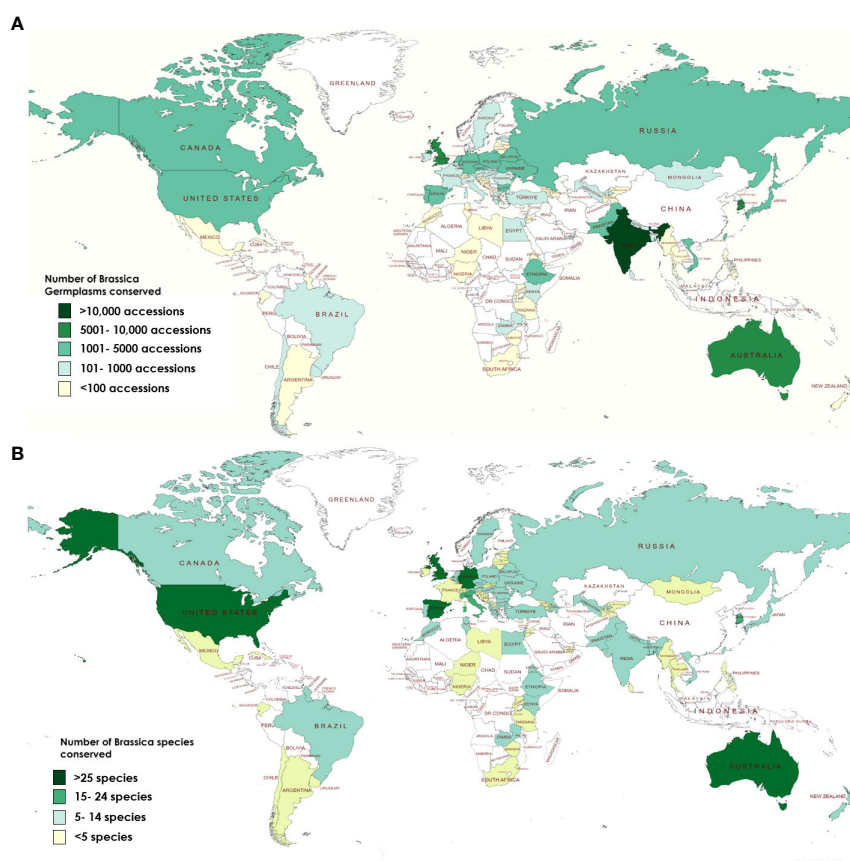


FIGURE 3
Conservation of *Brassica* biodiversity in terms of (A) Number of accessions stored (B) Species diversity across plant genebanks around the world.

year range is possible (Solberg et al., 2020). This explains the nature of storage of *Brassica* seeds, where a majority of germplasms are currently held in long term storage (Figure 2C) (Weise et al., 2017). However, other types of storages such as *in vitro* collection, DNA collection, cryopreservation are also employed to preserve *Brassica* germplasm, but scarcely used. Further studies or research are warranted on conserved germplasms to not only understand the characteristics and traits of stored seeds, but also to improve the currently followed storage practices (Acker et al., 2017; Panis et al., 2020). Several studies have been conducted in the past on monitoring dormancy, re-generation and suitability/effects of different temperatures as well as periods of storage for varied plants (Ellis et al., 2018; van Treuren et al., 2018; Everingham et al., 2021). Germplasms of *Brassica* conserved in gene banks have also been subjected to studies to optimize their storage conditions, regulation of seed dormancy and elucidation of useful characteristics such as metabolite profiling.

Seeds from members of family *Brassicaceae*, have been reported to commonly have a half-life (P_{50}) of 54 years and some plants such as cabbage (*Brassica oleracea* L.) can even be maintained at +20°C and 50% RH up to 5 years in paper bags (Nagel and Börner, 2010; Solberg et al., 2020). Seeds of *B. villosa* ssp. *drepanensis* exhibited a high germination rate (92%) when stored at -20°C with a moisture content of 3% to 6% for 16 years. However, room temperature stored seeds succumbed to rapid deterioration after 3 to 4 years (Scialabba et al., 2016). Study of long term (36 years) stored *Brassicaceae* members showed that desiccation to reduce seed water content to less than 2.5% of fresh weight significantly improved germination at the end of storage. This effect was enhanced by enrichment of the storage environment with carbon dioxide (CO_2) (Gonzalez-Benito et al., 2011). Hermitic storage of 37 *Brassicaceae* members including *B. napus*, *B. fruticulosa* and *B. fruticulosa glaberrima* at -5°C to -10°C with seed moisture content of 0.3 to 3%, 20% RH, indicated sustenance of seed viability over a period of 38-40 years (Pérez-García et al., 2007). Similar studies on seeds of *Brassica fruticulosa* and *B. repanda* stored at -5°C to -10°C with seed moisture content of <3%, 20% RH, over a period of 40 years from 1966 to 2006 also indicated sustained viability, displaying a higher germination rate of over 70% without use of dormancy-breaking agents such as gibberellic acid (Pérez-García et al., 2009). A study comparing *Brassica rapa* L. ssp. *pekinensis* and *Capsicum annuum* L. long term (10 years) seed storage using different packaging material indicated that *B. rapa* L. ssp. *pekinensis* was the most affected by storage at ambient temperatures even when packaged in vacuum-sealed aluminum pouches, but could be successfully stored for up to 10 years at 5°C/ RH 30, using vacuum-sealed aluminum pouches (Soh et al., 2014). From these studies it can be understood that while some *Brassica* plants such as Cabbage (*Brassica oleracea* L.) can be stored under ambient temperatures (+20°C) it may not be suitable for others plants such as *B. rapa* L. ssp. *pekinensis* and *B. villosa* ssp. *drepanensis* which are to be stored at much lower temperatures of -5°C to -20°C. Other optimal conditions reported for *Brassica* members include a relative humidity of less than 50%, moisture

content of <6%, use of CO_2 if stored in sealed containers or use of vacuum sealed aluminum pouches. Also, extreme desiccation of seeds during storage have been seen to have a detrimental effects on seed longevity on plants such as *B. repanda* (Mira et al., 2015).

Other studies on *Brassica* species conserved in gene banks are found to be on the biochemical changes that occur in seeds during long term storage and screening of collections for their metabolites. Studies on seed lipid peroxidation and membrane permeability in *Brassicaceae* members during long term storage and their effect on seed germination, vigor indicated that increased membrane permeability to occur in *Brassica repanda* seeds which led to reduced germination % as well as vigor loss (Mira et al., 2011). In another study, a seed lipid thermal fingerprinting using differential scanning calorimetry (DSC) was proposed, which used several biophysical markers to predict performance of oily seeds of *Brassicaceae* during long-term storage (8-44 years) (Mira et al., 2019). A study on 168 accessions of *Brassica rapa* on understanding relationships between genetic markers and metabolites developed a set of genetic markers could be used for rapid selection of specific metabolite producing *B. rapa* genotypes (Pino Del Carpio et al., 2011).

4 Challenges in *Brassica* germplasm conservation

4.1 Taxonomic predicament of *Brassica* germplasm in genebanks

A major concern with worldwide *Brassica* germplasm collections often is the uncertainty associated with the naming of plant genetic resources. Descriptions of the origins, phenotypic and genotypic diversity of several *Brassica* members have been published (Dixon, 2006; Bonnema et al., 2011). However, naming of the *Brassica* members hitherto remains complex and often ambiguous due to differences in the naming conventions. Proper phylogenetic organization of the members of *Brassica* has been arduous due to extensive crossing and hybridization over centuries which occurred across geographically isolated microenvironments causing difficulties in tracing and comparison (Bonnema et al., 2011; Fahey, 2016). There is also an overlap of plant forms and morphological traits across species which further complicates phenotypical phylogenetic assignment and organization (OECD, 2016). The origins of modern *Brassica* vegetables and their taxonomy continues to be unsettled and regarded to be in a state of constant flux (Fahey, 2016).

At present, only 13 subordinate taxa in the rank of species are recognized by the Integrated Taxonomic Information System (ITIS) in the genus *Brassica* (ITIS, 2022) (as on 05th October, 2022) (Table S2). The ITIS is a taxonomic database maintained by the United States government which partners with other governmental agencies of Canada and Mexico as well as organizations dedicated to biodiversity such as GBIF (Global Biodiversity Information Facility) (GBIF, 2020). However, other well-recognized online catalogues as well as germplasm collections do not necessarily

follow these naming conventions. For example, as on 05th October, 2022, the Plants of the World Online-Royal Botanical Gardens, Kew in collaboration with World Flora Online project, provides taxonomic recognition of 41 species of *Brassica* (Borsch et al., 2020; POWO, 2022) (Table S2). These names are accepted and added to its 'World Checklist of Seed Plants' (Govaerts et al., 2022). Another repository BrassiBase, dedicated towards developing taxonomic, evolutionary, systematics and germplasm knowledge systems specifically for *Brassicaceae*, reports 45 species of *Brassica* (Kiefer et al., 2013) (Table S2). The GRIN-taxonomy resource which provides taxonomic information on plants conserved by the United States National Plant Germplasm System (USNPGS) remains as the only provider of taxonomic information along with conserving germplasm. The GRIN-taxonomy also currently recognizes 41 species of *Brassica*. But they also are not completely in agreement with the accepted list of *Brassica* species given the World Checklist of Seed plants. On a comparison of the *Brassica* taxa recognized by the plant taxonomy authorities, online PGR directories and plant genebanks we recognized issues related with the naming of *Brassica* taxa at species level (Figure 4). There are 18 *Brassica* species that are present only among the plant taxonomy lists (ITIS, POWO, BrassiBase & GRIN-Taxonomy) which are not conserved under the designated name at any germplasm conservation center or present in the list of *Brassica* species in the online repositories (Genesys & EURISCO) (Figure 4A). Similarly, it can also be observed that there are 15 *Brassica* species which are listed only in the online PGR directories (Genesys & EURISCO). This could be solved as they all of them are subspecies of *Brassica*, particularly *Brassica oleracea* and *B. rapa* which have been misnamed or updated to a new taxon (Figure 4B). Same is the case of *Brassica sylvestris* which was formerly designated as *Brassica rapa* subsp. *sylvestris* which is now recognized only in plant genebanks but not by plant taxonomy authorities or online PGR directories. The NCBI taxonomy database under *Brassica*

(Taxonomy ID:3705) indicates 129 distinct taxa which after removing species crosses and taxa marked ambiguous names such as groups provide 87 unique taxa (Figure 5). These 87 unique taxa fail to include all the *Brassica* species conserved in plant genebanks.

The ambiguity of *Brassica* naming also continues to exist at the ground level. To begin with, the plant genetic resources storage centers around the world which are responsible for conservation and distribution of germplasm do not reflect the diverse species richness in *Brassica* germplasm conserved by them. The Gene bank Information System of the Leibniz Institute of Plant Genetics and Crop Plant Research (GBIS-IPK) indicates storing of accessions belonging to 31 species of *Brassica* whereas the United States Department of Agriculture- Germplasm Resource Information Network (USDA-GRIN) stores 24 different species of *Brassica* (excluding *Brassica* sp.) (Brassica-dataset, 2015; Oppermann et al., 2015) (Table S2). It is understandable that plant genebanks concentrate towards conserving local resources and rely on crop nomenclature. However, as most of them follow the *ex situ* model of conservation, they can also be used as an effective tool in protecting and documenting the *Brassica* biodiversity. For example, as several of the accepted taxa of *Brassica* such as *Brassica assyriaca*, *B. baldensis*, *B. beytepeensis* etc. have been officially added as verified and accepted species (Govaerts et al., 2022), plant genetic resource centers around the world can attempt to work towards taxonomic reclassification and updating of information on the bioresources which they conserve.

The problem of taxonomic uncertainties becomes much more complex across genebanks all around the world when describing the status at much lower taxonomic levels such as subspecies, cultivars and varieties which are often interchanged depending on the author/ gene bank accessed. Such issues at times can complicate the taxonomic assignment after breeding or during research. Introduction of universal code for naming at this juncture, would be a tedious as well as an incredible task. A polymerase chain reaction

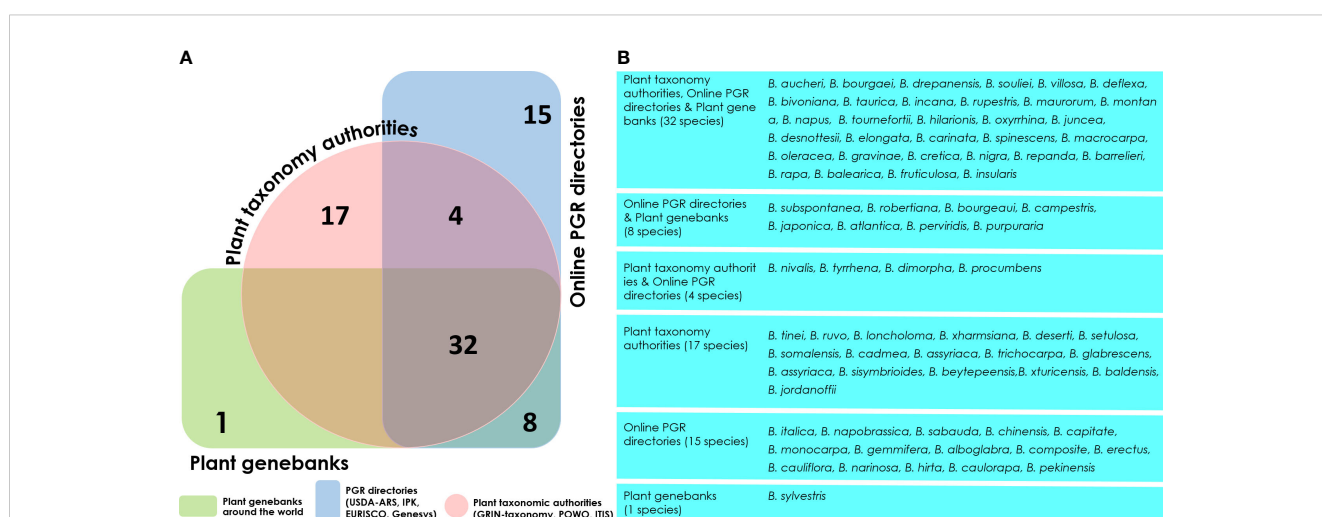


FIGURE 4

Venn diagram illustrating the distribution of *Brassica* species. (A) number of *Brassica* species available across Plant taxonomy authorities (Integrated Taxonomic Information System-ITIS, Plants of the world online-POWO, GRIN taxonomy), online plant genetic resource directories (WIEWS database, Genesys and EURISCO) and plant genebanks. (B) List of *Brassica* taxa in groups specified in the Venn diagram.

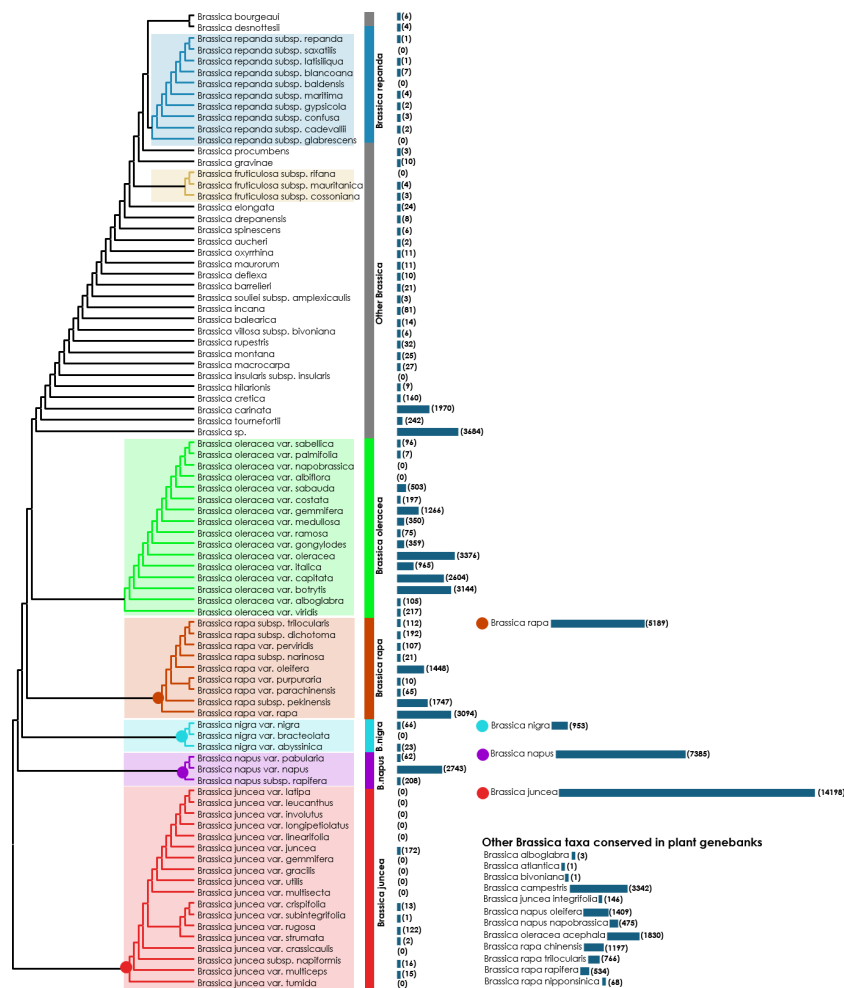


FIGURE 5

A common tree drawn using the NCBI taxonomy database (Sayers et al., 2019; Schoch et al., 2020). The base Newick's tree was downloaded from the NCBI taxonomy database and redrawn using MEGA v. 11. Bars along with numbers in the parentheses indicate the number of accessions available according to the WIEWS database.

(PCR) based approach for identification of *Brassica* species has been proposed albeit its efficacy is limited within the species from the U triangle (Koh et al., 2017). There is a possibility that this problem could be solved by help of genetic sequencing approaches. Since the advancement of sequencing technologies, genomic component based typing followed by distinctive universal taxonomic assignments have helped to overcome such barriers. This would also shorten the time required by traditional identification using phenotyping. The current requirement on *Brassica* taxonomy would therefore involve a polyphasic approach for phylogenetic organization which would combine more than one of the following: phenotyping or studying morphological traits; use of cytological markers, karyotyping; phytochemical, storage protein profiling; studying isozyme markers; microsatellite markers such as SSRs (short tandem repeats), SNPs (single nucleotide polymorphisms); and sequencing methods such as specific-locus amplified fragment sequencing (SLAF-seq) and using large scale genomic analyses and fingerprinting to categorize *Brassica* species (Pino Del Carpio et al., 2011; Liu et al., 2013; El-Esawi, 2017; Fotev et al., 2018; Yang et al., 2018; Bhandari et al., 2020; Chao et al.,

2020; He et al., 2021; Rakshita et al., 2021; Hong et al., 2022; Li et al., 2022)

4.2 Diverse morphotypes and cytogenetics of *Brassica*

The richness of both species and morphotypes in *Brassica* can be understood to have arisen after the whole genome triplication (WGT) event that has been accepted (Arias et al., 2014; Cheng et al., 2014). Among the speciation concepts of *Brassica*, the U triangle is the most famous and accepted concept of cytogenetic relationships between the species belonging to the genus *Brassica* (Figure S1). The genome types of the ancestral diploid and tetraploid as explained in the triangle of U has been well discussed (Branca and Cartea, 2011). In addition to the six species of *Brassica* explained by the U triangle, other species of *Brassica* have also been continually described (Govaerts et al., 2022) (Figure 6). These are morphologically distinct compared to the six species and not much information is

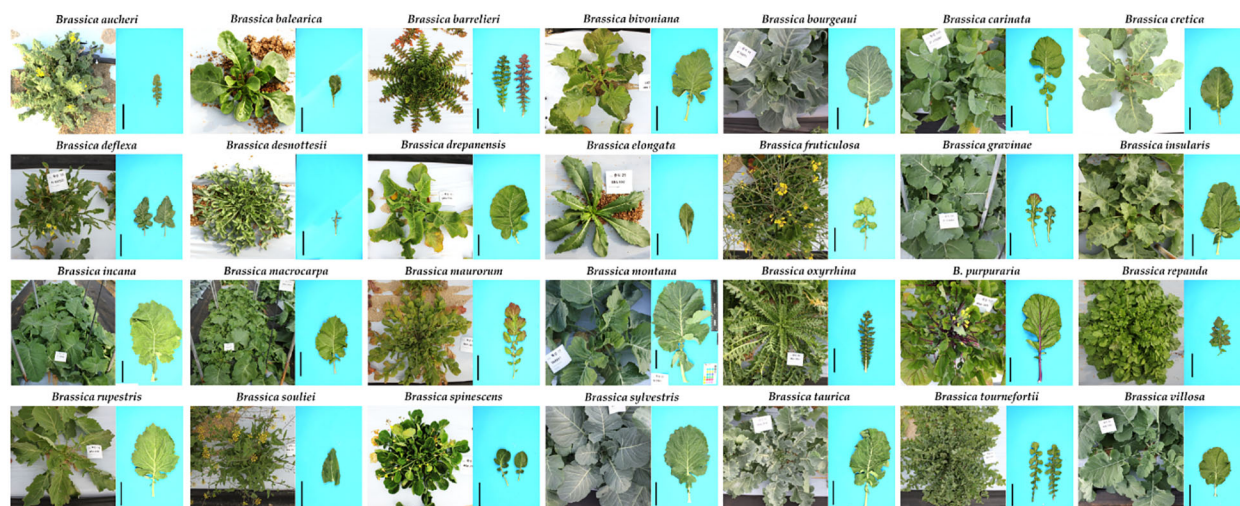


FIGURE 6
Morphologically varied wild type species of *Brassica*. The scale bars (in black) adjacent to the individual leaves represent a length of 10 cm and were included using ImageJ software (Unpublished data).

available about them for scientific research. There is a marked morphological diversity among various species of *Brassica* where some form heads or curd or stem and lead modifications. This variation among the species can be observed in their leaves – size, shape, margins, presence of anthocyanin; stems; habit; flower color and roots. To give few examples, the *Brassica* species *B. spinescens* and *B. balearica* form slight succulent leaves which is unique among various *Brassica* and *B. maurorum* produces minute flowers which look like an inflorescence (Figure 6). Some plants grow more than 2 meters in height (*B. rapa*) where are some (*B. elongata*, *B. balearica* and *B. maurorum*) do not grow more than 30 cm in height. Despite these plants having a long history of documentation and exhibiting significant difference in their morphology, habit and growth periods, they have not been studied in detail or conserved in large number in plant genebanks. In the recent decades, adequate emphasis has been given in literature to the importance of conservation of wild type species. These other *Brassica* species are vital as they can help to deal with the bottleneck problems associated with *Brassica* breeding which warrants the need for conservation of these wild type species of *Brassica*.

4.3 Technical hurdles in cultivation of *Brassica* species

Members of the family *Brassicaceae* family especially the vegetable crops, prefer to grow under cooler temperate conditions (15–20°C), with high humidity, but some species are also found to grow also under subtropical and even tropical zones (Fahey, 2016; Lin et al., 2018; Razzaq et al., 2021). For example Chinese cabbage plants grow well at a temperatures range of 13–20°C but prefer warmer growing conditions and often bolt at lower temperatures (Salehi et al., 2021). *Brassica juncea* or mustard plants widely grown on northern parts of India also prefer warmer temperatures as high as 27°C and can tolerate temperatures as low as 6°C (Shekhawat

et al., 2012). Among other species, Brussels sprouts and kale are the most resistant to low temperatures, as they can tolerate temperatures as low as –12°C. In addition to preference of longer photoperiods for optimized growth, *Brassica* vegetables such as kohlrabi, Chinese cabbage, cabbage have a weaker root system and require humus rich or loamy soil with pH near neutral (6.2–7.0) and high water retention properties (Dhaliwal, 2017; Razzaq et al., 2021). Therefore, propagation of diverse *Brassica* species requires environment controllable facilities and not all of them can be multiplied in outdoor fields in all countries.

It has been reported that some *Brassica* species are self-fertile but most of them are self-incompatible and may require cross pollination to obtain seed (Stewart, 2002). This pollination can occur through wind, animals or insects or through human intervention. Though the human intervention aspects are related to breeding, cross-pollination by wind or insects such as bees are undesirable and can contaminate varieties. In case of wind pollination, isolation of plants is very helpful to avoid cross-pollination and incase of *B. napus* outcross could be reduced to less than 4% on separating plants by 120 cm; in oil seed rape almost 10% of pollen could be detected at a distance of even 360 m from the edge of the field; in radish also, the minimal distance for 0% crossing was at least 50 m (Stewart, 2002). The plant *B. napus* is highly prone to cross-pollination and spontaneous inter-specific hybridization of *B. napus* is possible with sexually compatible species (Meglic and Pipan, 2018). To produce pure seeds, *Brassica* plants need to be grown in isolation and cross-pollination risk reduces with the distance or introducing physical barriers in the flowering plants. However, as explained earlier, cross pollination is not a common phenomenon and rates are very low among *Brassica* spp. (Stewart, 2002). Avoidance of cross pollination can be carried by using a number of strategies including creating topographical barriers such as edges that would restrict or limit wind and insects, shifting of bee hives, establishment of border zones at the edges of the where seeds will not be collected from the plants in the zones and planting fields

in such a way that the flowering time and period do not overlap (Stewart, 2002).

4.4 Diseases and pests

Almost all the *Brassica* vegetables are relatively susceptible to the same or similar diseases as well as insect pests (Fahey, 2016). Overall, pests are a serious threat to agricultural productivity and lately, with the exponential growth of *Brassica* cultivation, the distribution of *Brassica* vegetable pests have also increased (Witzel et al., 2021). The common microbial diseases occurring in *Brassica* species are tabulated (Table S3). The most common microbial infections in *Brassica* species are most caused by fungi especially *Leptosphaeria* (Blackleg disease), *Sclerotinia* (rot) and *Verticillium* (wilt). However, clubroot disease caused by *Plasmodiophora* is the highly reported disease among *Brassica* plants. Among bacterial pathogens *Pseudomonas* and *Pectobacterium* are the most common and Turnip yellow virus is the most common viral pathogen. In addition to diseases, *Brassica* crops are also attacked by a variety of pests that affect the growth as well as yield of *Brassica* members (Table S4). These causes a variety of issues ranging from feeding on the leaf tissues to killing the whole plants by feeding on plant roots. The most common insect pests of *Brassica* include aphids, worms, mites and weevils. However, there are also several beneficial microorganisms and insects that aid growth of *Brassica* species. These help the plants grow and prevent predators from feeding on the plants. Microorganisms including bacteria and fungi have been reported to help *Brassica* species by stimulating plant growth through mechanisms such as nitrogen fixation, siderophore production, alleviate chilling and heavy metal stress, and more importantly act as biocontrol agents against several bacterial and fungal diseases (Card et al., 2015). In addition to microbial interactions leading to enhanced growth of *Brassica* plants, some of the insects have also been reported to aid *Brassica* plants by acting as major predators of pests and feed on eggs of pests, aphids, small caterpillars, thrips and mites that attack the plants (Table S5).

5 Importance of conserving *Brassica* biodiversity

The culinary and therapeutic use of species in the U triangle has been well established and needs no emphasis. However, potential uses of other species are also currently gaining importance. For example, wild type accessions of *Brassica montana* and *B. balearica* were reported to exhibit resistance against the black rot caused by *Xanthomonas campestris* which particularly affects *B. oleracea* crops (Sheng et al., 2020). Similarly, *B. fruticulosa*, *B. hilarionis*, *B. macrocarpa*, *B. montana*, *B. spinescens* and *B. villosa* have been reported to exhibit antibiosis against the cabbage root fly (*Delia radicum*) which is a major threat to cabbage production in parts of Western Europe, the United States and Canada (Shuhang et al., 2016). Moderate resistance against other plant pathogen *Sclerotinia sclerotiorum* have reported in other *Brassica* species including *B. atlantica*, *B. bourgeai*, *B. incana*, *B. macrocarpa*, and *B. villosa*

(Taylor et al., 2018). Presence and applications of glucosinolates from *Brassica* has been well established in *B. oleracea*, *B. juncea* and *B. rapa*. The species *B. bourgeai* has also been studied for their glucosinolates and have been reported to contain significantly high total glucosinolates than several *B. oleracea* crops in particular gluconapin, glucoraphanin and neoglucobrassicin (Tortosa et al., 2017). Glucosinolates have also been studied in *B. desnottesii*, *B. drepanensis*, *B. elongata*, and *B. gravinae* (Montaut et al., 2017; Malfa et al., 2022). All of the above studies have been made on germplasm obtained from plant genebanks and therefore, further research on the prospects of wild type *Brassica* can be significantly improved if these accessions are locally available for researchers.

6 Conclusions

The genus *Brassica* has given rise to a number of agriculturally important vegetable crops since ancient times. Moreover, it also provides the second major oilseed crop next only to soybean (Gupta, 2016). Conservation of *Brassica* germplasm and its biodiversity has still room for progress at the current scenario. *Brassica* vegetables has consistently been gaining popularity especially in the health food sector and are currently consumed across the Americas, Asia, and several European countries. Several parts of the plants are economically important depending on the species and those include leaves, seeds, oils. Thus their economic importance has multiplied and are now grown along with cereal crops to meet global demand (Salehi et al., 2021). With respect to medicine, members of genus *Brassica* are rich in biologically active metabolites which have been reported to possess potential health-promoting properties (Lee et al., 2013; Solov'yeva et al., 2013; Lee et al., 2014; Klopsch et al., 2017; Bhandari et al., 2020). Bioactive metabolites other than glucosinolates have also been identified and reported in *Brassica* which include alkaloids, anthocyanins, carotenoids, flavonoids, folates, other phenols, phytoalexins, phytosteroids and tocopherols (Pino Del Carpio et al., 2011; Branca et al., 2018; Lotti et al., 2018; Ramirez et al., 2020; Salehi et al., 2021). In addition to the health benefits of *Brassica* plants, in recent times, their commercial importance has grown in the food industry in which their morphological appeal to the consumers remains an important factor. This includes the appetizing appearance which is a combination of properties such as color, size, shape of the edible parts of the plant. Gene bank germplasm can therefore be screened for accessions with such attractive morphological traits. Earlier screening has been carried out for form/appearance significant *Brassica* vegetables including Kale and cauliflower (Thorwarth et al., 2018; Witzel et al., 2021). Further studies have also been made which included morphological properties of Chinese broccoli (*B. oleracea alboglabra*), and *B. rapa* (Fotev et al., 2018; Witzel et al., 2021). As crop wild relatives can serve as potential source of both biomedical compounds as well to improve commercial important phenotypic characteristics, conservation of the available biodiversity of *Brassica* becomes a prerequisite. In addition, they can also serve as a source of several useful traits which can be exploited to adapt crops to changing climate and improve biotic as well as abiotic resistances in *Brassica* crop species (Quezada-Martinez et al., 2021). Current

status indicates conserved *Brassica* to be highly concentrated on few species and unevenly distributed across countries. Plant genebanks can provide an excellent solution to preserve the *Brassica* biodiversity for research, breeding, and commercial purposes.

Author contributions

B-SH was involved in the conceptualization of the manuscript. PS and B-SH were involved in writing of the manuscript. S-HK was involved with data analysis, preparation of figures and tables. All authors contributed to the article and submitted and approved the submitted section.

Funding

The study was supported by the grant (PJ01672201) funded by the National Agrobiodiversity Center, Rural Development Administration, Republic of Korea. This study was also supported by the 2023 Postdoctoral Fellowship Program (P.S.) of the National Institute of Agricultural Sciences, RDA, Republic of Korea.

References

- Acker, J. P., Adkins, S., Alves, A. A. C., Horna, D., and Toll, J. (2017). *Feasibility study for a Safety back-up Cryopreservation facility*. Independent expert report: July 2017. Rome (Italy): Bioversity International. 100p.
- Arias, T., Beilstein, M. A., Tang, M., McKain, M. R., and Pires, J. C. (2014). Diversification times among *Brassica* (*Brassicaceae*) crops suggest hybrid formation after 20 million years of divergence. *Am. J. Bot.* 101 (1), 86–91. doi: 10.3732/ajb.1300312
- Bhandari, S. R., Rhee, J., Choi, C. S., Jo, J. S., Shin, Y. K., and Lee, J. G. (2020). Profiling of Individual Desulfo-Glucosinolate Content in Cabbage Head (*Brassica Oleracea* var. capitata) Germplasm. *Molecules* 25 (8) 1860. doi: 10.3390/molecules25081860
- Bonnema, A., Del Carpio, D., and Zhao, J. (2011). Diversity analysis and molecular taxonomy of *Brassica* vegetable crops, in *Genetics, Genomics and Breeding of Vegetable Brassicas*. New York: CRC Press. pp. 81–124.
- Börner, A., and Khlestkina, E. K. (2019). *Ex-situ* genebanks—Seed treasure chambers for the future. *Russian J. Genet.* 55 (11), 1299–1305. doi: 10.1134/S1022795419110036
- Borsch, T., Berendsohn, W., Dalcin, E., Delmas, M., Demissew, S., Elliott, A., et al. (2020). World Flora Online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. *TAXON* 69 (6), 1311–1341. doi: 10.1002/tax.12373
- Branca, F., and Cartea, E. (2011). "Brassica," in *Wild Crop Relatives: Genomic and Breeding Resources: Oilseeds*. Ed. C. Kole (Berlin, Heidelberg: Springer Berlin Heidelberg), 17–36.
- Branca, F., Chiarenza, G. L., Cavallaro, C., Gu, H., Zhao, Z., and Tribulato, A. (2018). Diversity of Sicilian broccoli (*Brassica oleracea* var. italica) and cauliflower (*Brassica oleracea* var. botrytis) landraces and their distinctive bio-morphological, antioxidant, and genetic traits. *Genet. Resour. Crop Evol.* 65 (2), 485–502. doi: 10.1007/s10722-017-0547-8
- Brassica-dataset (2015). *USDA Agricultural Research Service, (2015)* (Germplasm Resources Information Network (GRIN)). USDA Agricultural Research Service).
- Card, S. D., Hume, D. E., Roodi, D., McGill, C. R., Millner, J. P., and Johnson, R. D. (2015). Beneficial endophytic microorganisms of *Brassica* – A review. *Biol. Control* 90, 102–112. doi: 10.1016/j.biocontrol.2015.06.001
- CGIAR (2022) CGIAR: Results dashboard - Genebanks. Available at: <https://www.cgiar.org/food-security-impact/results-dashboard/>.
- Chao, H., Li, T., Luo, C., Huang, H., Ruan, Y., Li, X., et al. (2020). BrassicaEDB: A gene expression database for *Brassica* crops. *Int. J. Mol. Sci.* 21 (16), 5831. doi: 10.3390/ijms21165831
- Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of *Brassica* plants. *Hortic. Res.* 1 (1), 14024. doi: 10.1038/hortres.2014.24
- Cornblatt, B. S., Ye, L., Dinkova-Kostova, A. T., Erb, M., Fahey, J. W., Singh, N. K., et al. (2007). Preclinical and clinical evaluation of sulforaphane for chemoprevention in the breast. *Carcinogenesis* 28 (7), 1485–1490. doi: 10.1093/carcin/bgm049
- Dhaliwal, M. S. (2017). *Handbook of vegetable crops 3rd Edn.* (New Delhi: Kalyani Publishers).
- Dixon, G. R. (2006). *Origins and diversity of Brassica and its relatives* (Wallingford: CABI), 1–33.
- El-Esawi, M. A. (2017). Genetic diversity and evolution of *Brassica* genetic resources: from morphology to novel genomic technologies – a review. *Plant Genet. Resour.* 15 (5), 388–399. doi: 10.1017/S1479262116000058
- Ellis, R. H., Nasehzadeh, M., Hanson, J., and Woldemariam, Y. (2018). Medium-term seed storage of 50 genera of forage legumes and evidence-based genebank monitoring intervals. *Genet. Resour. Crop Evol.* 65 (2), 607–623. doi: 10.1007/s10722-017-0558-5
- Everingham, S. E., Offord, C. A., Sabot, M. E. B., and Moles, A. T. (2021). Time-traveling seeds reveal that plant regeneration and growth traits are responding to climate change. *Ecology* 102 (3), e03272. doi: 10.1002/ecy.3272
- Fahey, J. W. (2016). "Brassica: characteristics and properties," in *Encyclopedia of Food and Health*. Eds. B. Caballero, P. M. Finglas and F. Toldrá (Oxford: Academic Press), 469–477.
- FAOSTAT (1997). *FAOSTAT statistical database* (Rome: FAO), c1997.
- Food and Agriculture Organization of the United Nations (2013). *Genebank standards for plant genetic resources for food and agriculture* (Rome: Food and Agriculture Organization of the United Nations).
- Fotev, Y. V., Artemyeva, A., Fateev, D., Bugrovskaya, G., Belousova, V., and Kukushkina, T. (2018). Results of SSR analysis, properties of plant morphology and biochemical composition of Chinese broccoli—a new vegetable crop for Russia. *Veg. Crops Russia* 1, 12–19. doi: 10.18619/2072-9146-2018-1-12-19
- GBIF (2020). *What is GBIF? The Global Biodiversity Information Facility*.
- Genesys (2017). *Genesys Global Portal on Plant Genetic Resources*. Crop Trust. <https://www.genesys-pgr.org/>.
- Gonzalez-Benito, M. E., Pérez-García, F., Tejeda, G., and Gomez-Campo, C. (2011). Effect of the gaseous environment and water content on seed viability of four *Brassicaceae* species after 36 years storage. *Seed Sci. Technol.* 39, 443–451. doi: 10.15258/sst.2011.39.2.16

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.2023.1220134/full#supplementary-material>

- Govaerts, R., Dransfield, J., Zona, S., Hodel, D. R., and Henderson, A. (2022). *World Checklist of Selected Plant Families* (Royal Botanic Gardens, Kew).
- Gupta, S. K. (2016). "Chapter 3 - Brassicas," in *Breeding Oilseed Crops for Sustainable Production*. Ed. S. K. Gupta (San Diego: Academic Press), 33–53.
- Hagos, R., Shaibu, A. S., Zhang, L., Cai, X., Liang, J., Wu, J., et al. (2020). Ethiopian mustard (*Brassica carinata* A. Braun) as an alternative energy source and sustainable crop. *Sustainability* 12 (18), 7492. doi: 10.3390/su12187492
- He, Z., Ji, R., Havlickova, L., Wang, L., Li, Y., Lee, H. T., et al. (2021). Genome structural evolution in *Brassica* crops. *Nat. Plants* 7 (6), 757–765. doi: 10.1038/s41477-021-00928-8
- Hong, S., Choi, S. R., Kim, J., Jeong, Y.-M., Kim, J.-S., Ahn, C.-H., et al. (2022). Identification of accession-specific variants and development of KASP markers for assessing the genetic makeup of *Brassica rapa* seeds. *BMC Genomics* 23 (1), 326. doi: 10.1186/s12864-022-08567-9
- ITIS (2022). *Intergrated Taxonomic Information System*. National Museum of Natural History, Smithsonian Institution. <https://www.itis.gov/>.
- Kameswara Rao, N., Dulloo, M. E., and Engels, J. M. M. (2017). A review of factors that influence the production of quality seed for long-term conservation in genebanks. *Genet. Resour. Crop Evol.* 64 (5), 1061–1074. doi: 10.1007/s10722-016-0425-9
- Kellingray, L., Tapp, H. S., Saha, S., Doleman, J. F., Narbad, A., and Mithen, R. F. (2017). Consumption of a diet rich in *Brassica* vegetables is associated with a reduced abundance of sulphate-reducing bacteria: A randomised crossover study. *Mol. Nutr. Food Res.* 61 (9). doi: 10.1002/mnfr.201600992
- Kiefer, M., Schmickl, R., German, D. A., Mandáková, T., Lysak, M. A., Al-Shehbaz, I. A., et al. (2013). BrassiBase: introduction to a novel knowledge database on *Brassicaceae* evolution. *Plant Cell Physiol.* 55 (1), e3–e3. doi: 10.1093/pcp/pct158
- Klopsch, R., Witzel, K., Börner, A., Schreiner, M., and Hanschen, F. S. (2017). Metabolic profiling of glucosinolates and their hydrolysis products in a germplasm collection of *Brassica rapa* turnips. *Food Res. Int.* 100, 392–403. doi: 10.1016/j.foodres.2017.04.016
- Koh, J. C. O., Barbulescu, D. M., Norton, S., Redden, B., Salisbury, P. A., Kaur, S., et al. (2017). A multiplex PCR for rapid identification of *Brassica* species in the triangle of U. *Plant Methods* 13 (1), 49. doi: 10.1186/s13007-017-0200-8
- Langridge, P., and Waugh, R. (2019). Harnessing the potential of germplasm collections. *Nat. Genet.* 51 (2), 200–201. doi: 10.1038/s41588-018-0340-4
- Lee, J. G., Bonnema, G., Zhang, N., Kwak, J. H., de Vos, R. C. H., and Beekwilder, J. (2013). Evaluation of glucosinolate variation in a collection of turnip (*Brassica rapa*) germplasm by the analysis of intact and desulfo glucosinolates. *J. Agric. Food Chem.* 61 (16), 3984–3993. doi: 10.1021/jf400890p
- Lee, M.-K., Chun, J.-H., Byeon, D. H., Chung, S.-O., Park, S. U., Park, S., et al. (2014). Variation of glucosinolates in 62 varieties of Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) and their antioxidant activity. *LWT - Food Sci. Technol.* 58 (1), 93–101. doi: 10.1016/j.lwt.2014.03.001
- Li, G., Yue, L., Cai, X., Li, F., Zhang, H., Zhang, S., et al. (2022). Fingerprint construction through genotyping by sequencing for applied breeding in *Brassica rapa*. *Genome* 65 (2), 105–113. doi: 10.1139/gen-2021-0021%M34648727
- Lin, Y.-r., Lee, J.-y., Tseng, M.-c., Lee, C.-y., Shen, C.-h., Wang, C.-s., et al. (2018). Subtropical adaptation of a temperate plant (*Brassica oleracea* var. *italica*) utilizes non-vernalization-responsive QTLs. *Sci. Rep.* 8 (1), 13609. doi: 10.1038/s41598-018-31987-1
- Liu, U., Cossu, T. A., Davies, R. M., Forest, F., Dickie, J. B., and Breman, E. (2020). Conserving orthodox seeds of globally threatened plants ex situ in the Millennium Seed Bank, Royal Botanic Gardens, Kew, UK: the status of seed collections. *Biodiver. Conserv.* 29 (9), 2901–2949. doi: 10.1007/s10531-020-02005-6
- Liu, B., Wang, Y., Zhai, W., Deng, J., Wang, H., Cui, Y., et al. (2013). Development of InDel markers for *Brassica rapa* based on whole-genome re-sequencing. *Theor. Appl. Genet.* 126 (1), 231–239. doi: 10.1007/s00122-012-1976-6
- Lotti, C., Iovieno, P., Centomani, I., Marcotrigiano, A. R., Fanelli, V., Mimola, G., et al. (2018). Genetic, bio-agronomic, and nutritional characterization of kale (*Brassica oleracea* L. var. *Acaphala*) diversity in Apulia, Southern Italy. *Diversity* 10 (2), 25. doi: 10.3390/d10020025
- Lusty, C., van Beem, J., and Hay, F. R. (2021). A performance management system for long-term germplasm conservation in CGIAR genebanks: aiming for quality, efficiency and improvement. *Plants* 10 (12), 2627. doi: 10.3390/plants10122627
- Malfa, G. A., De Leo, M., Tundis, R., Braca, A., Loizzo, M. R., Di Giacomo, C., et al. (2022). Biological Investigation and Chemical Study of *Brassica villosa* subsp. *drepanensis* (*Brassicaceae*) Leaves. *Molecules* 27 (23), 8447. doi: 10.3390/molecules27238447
- Meglic, V., and Pipan, B. (2018). *Spatial and Temporal Assessment of Brassica napus L. Maintaining Genetic Diversity and Gene Flow Potential: An Empirical Evaluation*. Intechopen Publishers. <https://www.intechopen.com/chapters/59678>.
- Mira, S., Estrelles, E., and González-Benito, M. E. (2015). Effect of water content and temperature on seed longevity of seven *Brassicaceae* species after 5 years of storage. *Plant Biol.* 17 (1), 153–162. doi: 10.1111/plb.12183
- Mira, S., Estrelles, E., González-Benito, M. E., and Corbier, F. (2011). Biochemical changes induced in seeds of *Brassicaceae* wild species during ageing. *Acta Physiologiae Plantarum* 33 (5), 1803–1809. doi: 10.1007/s11738-011-0719-7
- Mira, S., Nadarajan, J., Liu, U., González-Benito, M. E., and Pritchard, H. W. (2019). Lipid thermal fingerprints of long-term stored seeds of *Brassicaceae*. *Plants* 8 (10), 414. doi: 10.3390/plants8100414
- Montaut, S., Blažević, I., Rušćić, M., and Rollin, P. (2017). LC–MS profiling of glucosinolates in the seeds of *Brassica elongata* Ehrh., and of the two stenoendemic *B. botteri* Vis and *B. cazzae* Ginz. & Teyber. *Natural Product Res.* 31 (1), 58–62. doi: 10.1080/14786419.2016.1212032
- Nagel, M., and Börner, A. (2010). The longevity of crop seeds stored under ambient conditions. *Seed Sci. Res.* 20 (1), 1–12. doi: 10.1017/S0960258509990213
- OECD (2016). *Safety Assessment of Transgenic Organisms in the Environment*. Organisation for Economic Co-operation and Development, Vol. 5. <https://www.oecd.org/chemicalsafety/safety-assessment-of-transgenic-organisms-in-the-environment-volume-5-9789264253018-en.htm>.
- Oppermann, M., Weise, S., Dittmann, C., and Knüpfer, H. (2015). GBIS: the information system of the German Genebank. *Database* 2015. doi: 10.1093/database/bav021
- Panis, B., Nagel, M., and Van den houwe, I. (2020). Challenges and prospects for the conservation of crop genetic resources in field genebanks, in *in vitro* collections and/or in liquid nitrogen. *Plants* 9 (12), 1634. doi: 10.3390/plants9121634
- Pathirana, R., and Carimi, F. (2022). Management and utilization of plant genetic resources for a sustainable agriculture. *Plants* 11 (15), 2038. doi: 10.3390/plants11152038
- Pérez-García, F., Gómez-Campo, C., and Ellis, R. (2009). Successful long-term ultra dry storage of seed of 15 species of *Brassicaceae* in a genebank: variation in ability to germinate over 40 years and dormancy. *Seed Sci. Technol.* 37 (3), 640–649. doi: 10.15258/sst.2009.37.3.12
- Pérez-García, F., González-Benito, M. E., and Gómez-Campo, C. (2007). High viability recorded in ultra-dry seeds of 37 species of *Brassicaceae* after almost 40 years of storage. *Seed Sci. Technol.* 35, 143–153. doi: 10.15258/sst.2007.35.1.13
- Pino Del Carpio, D., Basnet, R. K., De Vos, R. C. H., Maliepaard, C., Paulo, M. J., and Bonnema, G. (2011). Comparative methods for association studies: A case study on metabolite variation in a *Brassica rapa* core collection. *PLoS One* 6 (5), e19624. doi: 10.1371/journal.pone.0019624
- Postman, J., Hummer, K., Ayala-Silva, T., Bretting, P., Franko, T., Kinard, G., et al. (2010). GRIN-Glob: An international project to develop a global plant genebank information management system. *Acta Hort.* 859, 49–55. doi: 10.17660/ActaHortic.2010.859.4
- POWO (2022). *Plants of the World Online*. Royal Botanic Gardens, Kew. <https://powo.science.kew.org/>.
- Quezada-Martinez, D., Addo Nyarko, C. P., Schiessl, S. V., and Mason, A. S. (2021). Using wild relatives and related species to build climate resilience in *Brassica* crops. *Theor. Appl. Genet.* 134 (6), 1711–1728. doi: 10.1007/s00122-021-03793-3
- Rakshita, K. N., Singh, S., Verma, V. K., Sharma, B. B., Saini, N., Iqbal, M. A., et al. (2021). Agro-morphological and molecular diversity in different maturity groups of Indian cauliflower (*Brassica oleracea* var. *botrytis* L.). *PLoS One* 16 (12), e0260246. doi: 10.1371/journal.pone.0260246
- Ramirez, D., Abellán-Victorio, A., Beretta, V., Camargo, A., and Moreno, D. A. (2020). Functional ingredients from *Brassicaceae* species: overview and perspectives. *Int. J. Mol. Sci.* 21 (6), 1998. doi: 10.3390/ijms21061998
- Razzaq, H., Armstrong, S. J., and Saleem, H. (2021). "Chapter 10 - Brassicas: A complete guide to the potential of their wild relatives," in *Wild Germplasm for Genetic Improvement in Crop Plants*. Eds. M. T. Azhar and S. H. Wani (Academic Press), 187–199.
- Reed, B. M., Engelmann, F., Dulloo, E., and Engels, J. (2004). *Technical guidelines for the management of field and in vitro germplasm collections*. Handbooks for Genebanks. n.7, 106 p. International Plant Genetic Resources Institute, 95. <https://hdl.handle.net/10568/105045>.
- Salehi, B., Quispe, C., Butnariu, M., Sarac, I., Marmouzi, I., Kamle, M., et al. (2021). Phytotherapy and food applications from *Brassica* genus. *Phytother. Res.* 35 (7), 3590–3609. doi: 10.1002/ptr.7048
- Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Res.* 47 (D1), D94–D99. doi: 10.1093/nar/gky989
- Schoch, C. L., Cufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020. doi: 10.1093/database/baaa062
- Scialabba, A., Giorgetti, L., and Bellani, L. M. (2016). Stress integrated tests and cytological analyses reveal *Brassica villosa* subsp. *drepanensis* seed quality decrease upon long-term storage. *Plant Biosyst. - Int. J. Dealing all Aspects Plant Biol.* 150 (4), 757–766. doi: 10.1080/11263504.2014.991359
- Shekhawat, K., Rathore, S. S., Premi, O. P., Kandpal, B. K., and Chauhan, J. S. (2012). Advances in agronomic management of indian mustard (*Brassica juncea* (L.) Czernj. Cosson): an overview. *Int. J. Agron.* 2012, 408284. doi: 10.1155/2012/408284
- Sheng, X.-G., Branca, F., Zhao, Z.-Q., Wang, J.-S., Yu, H.-F., Shen, Y.-S., et al. (2020). Identification of black rot resistance in a wild *Brassica* species and its potential transferability to cauliflower. *Agronomy* 10 (9), 1400. doi: 10.3390/agronomy10091400

- Shuhang, W., Voorrips, R. E., Steenhuis-Broers, G., Vosman, B., and van Loon, J. J. A. (2016). Antibiosis resistance against larval cabbage root fly, *Delia radicum*, in wild *Brassica*-species. *Euphytica* 211 (2), 139–155. doi: 10.1007/s10681-016-1724-0
- Soh, E. H., Lee, W. M., Park, K. W., Choi, K.-J., and Yoon, M. K. (2014). Change of germination rate for chili pepper and chinese cabbage seed in relation to packaging materials and storage conditions over 10 years. *Hortic. Sci. Technol.* 32 (6), 864–871. doi: 10.7235/hort.2014.14091
- Solberg, S. Ø., Yndgaard, F., Andreassen, C., von Bothmer, R., Loskutov, I. G., and Asdal, Å. (2020). Long-term storage and longevity of orthodox seeds: A systematic review. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01007
- Solov'yeva, A. E., Artem'yeva, A. M., and Schuetze, W. (2013). Peculiar properties of glucosinolate accumulation in the *Brassicaceae* family. *Russian Agric. Sci.* 39 (5), 419–422. doi: 10.3103/S1068367413050169
- Stewart, A. (2002). A review of *Brassica* species, cross-pollination and implications for pure seed production in New Zealand. *Agronomy New Zealand* 32/33, 63–82.
- Szczygłowska, M., Piekarska, A., Konieczka, P., and Namieśnik, J. (2011). Use of *Brassica* plants in the phytoremediation and biofumigation processes. *Int. J. Mol. Sci.* 12 (11), 7760–7771. doi: 10.3390/ijms12117760
- Taylor, A., Rana, K., Handy, C., and Clarkson, J. P. (2018). Resistance to *Sclerotinia sclerotiorum* in wild *Brassica* species and the importance of *Sclerotinia subarctica* as a *Brassica* pathogen. *Plant Pathol.* 67 (2), 433–444. doi: 10.1111/ppa.12745
- Thorwarth, P., Yousef, E. A., and Schmid, K. J. (2018). Genomic prediction and association mapping of curd-related traits in gene bank accessions of cauliflower. *G3: Genes Genomes Genet.* 8 (2), 707–718. doi: 10.1534/g3.117.300199
- Tortosa, M., Velasco, P., Afonso, D., Padilla, G., Ríos, D., and Soengas, P. (2017). Characterization of a Spanish *Brassica oleracea* collection by using molecular and biochemical markers. *Scientia Hort.* 219, 344–350. doi: 10.1016/j.scienta.2017.03.021
- U, N. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese J. Bot.* 7, 389–452.
- van Treuren, R., Bas, N., Kodde, J., Groot, S. P. C., and Kik, C. (2018). Rapid loss of seed viability in *ex situ* conserved wheat and barley at 4°C as compared to –20°C storage. *Conserv. Physiol.* 6 (1). doi: 10.1093/conphys/coy033
- Volk, G. M., Namuth-Covert, D., and Byrne, P. F. (2019). Training in plant genetic resources management: A way forward. *Crop Sci.* 59 (3), 853–857. doi: 10.2135/cropsci2018.11.0689
- Walters, C., Wheeler, L., and Stanwood, P. C. (2004). Longevity of cryogenically stored seeds. *Cryobiology* 48 (3), 229–244. doi: 10.1016/j.cryobiol.2004.01.007
- Weise, S., Lohwasser, U., and Oppermann, M. (2020). Document or lose it—On the importance of information management for genetic resources conservation in genebanks. *Plants* 9 (8), 1050. doi: 10.3390/plants9081050
- Weise, S., Oppermann, M., Maggioni, L., van Hintum, T., and Knüpfner, H. (2017). EURISCO: The European search catalogue for plant genetic resources. *Nucleic Acids Res.* 45 (D1), D1003–D1008. doi: 10.1093/nar/gkw755
- WIEWS (2022). *WIEWS - World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture* (Food and Agriculture Organisation of the United Nations).
- Wijaya, H., Rouw, R., and Kadir, A. (2020). *Brassica* box food products as a healthy local food innovation in The Covid-19 pandemic period. *IOP Conf. Series: Earth Environ. Sci.* 575, 12011. doi: 10.1088/1755-1315/575/1/012011
- Witzel, K., Kurina, A. B., and Artemyeva, A. M. (2021). Opening the Treasure Chest: The Current Status of Research on *Brassica oleracea* and *B. rapa* Vegetables From *ex situ* Germplasm Collections. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.643047
- Yang, J., Zhang, C., Zhao, N., Zhang, L., Hu, Z., Chen, S., et al. (2018). Chinese Root-type Mustard Provides Phylogenomic Insights into the Evolution of the Multi-use Diversified Allopolyploid *Brassica juncea*. *Mol. Plant* 11 (3), 512–514. doi: 10.1016/j.molp.2017.11.007



OPEN ACCESS

EDITED BY

Maarten Van Zonneveld,
World Vegetable Center, Taiwan

REVIEWED BY

Parthiban Subramanian,
National Agrobiodiversity Center,
Republic of Korea
Rodrigo R. Amadeu,
Bayer Crop Science, United States

*CORRESPONDENCE

Eder Jorge de Oliveira
✉ eder.oliveira@embrapa.br

RECEIVED 29 June 2023

ACCEPTED 21 August 2023

PUBLISHED 06 September 2023

CITATION

Santos CCd, Andrade LRBd, Carmo CDd
and Oliveira EJd (2023) Development of
cassava core collections based on
morphological and agronomic traits
and SNPS markers.
Front. Plant Sci. 14:1250205.
doi: 10.3389/fpls.2023.1250205

COPYRIGHT

© 2023 Santos, Andrade, Carmo and
Oliveira. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). 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.

Development of cassava core collections based on morphological and agronomic traits and SNPS markers

Caroline Cardoso dos Santos ¹,
Luciano Rogerio Braatz de Andrade ²,
Cátia Dias do Carmo ¹ and Eder Jorge de Oliveira ^{2*}

¹Centro de Ciências Agrárias, Ambientais e Biológicas, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, Bahia, Brazil, ²Embrapa Mandioca e Fruticultura, Nubem, Cruz das Almas, Bahia, Brazil

Cassava (*Manihot esculenta* Crantz) holds significant importance as one of the world's key starchy crop species. This study aimed to develop core collections by utilizing both phenotypic data (15 quantitative and 33 qualitative descriptors) and genotypic data (20,023 single-nucleotide polymorphisms) obtained from 1,486 cassava accessions. Six core collections were derived through two optimization strategies based on genetic distances: Average accession-to-nearest-entry and Average entry-to-nearest-entry, along with combinations of phenotypic and genotypic data. The quality of the core collections was evaluated by assessing genetic parameters such as genetic diversity Shannon-Weaver Index, inbreeding (*Fis*), observed (*Ho*), and expected (*Hs*) heterozygosity. While the selection of accessions varied among the six core collections, a seventh collection (consolidated collection) was developed, comprising accessions selected by at least two core collections. Most collections exhibited genetic parameters similar to the complete collection, except for those developed by the Average accession-to-nearest-entry algorithm. However, the variations in the maximum and minimum values of *Ho*, *Hs*, and *Fis* parameters closely resembled the complete collection. The consolidated collection and the collection constructed using genotypic data and the Average entry-to-nearest-entry algorithm (GenEN) retained the highest number of alleles (>97%). Although the differences were not statistically significant (above 5%), the consolidated collection demonstrated a distribution profile and mean trait values most similar to the complete collection, with a few exceptions. The Shannon-Weaver Index of qualitative traits exhibited variations exceeding $\pm 10\%$ when compared to the complete collection. Principal component analysis revealed that the consolidated collection selected cassava accessions with a more uniform dispersion in all four quadrants compared to the other core collections. These findings highlight the development of optimized and valuable core collections for efficient breeding programs and genomic association studies.

KEYWORDS

Manihot esculenta Crantz, germplasm, breeding, genetic resources, SNP

1 Introduction

Cassava (*Manihot esculenta* Crantz) plays a pivotal role in ensuring global food security as it is a staple food consumed by thousands of people in countries across Africa, Asia, and Latin America (Lebot, 2009). This economically important crop possesses genetic variability. The formal documentation of cassava breeding and *ex situ* conservation of plant genetic resources began in the mid-1930s at the Instituto Agronômico de Campinas in São Paulo (Fukuda et al., 2002). Presently, numerous cassava germplasm collections exist worldwide, with the objective of documenting, evaluating, preserving, and making available the existing genetic diversity of the species for breeding programs. In Brazil, the Empresa Brasileira de Pesquisa e Agropecuária (EMBRAPA) maintains approximately 4,000 *ex situ* conserved cassava accessions in the field and *in vitro* (Hershey, 2017).

The management of these databases is a challenge due to the large number of accessions and the high maintenance cost (Van Hintum et al., 2000). The future of cassava breeding faces potential challenges that could jeopardize its progress, stemming from various factors, including: i) under-representation of the diversity present in certain biomes, ii) the existence of accession duplicates, iii) limited use of germplasm by end users, iv) insufficient regeneration of preserved materials, v) incomplete morpho-agronomic characterization, vi) low investment in the collection and maintenance of resources genetic factors (Diez et al., 2018). Furthermore, the growing demand for the conservation and preservation of genetic materials and the lack of fast methodologies for verifying the existence or not of additional diversity of new germplasms, cause the continuous growth of collections.

Efficient access to genetic variability is crucial for the genetic improvement of successful plants, and large germplasm collections are valuable for preserving genetic diversity (Frankel and Bennett, 1970). However, there is a significant risk that the usefulness and accessibility of these collections will decrease as their size increases (Frankel et al., 1981). Consequently, breeding programs often utilize only a fraction of the available genetic diversity. Germplasm banks play a vital role in storing the genetic variation necessary for continual enhancements in productivity, stress resistance, and nutritional quality through breeding programs (Wang et al., 2017). Nevertheless, cassava germplasm collections consist of thousands of genotypes, including numerous duplicate accessions, with limited characterization and understanding of their potential as parents or for direct use in commercial production systems. This limitation has impeded the utilization of cassava genetic resources to unlock the crop's full productivity potential and address challenges arising from global climate change.

To address these challenges, various strategies have been proposed to enhance the management of large collections, such as the creation of core collections. Core collections involve selecting representative sets of samples that capture the genetic variability of the entire collection while minimizing redundancy (Brown and Spillane, 1999). Due to their smaller size, core collections can undergo comprehensive phenotyping for key descriptors that define potential applications in species improvement. By

characterizing and evaluating a small portion of accessions in detail, core collections can effectively represent the morpho-agronomic and molecular diversity of the complete collection. This approach can encourage researchers and producers to incorporate new germplasms into breeding programs and even directly into production systems (Boczkowska et al., 2016).

Core collections are crucial in maximizing diversity and minimizing duplication within the complete collection, leading to improved management and efficiency in the conservation and utilization of genetic resources for a particular species. In the context of cassava, the formation of core collections was first documented by the International Center for Tropical Agriculture (CIAT) (Hershey, 1994). Bhattacharjee et al. (2012) used 40 morpho-agronomic traits evaluated in two different locations, selecting 428 accessions that captured 90% of the total variation to compose the core collection of the International Institute of Tropical Agriculture (IITA). Similarly, Oliveira et al. (2014) employed 354 single-nucleotide polymorphism (SNP) markers to create core collections with varying numbers of accessions. However, there are limited reports on the formation of cassava core collections that encompass broader genomic coverage and integration of phenotypic and genotypic information.

While phenotypic variation plays a crucial role in practical selection within genetic improvement programs, the establishment of cassava core collections based on a substantial number of morphological and agronomic descriptors evaluated over multiple cultivation years, along with a large set of molecular markers, has not yet been proposed. Therefore, the objectives of this study are as follows: i) develop cassava core collections based on quantitative, qualitative, and molecular data descriptors, both individually and in combination; ii) assess the effectiveness of different selection methods for cassava core collections in retaining maximum genetic diversity, variance, and other genetic parameters compared to the complete collection; and iii) generate a consolidated core collection that represents the highest phenotypic and molecular variability among cassava genotypes.

2 Materials and methods

2.1 Plant material

A comprehensive evaluation was conducted on a total of 1,486 accessions from the cassava germplasm bank of Embrapa Mandioca e Fruticultura. These accessions originate from various regions in Brazil and some others acquired through exchanges with countries such as Colombia, Venezuela, Nigeria, Mexico and Uganda (Table S1). The collection encompasses both local and improved varieties, which have been obtained through breeding techniques such as crossings, mass selection, and identification by producers or research institutions.

The evaluation and characterization of the germplasm bank accessions took place between 2011 and 2021 in three cities in the State of Bahia, Brazil: Cruz das Almas, Laje, and Valença (Table S2), according to Fukuda et al. (2010). The climate in this region is classified as type Af according to the Köppen classification, characterized as tropical with an average annual temperature of 24.2°C, approximately

80% humidity, and an average annual precipitation of 1,300 mm. The wettest months typically occur from March to July, while October and January are considered the driest periods. Detailed information regarding soil type, geographic coordinates, and evaluation years for each location can be found in [Table S2](#).

2.2 Morpho-agronomic descriptors

The characterization of the cassava accessions involved the use of standardized scales for morpho-agronomic descriptors, which encompassed various aspects of the plant including leaf, stem, root, flower, and agronomic traits. The descriptors were categorized into qualitative and quantitative variables specific to the cassava crop, as outlined in [Table S3](#). The characterization process followed the guidelines established by [Fukuda et al. \(2010\)](#), [Bradbury et al. \(1999\)](#), and [Kawano et al. \(1987\)](#).

2.3 Genotyping of the cassava accessions

DNA extraction from young cassava leaves was carried out using the CTAB method (cetyltrimethylammonium bromide), following the protocol described by [Doyle and Doyle \(1990\)](#) with certain modifications. These modifications included increasing the concentration of 2-mercaptoethanol to 0.4% and incorporating polyvinylpyrrolidone (PVP). The quality of the extracted DNA was assessed by running the samples on a 1% agarose gel stained with ethidium bromide (1.0mg/L).

For genotyping, the DNA samples were sent to Cornell University's Genomic Diversity Facility, where the Genotyping-by-Sequencing (GBS) protocol ([Elshire et al. 2011](#)) was employed. Initially, the samples were digested using the restriction enzyme *ApeKI*, a type II restriction endonuclease that recognizes a degenerate sequence of 5 bases (GCWGC, where W represents A or T) with fragment lengths of 100 bp ([Hamblin and Rabbi, 2014](#)). After digestion and ligation of the *ApeKI* cleavage fragments with adapters, sequencing was performed in a multiplex system with 192 samples. The Genome Analyzer 2000 genotyping platform (Illumina, Inc., San Diego, CA) was used for the sequencing process. The obtained reads were aligned to the cassava v.6 reference genome ([Bredeson et al., 2016](#)) using the BWA software ([Li and Durbin, 2009](#)).

A total of 20,023 SNP markers were obtained, and these markers are distributed across the 18 cassava chromosomes. Sequence analysis and quality filtering were performed using Tassel software version 5.2.37 ([Bradbury et al., 2007](#)). The filtering steps involved removing markers with a minimum allele frequency (MAF) and a high rate of missing data (Call Rate) below 5% and above 80%, respectively. Any remaining missing data were subsequently imputed using the Beagle 4.1 software ([Browning and Browning, 2016](#)).

2.4 Data analysis

The accessions in the cassava germplasm bank underwent evaluations in various trials conducted from 2011 to 2021, leading

to minor variations in the qualitative descriptors across different classes. To address this variability, we employed the mode as an indicator of the prevailing trend for accessions in terms of qualitative characteristics. The mode signifies the class that exhibited the highest frequency of observations recorded over multiple years.

The quantitative dataset was analyzed using linear mixed models. The dataset included information about the year and location of assessment, referred to as “environments” in this context. A analysis considering all environments for each quantitative descriptor was performed using the following statistical model: $y = Zg + Wb + Ti + e$, where y is the vector of phenotypic observations, g represents the genotypic effects considered as random effect $g \sim N(0, \sigma_g^2)$; b is the aligned effects of blocks within trials considered as random $b \sim N(0, \sigma_b^2)$; i represents the effects of the genotype-trial interaction considered as random effect $i \sim N(0, \sigma_i^2)$; and e represents the error effects considered as random effect $e \sim N(0, \sigma_e^2)$. Z , W , and T are the incidence matrices for the corresponding effects. This model was used to estimate the genetic values of the genotypes based on the evaluation of experiments conducted under an incomplete block design across multiple trials. The mixed linear model analyses were performed using the sommer package version 4.1.8 ([Covarrubias-Pazarán, 2016](#)) within the R version 4.3.0 environment ([R Development Core Team, 2023](#)).

2.5 Development of the cassava core collections

Six core collections were generated using different criteria based on distances between the accessions of the complete collection and the core collection ([Table 1](#)). The core collections were developed by using the stochastic parallel tempering algorithm ([Thachuk et al., 2009](#)) in the Core Hunter 3 package version 3.2.2 ([Beukelaer and Davenport, 2018](#)) in R version 4.3.0 ([R Development Core Team, 2023](#)), selecting 10% of the accessions relative to the size of the complete collection. Two strategies based on genetic distances were employed to optimize the collections: AN (Average accession-to-nearest-entry) and EN (Average entry-to-nearest-entry) as described by [Odong et al. \(2013\)](#). Under the AN criterion, the average distance between the accessions of the complete collection and their closest entry in the core collection is calculated. For the EN criterion, the objective is to maximize the average distance between each entry and its nearest neighboring entry in the complete collection, ensuring that each entry is as distinct as possible from the others. Different datasets were used for each optimization method, including phenotypic data, genotypic data, and a combination of both (phenotypic + genotypic data), each with their respective related distances.

In the case of collections based solely on phenotypic data, the Gower dissimilarity matrix ([Gower, 1971](#)) was employed as a criterion to define the core collections. The Gower matrix allows for the combined analysis of numerical and categorical variables, encompassing both quantitative and qualitative data from the evaluated descriptors. For each variable (j), a similarity coefficient

TABLE 1 Summary of strategies used to obtain cassava core collections based on 20,023 SNPs markers and 48 morpho-agronomic descriptors, divided into qualitative (33) and quantitative (15).

Core collection	Entry data	Algorithm	Genetic distance
GenAN	SNPs	AN	Modified Rogers
GenEN		EN	
PhenAN	morpho-agronomic	AN	Gower
PhenEN		AN	
GPmAN	SNPs + morpho-agronomic	AN	Czekanowski (Manhattan) + Gower
GPmEN		EN	

GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively.

(s_j) within the range of [0,1] is considered. The similarity between elements (l and k) is then calculated as follows: $d(l, k) = (\frac{\sum_{j=1}^{p+q} l_j(l, k) s_j(l, k)}{\sum_{j=1}^{p+q} l_j(l, k)})$, where $l_j(l, k)$ is a variable that equals 1 if l and k can be compared with variable X_j .

For collections obtained solely using molecular marker data, the Modified Rogers distance (Wright, 1978) was utilized. The formula for this distance is given as $(MR_{ij} = \frac{1}{2L} \sqrt{\sum_{l=1}^L \sum_{a=1}^2 (p_{ila} - p_{jla})^2})$, where L represents the total number of markers, p_{ila} is the allele frequency of allele a of marker l for accession i , p_{jla} is the allele frequency of allele a of marker l for accession j , and m_l denotes the number of matching alleles for marker l .

To form the pooled core collection, both genotypic and phenotypic data were utilized to generate distance matrices. Gower distance was employed for phenotypic data, while Manhattan distance was applied for genotypic data. The Czekanowski distance (calculated using the Manhattan formula) is given by the equation: $d_{cz}(A, B) = \frac{1}{2L} \sum_{i=1}^L |x_i - y_i|$, where x_i and y_i are the allele frequencies at locus i for individuals A and B , respectively; L denotes the number of loci for which x_i and y_i are available. The implementation of the Czekanowski distance utilized the dartR package version 2.7.2 (Mijangos et al., 2022) in R version 4.3.0 (R Development Core Team, 2023).

2.6 Assessing diversity: analysis, comparison, and validation of methods for cassava core collections

The coincidence between the different methods of forming cassava core collections was assessed using the Kappa index (Cohen, 1960). A binary code was employed to represent selected and unselected individuals, where selected individuals were assigned a code of 1 and unselected individuals a code of 0. The coincidence of accessions between collections was then analyzed based on this binary representation.

To evaluate the genetic diversity within core collections, consolidated collection, and complete collection, several parameters were considered. The observed heterozygosity (Ho) was calculated using the formula $Ho = 1 - \sum_k \sum_i P_{kii}/np$, where P_{kii} represents the proportion of homozygote i in sample k and np the number of samples. The genetic diversity within the population

(Hs) was determined using the formula $Hs = \bar{n}/(\bar{n} - 1)[1 - \sum \bar{p}_i^2 - Ho/2\bar{n}]$, where $\bar{n} = np/\sum_k 1/n_k$ and $\bar{p}_i^2 = \sum P_{ki}^2/np$. The inbreeding coefficient (Fis) was calculated as $Fis = 1 - Ho/Hs$. These calculations were performed using the hierfstat package version 0.5.11 (Goudet, 2005) in R version 4.3.0 (R Development Core Team, 2023).

The comparison between different core collections, the consolidated collection, and the complete collection for phenotypic data was conducted by analyzing the dispersion of quantitative and qualitative traits. The Shannon-Weaver diversity indices were calculated for each trait in the complete collection and individual collections using the formula $H' = -\sum_{i=1}^n p_i \log_e p_i$, where p_i represents the observed frequency of class i for trait n , n is the number of phenotypic classes. All H' indices were normalized and divided by the maximum value ($\log_e n$) to ensure that the values ranged from 0 to 1, representing monomorphism to maximum phenotypic diversity. For qualitative characteristics, k denoted the number of classes or grades of the descriptor, while for quantitative characters, six classes were estimated based on the lower and upper limits observed in the complete collection for each trait (Table S4). These analyses were performed using R version 4.3.0 (R Development Core Team, 2023).

The structure of the core and consolidated collections was assessed in comparison to the complete collection using principal component analysis (PCA). Morpho-agronomic data was analyzed using the AMR package version 2.0.0 (Berends et al., 2022), while molecular data underwent PCA using the PCAtools package version 2.12.0 (Blighe and Lun, 2023) in R version 4.3.0 (R Development Core Team, 2023).

3 Results

3.1 Concordance in genotype selection of core collections using phenotypic, genotypic, and pooled data

The core collections were created by selecting 10% of the complete collection, resulting in 149 genotypes. Overall, there was a lack of significant overlap in the selected accessions among the core collections. The core collections formed using the EN algorithm, based on phenotypic data (PhenEN), and pooled data

TABLE 2 Kappa index considering different methodologies for forming core collections.

Kappa index	GenAN	GenEN	PhenAN	PhenEN	GPmAN	GPmEN
GenAN	1	-0.07	0.02	0.05	0.11	0.01
GenEN	-0.07	1	-0.04	0.04	-0.05	0.09
PhenAN	0.02	-0.04	1	-0.08	0.19	-0.09
PhenEN	0.05	0.04	-0.08	1	-0.08	0.48
GPmAN	0.11	-0.05	0.19	-0.08	1	-0.09
GPmEN	0.01	0.09	-0.09	0.48	-0.09	1

GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively.

markers (Table 3). In general, the majority of collections exhibited comparable genetic parameters to the complete collection, with the exception of those utilizing the AN algorithm, which maintained the same *Ho* value (0.403). Moreover, the variations in the maximum and minimum values for the three parameters (*Ho*, *Hs*, and *Fis*) closely resembled those of the complete collection. Notably, the core collections CCons and GenEN were able to preserve a substantial proportion (>97%) of the total number of alleles present in the complete collections.

The distribution of genetic parameters in the core collections exhibited patterns that were largely comparable to those observed in the complete collection (Figure 2). Notably, prominent similarities were identified between the consolidated collection and the complete collection in terms of the *Ho* parameter. Similarly, the GenAN and PhenAN collections displayed noticeable resemblances to the complete collection in relation to the *Hs* parameter.

3.4 Variation in morpho-agronomic descriptors from different core collections

The interquartile ranges of phenotypic traits showed variations among the core collections, although the means of most traits were

similar to those of the complete collection. However, some specific traits, such as length and width ratio of leaf lobes, cyanide content, thickness of the root cortex, root diameter, dry matter content, plant height, and harvest index, exhibited slight variations when compared to the complete collection (Figure 3). Among the core collections, the consolidated collection displayed a distribution profile and average characteristics that were most similar to the complete collection, with the exception of cyanide content in the roots, number of roots, and harvest index.

The core collections formed based on phenotypic data, whether used alone or in combination with genotypic data, exhibited minimum and maximum values of quantitative traits that were very similar to those of the complete collection. In contrast, collections based solely on genotypic data, such as the GenAN collection, showed greater variation in the mean and range of phenotypic data, particularly for traits related to leaf lobes (e.g., length of leaf lobe, width of leaf lobe, length and width ratio of leaf lobes) and petiole length. Accessions with extreme values or low harvest index were not included in the GenAN core collection.

For the majority of quantitative phenotypic traits, there was no significant difference (>5%) in means and variances between the core collections and the complete collection (Table S6). However, some variation was observed in the means of these traits. The GenAN

TABLE 3 Basic genetic diversity parameters calculated for the core collections formed using different approaches based on 20,023 SNP markers.

Collections	<i>Ho</i>		<i>Hs</i>		<i>Fis</i>		Total number of alleles
	Mean	Range	Mean	Range	Mean	Range	
Complete	0.403	(0.04 – 1.00)	0.301	(0.04 – 0.62)	-0.228	(-1.00/-0.01)	58,672
CCons	0.396	(0.01 – 1.00)	0.299	(0.01 – 0.62)	-0.220	(0.00/-1.00)	56,972
GenAN*	0.403	(0.02 – 1.00)	0.302	(0.02 – 0.62)	-0.226	(0.00/-1.00)	55,976
GenEN	0.390	(0.01 – 1.00)	0.296	(0.01 – 0.62)	-0.213	(0.00/-1.00)	57,338
PhenAN	0.403	(0.01 – 1.00)	0.302	(0.01 – 0.62)	-0.227	(0.00/-1.00)	55,464
PhenEN	0.396	(0.01 – 1.00)	0.298	(0.01 – 0.62)	-0.221	(0.00/-1.00)	56,021
GPmAN	0.403	(0.02 – 1.00)	0.302	(0.02 – 0.62)	-0.226	(-1.00/-0.01)	55,623
GPmEN	0.398	(0.01 – 1.00)	0.299	(0.01 – 0.62)	-0.220	(0.00/-1.00)	56,806

*GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches. The genetic diversity parameters assessed were *Ho* (observed heterozygosity), *Hs* (genetic diversity within population), and *Fis* (inbreeding coefficient).

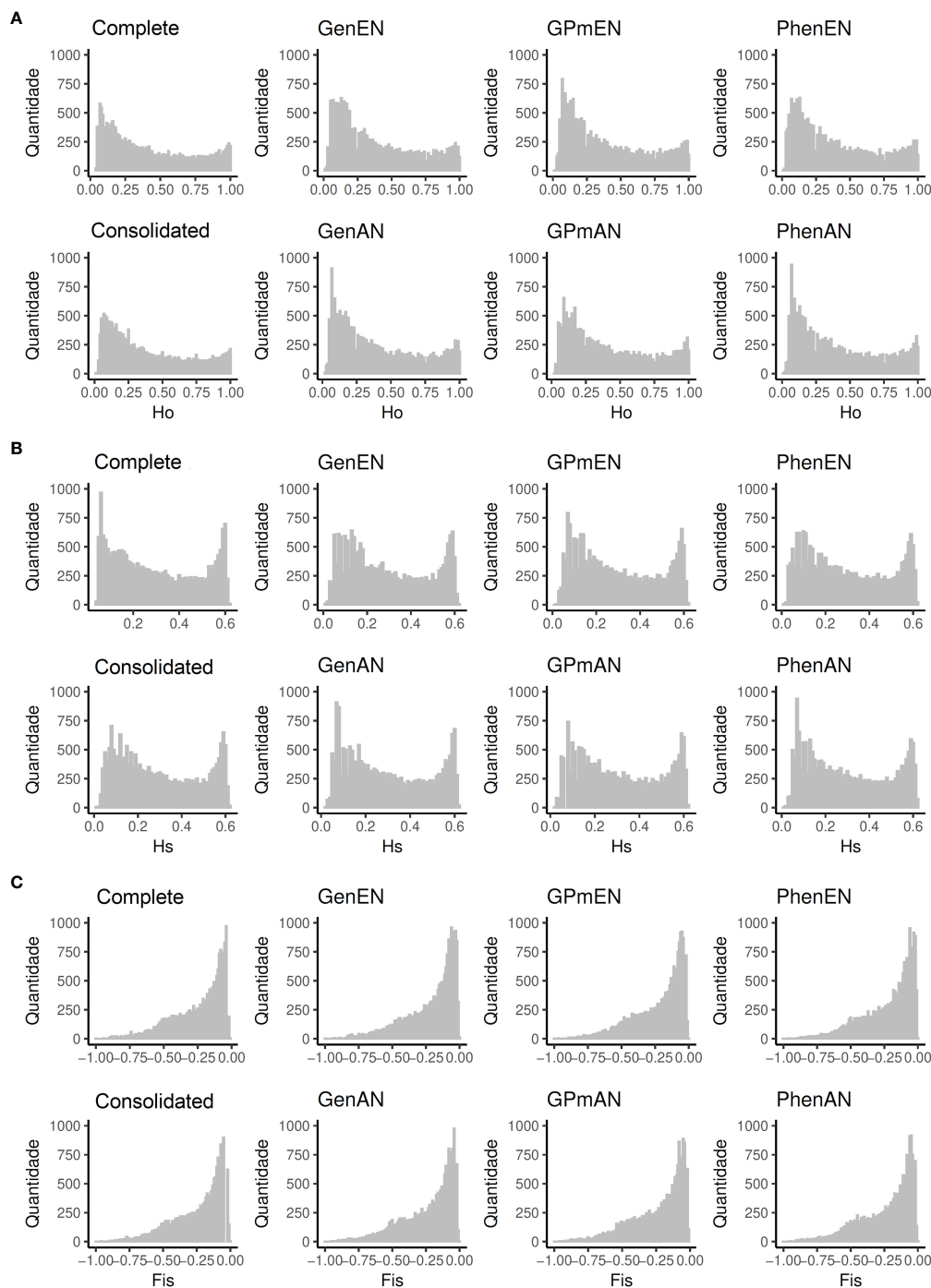


FIGURE 2

Genetic parameters analyzed for the complete collection (Complete), consolidated collection (CCons), and core collections derived from different data types. GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches. The genetic parameters evaluated included: (A) observed heterozygosity (H_o), (B) expected heterozygosity (H_s), and (C) inbreeding coefficient (F_{is}).

collection showed higher means compared to the complete collection for most traits, except for the length and width ratio of leaf lobes and thickness of the root cortex. Variance was higher than the complete collection (>50%) for certain traits, such as the root diameter, where the GPmEN (125.90) and PhenEN (135.00) collections exhibited

considerably higher variances than the complete collection (79.31). The harvest index also showed higher variances than the complete collection (45.95) in the GPmEN (67.00) and PhenEN (66.19) collections. In contrast, the PhenAN collection displayed lower variances than the complete collection for all traits.

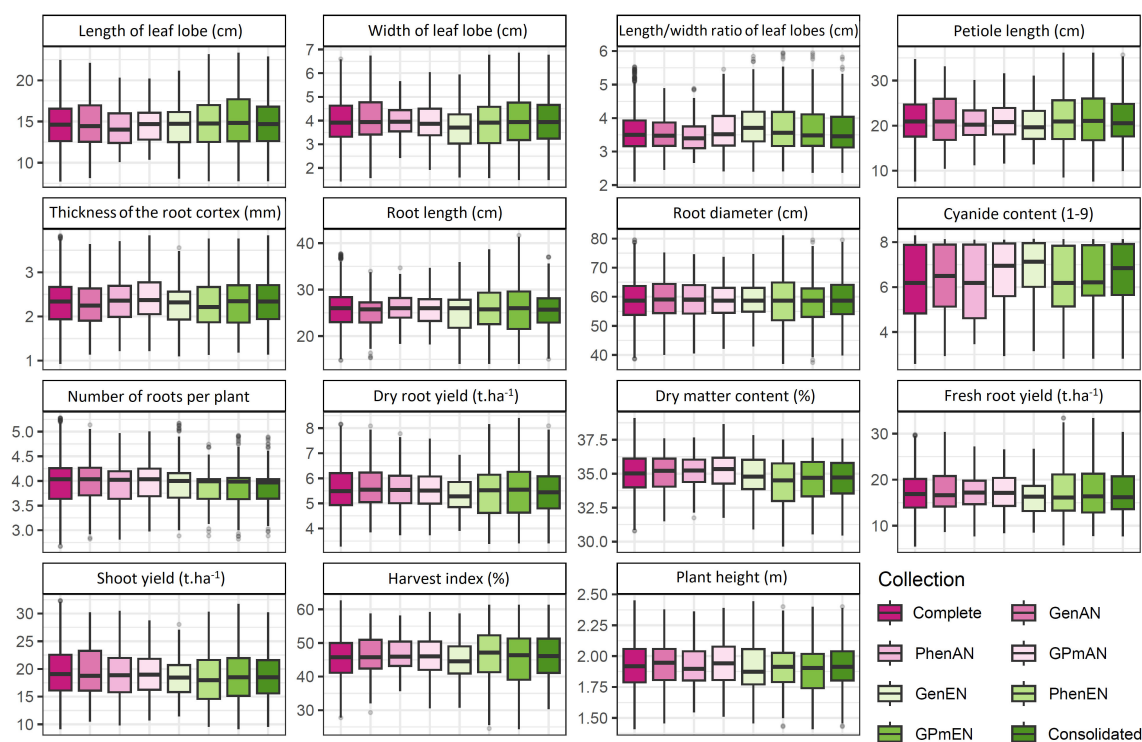


FIGURE 3

Comparative boxplot analysis of quantitative cassava descriptors in different core collections. GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches.

The qualitative data were analyzed based on the efficiency of the core collections in encompassing all classes of each evaluated trait (Figure 4). The core collections that showed a better balance in representing the classes were CCons, GPmEN, and PhenEN, especially for the traits of color of leaf vein, number of lobes, petiole position, root pulp color, root position, stipule margin, growth habit of stem, branching angle, and external stem skin color.

Some characteristics were not well represented in all core collections, possibly because some of the classes are rare occurrences, being seldom observed in the field. For example, only 28 cassava accessions (~1.88% of the complete collection) exhibit a zig-zag stem growth habit. Therefore, any variation in the method of forming the core collection can alter this frequency, as seen in the PhenAN collection, which was represented only by the straight growth habit. Other significant variations in the representativeness of the core collections based on qualitative data were identified for the stipule margin and root position traits, also due to the low frequency of certain classes in the complete collection.

3.5 Analysis of the phenotypic diversity of core collections

The quality assessment of core collections was conducted using the Shannon-Weaver Index (ISW). Comparisons were made between

the core collections and the complete collection, with variations greater than $\pm 10\%$ of the ISW considered significant for quantitative phenotypic traits (Table 4). Among the core collections, only the thickness of the root cortex showed a significant impact on the ISW, reaching 0.00 in the GenAN, PhenAN, GPmAN, and CCons collections, while being higher than the complete collection in the GenEN, PhenEN, and GPmEN collections. For other traits, the GenAN collection exhibited the smallest difference in ISW compared to the complete collection, while the PhenAN collection had the highest number of traits with lower ISW than the complete collection (length of leaf lobe, length and width ratio of leaf lobes, petiole length, root length, and harvest index). On the other hand, the PhenEN and GPmEN collections had a greater number of traits with higher ISW than the complete collection (length and width ratio of leaf lobes, petiole length, thickness of the root cortex, root length, root diameter, and harvest index). The consolidated collection showed minor differences in ISW for most traits, except for the length and width ratio of leaf lobes, root length and diameter, dry root yield, and harvest index, where the ISW differences exceeded 5%.

The ISW for qualitative traits exhibited variations greater than $\pm 10\%$ when compared to the complete collection, particularly in the GenEN, PhenAN, PhenEN, GPmAN, GPmEN, and CCons collections (Table 5). Similar to the quantitative traits, the GenAN collection demonstrated the lowest ISW variation for qualitative

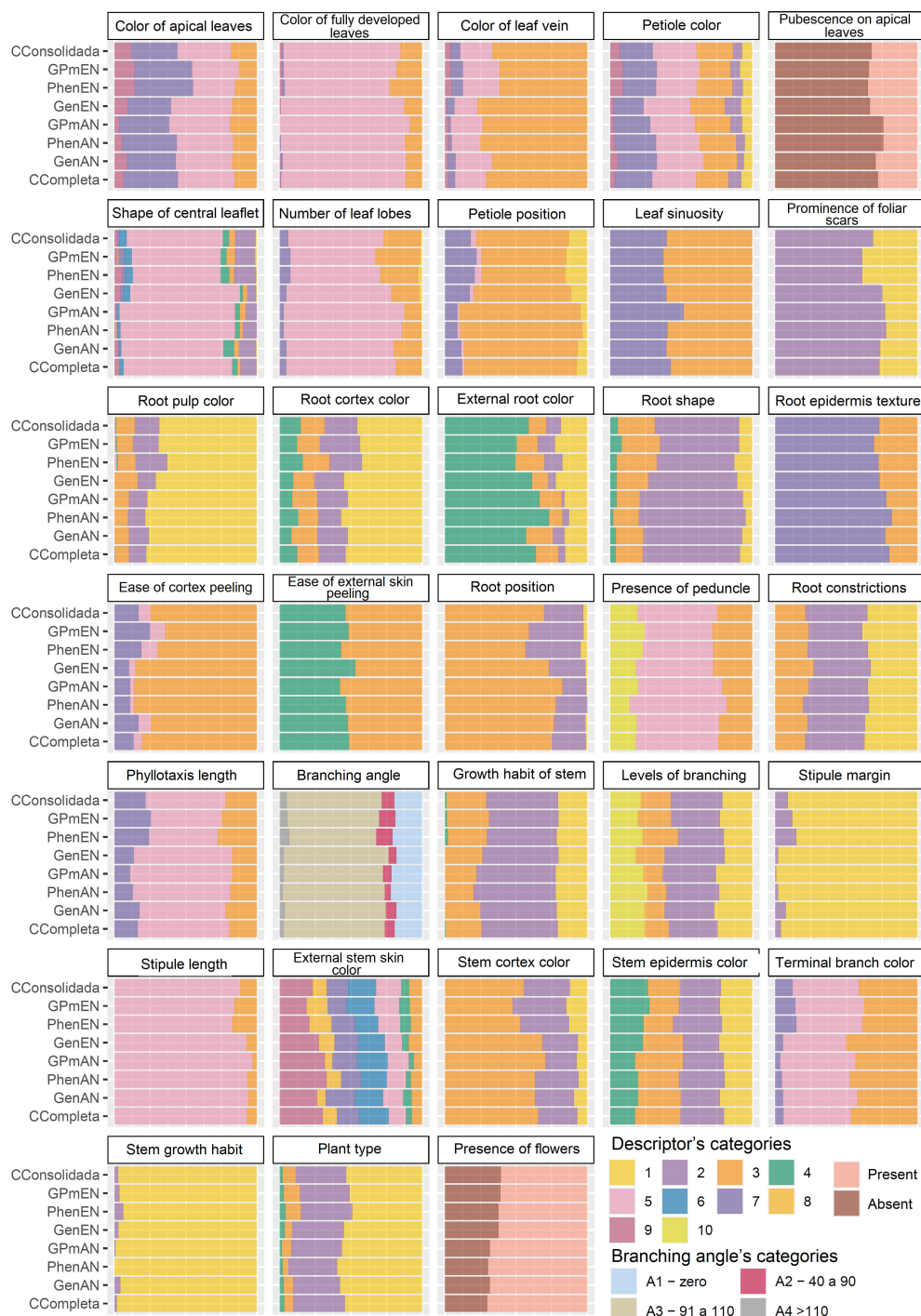


FIGURE 4

Comparative barplot analysis of different qualitative cassava descriptors across various core collections. GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches.

traits, except for growth habit of stem, stipule margin, external root color, and ease of cortex peeling, which exhibited higher ISW compared to the complete collection. However, stipule length showed a reduction in ISW. In the core collections PhenEN, GPmEN, and CCons, there was a trend towards an increase in

ISW compared to the complete collection for most traits (color of fully developed leaf, color of leaf vein, shape of the central leaflet, number of lobes, position of the petiole, prominence of foliar scars, stipule length, terminal branch color, phyllotaxis length, branching angle, stem cortex color, stem growth habit, stipule margin, root

TABLE 4 Shannon-Weaver indices obtained based on 15 quantitative descriptors evaluated in cassava accessions for the development of different core collections.

Trait	Collection							
	Complete	GenAN	GenEN	PhenAN	PhenEN	GPmAN	GPmEN	CCons
Length of leaf lobe	0.80	0.80	0.76	0.68	0.87	0.71	0.88	0.80
Width of leaf lobe	0.82	0.84	0.82	0.76	0.86	0.77	0.90	0.85
Length and width ratio leaf lobes	0.50	0.49	0.59	0.38	0.58	0.52	0.57	0.55
Petiole length	0.70	0.75	0.65	0.62	0.79	0.64	0.80	0.72
Thickness of the root cortex	0.01	0.00	0.02	0.00	0.02	0.00	0.02	0.00
Root length	0.46	0.49	0.50	0.34	0.52	0.41	0.52	0.51
Root diameter	0.59	0.53	0.53	0.56	0.72	0.55	0.68	0.64
Cyanide content	0.86	0.86	0.79	0.85	0.86	0.83	0.83	0.84
Number of roots per plant	0.68	0.72	0.71	0.66	0.73	0.64	0.74	0.71
Dry root yield	0.73	0.68	0.67	0.70	0.75	0.68	0.76	0.69
Dry matter content	0.72	0.68	0.70	0.64	0.75	0.69	0.78	0.73
Fresh root yield	0.77	0.77	0.72	0.73	0.83	0.74	0.82	0.79
Shoot yield	0.79	0.78	0.77	0.72	0.82	0.75	0.81	0.77
Harvest index	0.77	0.77	0.78	0.67	0.86	0.74	0.87	0.83
Plant height	0.71	0.72	0.70	0.65	0.76	0.73	0.78	0.74

Complete collection of cassava germplasm (Complete), GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches.

pulp color, external root color, ease of external skin peeling, root shape, root epidermis texture, and presence of flowers). On the other hand, the PhenAN and GPmAN collections showed a tendency to reduce the ISW, especially for the color of leaf vein, shape of central leaflet, number of lobes, position of the petiole, stem growth habit, stipule margin, plant type, external root color, root cortex prominence, root shape, root position, and presence of root peduncle.

3.6 Validation of core collections

Principal Component Analysis (PCA) was utilized to evaluate the representation of diversity in the core collections based on phenotypic and molecular data (Figures 5, 6, respectively). The first and second principal components accounted for over 38% of the phenotypic variation in cassava accessions, indicating a good representation of phenotypic diversity (Figure 5). Overall, the selected accessions in the different core collections were well distributed across the quadrants of the phenotypic data PCA. However, the GPmEN collection exhibited a higher number of cassava accessions positioned at the extremes of the phenotypic data PCA quadrants, while the consolidated collection demonstrated a slightly more uniform dispersion of cassava accessions across all four quadrants compared to the other collections.

In the PCA analysis of SNPs, the first two principal components accounted for a smaller percentage of the molecular variation in the

data (8.72% and 4.42%, respectively) compared to the phenotypic data (Figure 6). Despite this, similar to the phenotypic data PCA, all core collections were well represented in the molecular data PCA. The consolidated collection and GPmEN exhibited accessions distributed across all quadrants and had a more representative distribution compared to the complete collection.

4 Discussion

4.1 Convergence of selection and diversity in phenotypic and molecular data of core collections

A core collection is a subset of accessions derived from larger germplasm collections with the goal of representing the maximum possible diversity of the original collection (Frankel & Brown, 1984). It is generally recommended to develop core collections that have at least 10% of the size and 70% of the genetic diversity of the original collection (Brown, 1989). Following this recommendation, several core collections of cassava have been constructed using phenotypic and genotypic data alone or in combination, along with a consolidated collection that includes accessions selected by at least two core collections. However, the selection of cassava accessions based on phenotypic and genotypic data did not show high agreement. This lack of correlation between morphological and molecular data has also been observed in potato

TABLE 5 Shannon-Weaver indices obtained based on 33 qualitative descriptors of leaf, stem, root and flower, evaluated in cassava accessions for the development of different core collections.

Trait		Collection							
		Complete	GenAN	GenEN	PhenAN	PhenEN	GPmAN	GPmEN	CCons
Leaf	Color of apical leaves	0.87	0.90	0.90	0.86	0.93	0.84	0.92	0.93
	Color of fully developed leaves	0.26	0.29	0.26	0.25	0.43	0.18	0.40	0.34
	Color of leaf vein	0.58	0.61	0.51	0.51	0.72	0.51	0.71	0.66
	Shape of central leaflet	0.42	0.47	0.43	0.35	0.59	0.34	0.60	0.52
	Petiole color	0.89	0.86	0.88	0.84	0.91	0.87	0.92	0.89
	Pubescence on apical leaves	0.85	0.88	0.92	0.80	0.93	0.80	0.93	0.91
	Number of leaf lobes	0.43	0.42	0.48	0.34	0.58	0.33	0.57	0.52
	Petiole position	0.50	0.50	0.66	0.39	0.77	0.40	0.75	0.71
	Leaf sinuosity	0.98	0.97	0.97	0.97	0.96	1.00	0.96	0.97
	Prominence of foliar scars	0.84	0.83	0.81	0.76	0.96	0.78	0.96	0.90
Stem	Stipule length	0.37	0.33	0.38	0.36	0.67	0.21	0.64	0.52
	Terminal branch color	0.79	0.78	0.79	0.77	0.92	0.74	0.91	0.96
	Phyllotaxis length	0.82	0.86	0.76	0.76	0.96	0.71	0.95	0.90
	Branching angle	0.61	0.62	0.57	0.56	0.76	0.61	0.72	0.69
	External stem skin color	0.92	0.93	0.94	0.91	0.97	0.89	0.97	0.95
	Stem cortex color	0.59	0.63	0.57	0.59	0.70	0.56	0.72	0.69
	Stem epidermis color	0.98	0.99	1.00	0.98	0.98	0.98	0.99	1.00
	Stem growth habit	0.12	0.24	0.18	0.00	0.33	0.06	0.21	0.17
	Stipule margin	0.24	0.38	0.14	0.14	0.60	0.06	0.53	0.43
	Plant type	0.71	0.69	0.70	0.64	0.78	0.67	0.76	0.72
	Growth habit of stem	0.75	0.73	0.73	0.70	0.80	0.72	0.79	0.77
	Levels of branching	0.98	0.97	0.95	0.96	0.99	0.96	0.98	0.98
Root	Root pulp color	0.43	0.45	0.50	0.42	0.61	0.43	0.55	0.55
	Root cortex color	0.86	0.85	0.84	0.83	0.95	0.86	0.92	0.92
	External root color	0.73	0.82	0.75	0.62	0.89	0.67	0.88	0.81
	Ease of cortex peeling	0.55	0.67	0.47	0.42	0.74	0.42	0.79	0.66
	Ease of external skin peeling	1.00	1.00	1.00	1.00	0.99	0.98	1.00	1.00
	Root shape	0.66	0.65	0.73	0.54	0.79	0.62	0.80	0.75
	Root epidermis texture	0.72	0.78	0.84	0.68	0.85	0.76	0.89	0.84
	Root position (RP)	0.55	0.54	0.58	0.48	0.76	0.45	0.71	0.65
	Presence of peduncle	0.88	0.88	0.91	0.76	0.95	0.87	0.96	0.90
	Root constrictions	0.97	0.98	0.99	0.95	0.97	0.98	0.98	0.89
Flower	Presence of flowers	0.88	0.90	0.96	0.88	0.96	0.90	0.96	0.97

GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively; GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively; CCons - consolidated collection that includes accessions selected by at least two of the previous approaches.

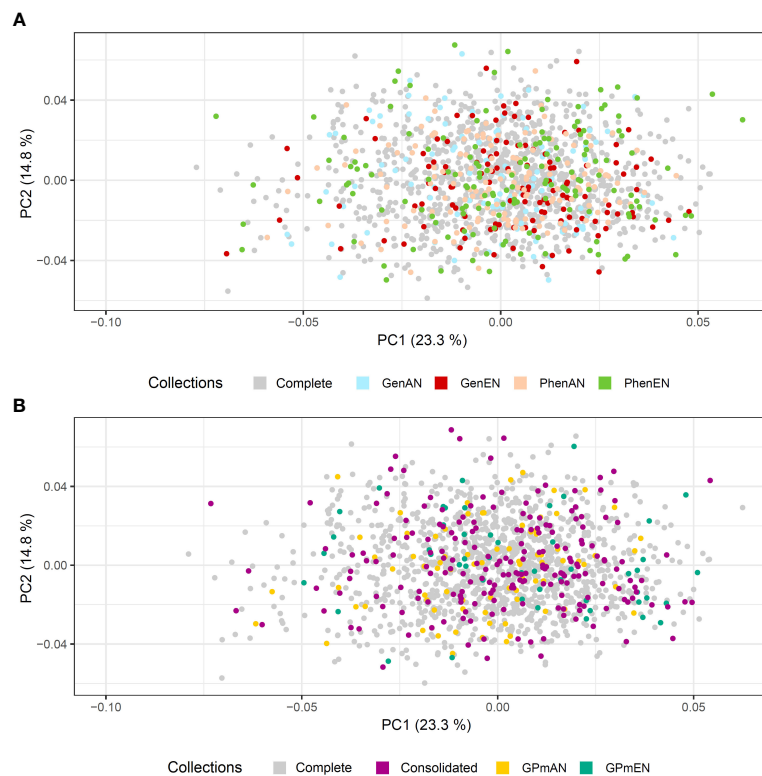


FIGURE 5

Principal component analysis (PCA) of the phenotypic data of 1,486 cassava accessions with dispersion of different core collections. **(A)** Complete collection (CComplete) and GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively. **(B)** Complete collection (CComplete), CCons - consolidated collection that includes accessions selected by at least two of the previous approaches, and GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively.

populations (*Solanum tuberosum* L.) (Berdugo-Cely et al., 2017). The discrepancy can be attributed to the selection pressures that populations undergo, as molecular markers are generally not subject to natural selection, while phenotypic traits are influenced by selection pressures and environmental factors. Another explanation for the low agreement in selection is the weak association between the genomic regions accessed by SNPs and the evaluated phenotypic traits (Oliveira et al., 2012).

Several core collections have been developed based on phenotypic data alone (Upadhyaya and Ortiz, 2001; Upadhyaya et al., 2009; Mahmoodi et al., 2019). Although phenotypic data is directly related to agronomic and yield attributes, it can be influenced by environmental factors, experimental errors, and genotype \times environment interactions. Therefore, it is recommended to construct core collections that incorporate both phenotypic and genotypic data to ensure maximum representativeness of the original collection for a wide range of data types and characteristics (Kumar et al., 2016; Xue et al., 2021), without losing important alleles for conservation and improvement purposes.

Due to the low agreement in the selection of cassava accessions among different core collections and the risk of excluding accessions with important phenotypic or molecular characteristics, a consolidated core collection was created by including accessions

selected by at least two methodological approaches (EN and AN) and different types of collections (Gen, Phen, and GPm). This slightly increased the number of selected clones (from 10% to ~14% of the complete collection), which is still manageable within the scope of genetic resources and species breeding programs.

Overall, the cassava core collections effectively retained a high number of SNP alleles from the complete collection, surpassing 94.5%. Notably, the consolidated and GenEN collections exhibited the highest allelic richness, retaining 97.1% and 97.73% of the alleles, respectively. This preservation of allelic richness in core collections holds significant importance for future studies on genomic associations, especially for traits controlled by rare alleles. Furthermore, the allelic richness retained in cassava core collections compares favorably to other species such as maize (93% - Todorovska et al., 2005) and tomato (92% - Martins et al., 2015), indicating promising results.

While minimal changes were observed in the analyzed genetic parameters, core collections constructed based on phenotypic and genotypic information separately exhibited greater deviations in diversity values and genetic parameters (H_o , H_s , and F_{is}) compared to the complete and consolidated collections. Methodologically, the collections obtained through the AN algorithm demonstrated H_o , H_s , and F_{is} values more similar to the complete collection, likely due to the algorithm's aim of achieving a similar representation of the

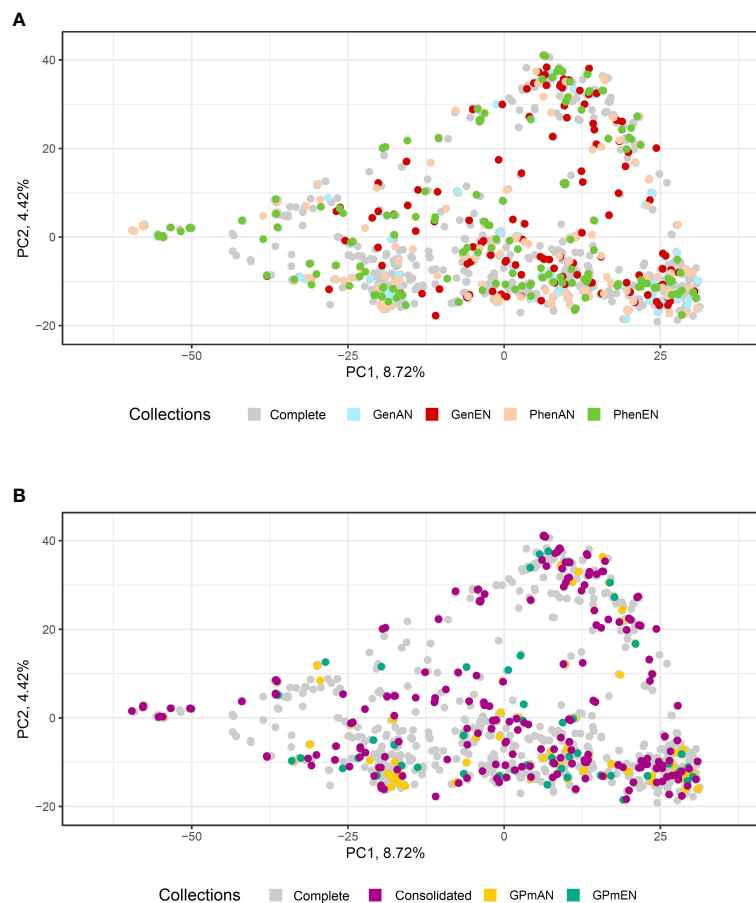


FIGURE 6

Principal component analysis (PCA) based on 20,023 single-nucleotide polymorphisms (SNP) of 1486 cassava accessions with dispersion of different core collections. **(A)** Complete collection (CComplete) and GenAN and GenEN - core collection formed by genotypic data and optimization strategy average accession-to-nearest-entry (AN) and average entry-to-nearest-entry (EN), respectively; PhenAN and PhenEN - Core collection formed by phenotypic data and optimization strategy AN and EN, respectively. **(B)** Complete collection (CComplete), CCons - consolidated collection that includes accessions selected by at least two of the previous approaches, and GPmAN and GPmEN - Collection formed by morpho-agronomic data + SNPs and optimization strategy AN and EN, respectively.

complete collection's characteristics in the core collection. In general, the variation of *Ho* (0.39 to 0.40) and *Hs* (0.29 to 0.30) observed in core collections for biallelic markers closely aligns with other cassava germplasms, such as the global cassava collection from IITA, which exhibited *Ho* values ranging from 0.33 to 0.39 and *Hs* ranging from 0.31 to 0.34 (Ferguson et al., 2019).

The distribution and representativeness of the core collections demonstrated relative similarity to the complete collection, considering the mean and variance estimates of the quantitative phenotypic data. However, the distribution profile of the means and variances of the phenotypic data more closely resembled that of the complete collection. Similar findings of few significant differences between the complete collection and core collections were reported in studies on maize landraces (Li et al., 2005) and rice (*Oryza sativa* L.) (Yan et al., 2009). These results highlight the high potential of these core collections to represent the complete collection without significant losses in genetic variability.

In specific cases, an increase in the variance of certain traits was observed, likely due to the removal of accessions that made redundant contributions to phenotypic diversity, with phenotypic values close to

the sample mean. For instance, the mean root diameter variance in the consolidated collections, GPmEN, and PhenEN increased by 31%, 49%, and 62%, respectively, compared to the complete collection. However, this increase in variance within core collections is a desirable attribute as it allows for a greater representation of the complete collection while minimizing redundancy (Hu et al., 2000). This is important for obtaining manageable collections that can be practically used in discovering new characteristics and incorporating them into the species improvement efforts.

The ISW was also utilized to assess the efficiency of core collections in representing cassava germplasm diversity. Overall, variations greater than $\pm 10\%$ of the ISW compared to the complete collection were observed for only a few quantitative phenotypic traits. The characteristics that were most affected by the ISW reaching 0.00 in the collections obtained with the AN algorithm were those with the greatest imbalance in their distribution, such as the thickness of the root cortex. On the other hand, the collections obtained with the EN algorithm exhibited higher ISW values than the complete collection. Similar results were identified in the analysis of qualitative phenotypic data, where the collections

based on the EN algorithm and the consolidated collection showed higher ISW compared to the complete collection for most traits. However, similar to other species with clonal propagation such as yam (Beukelaer and Davenport, 2018; Girma et al., 2018), cassava core collections maintained or even increased diversity based on ISW among the core collections.

4.2 Optimizing diversity and representation in core collections: strategies for effective utilization of germplasm resources

The construction of core collections involves various methodological considerations that can impact the representativeness of genetic diversity. In this study, two criteria described by Odong et al. (2013) were employed to create the core collections, each serving a distinct purpose. The Type 1 collection aimed to maximize the genetic diversity of the complete collection by encompassing all original diversity. This type of collection ensured a more balanced representation of phenotypic characteristics, including those with both low and high frequencies. The collections formed using the AN algorithm, which minimized the average distance between accessions in the complete dataset and the closest selected accession in the core collection, maintained similar levels of heterozygosity, genetic diversity, and inbreeding coefficients compared to the complete collection.

The second collection in this study, known as type 3 aimed to represent the distribution of accessions in the complete collection. Its objective was to ensure that the selected proportion of the complete collection reflects the numerical contributions of different categories in the core collection. The EN algorithm was used to form these collections, which selected accessions that were well-distributed, particularly at the extremes of the different quadrants of the PCA. This approach provided a better representation of the entire collection, resulting in more diverse collections where each selected individual was sufficiently different from others. As a result, subsets with low redundancy (Odong et al., 2013) and high representativeness of the descriptors used to form the collections were obtained. This increased sample diversity was evident when considering the ISW.

The consolidated collection was developed to address the issue of low coincidence in the selection of accessions among the core collections. It served as an alternative to better represent the cassava accessions among the six collections developed based on different types of data. The consolidated collection proved to be efficient not only in overcoming the low coincidence but also in improving allele retention. It resulted in less difference in genetic parameters among the collections and maintained maximum diversity in the ISW for all traits. Furthermore, it better represented the phenotypic and genotypic classes of the complete collection in the PCA.

4.3 Validating the effectiveness of core collections: enhancing representation and retention of genetic diversity

The distribution of selected accessions in the core and consolidated collections of cassava exhibited a remarkable level of

representativeness when compared to the complete collection, as evidenced by the PCA analysis conducted on both phenotypic and molecular data. Despite the presence of population structure in both data sets, the cassava accessions were well dispersed across different quadrants of the PCA, with notable emphasis on the GPmEN and consolidated collections. This resulted in the selection of cassava accessions with minimal redundancy within the core collections. Similar studies conducted on other species, such as *Lagenaria siceraria*, have also demonstrated that PCA analysis of core collections, utilizing various phenotypic data types, accurately represents the complete collection and preserves the geographic distribution of accessions (Wang et al., 2021).

It is important to acknowledge that there is no universally applicable ratio or fixed size for all core collections, as the research requirements vary among different species. Nevertheless, the consolidated collection outlined in this study, which comprises approximately 14% of the complete collection, exhibits an appropriate sample ratio considering the extensive breadth and complexity of cassava genetic resources. This consolidated collection serves as a valuable and comprehensive reference, forming a solid basis for the utilization of cassava germplasm resources in future breeding programs.

4.4 Cassava core collections for conservation, characterization and use of cassava genetic resources

The conservation of cassava genetic resources is crucial for research purposes and the discovery of genes with agronomic significance to be used in cassava breeding programs. However, Guo et al. (2014) highlighted the challenge of maintaining and utilizing the diversity of accessions in a germplasm bank. The entire process of conservation and characterization is labor-intensive, time-consuming, and requires substantial financial resources. In this context, the technological advancements developed in this study offer a relevant alternative for reducing costs associated with the conservation and characterization stages of cassava germplasm.

The main objective of developing the core collection, in addition to reducing the size of the set and maintaining genetic representativeness, is to define conservation priorities, prioritize and allocate efforts for characterizations and evaluations, facilitate access, and enhance knowledge of the available genetic structure in germplasm banks. The consolidated core collection will facilitate the handling of a more focused and detailed morphological and agronomic variability, enabling comprehensive characterization studies. These measures aim to optimize the conservation and utilization of cassava germplasm while ensuring the preservation of currently available genetic resources. Moreover, this core collection will be given priority for *in vitro* conservation, ensuring protection against environmental degradation and facilitating efficient exchange of the collection. It is important to note that genetic collections should be dynamic and periodically reviewed to incorporate additional accessions. This ensures that the most valuable genotypes are preserved and characterized, serving the purpose of conservation and species improvement.

5 Conclusion

This study highlights the possibility of using diverse methodological approaches and data types to construct core collections for cassava, effectively preserving the diversity and genetic parameters of the complete collection. However, the low overlap in the selection of accessions among different core collection formation algorithms necessitated the creation of an alternative collection called the consolidated collection. This collection incorporated cassava accessions selected by at least two different algorithms, combining phenotypic and genotypic data.

The consolidated collection demonstrated less variation in the analyzed genetic parameters compared to the complete collection. It retained over 97% of the allelic richness observed in the complete collection, even with the inclusion of accessions selected based on different types of information. Additionally, the consolidated collection exhibited similar data dispersion and representation of classes in both quantitative and qualitative characteristics when compared to the complete collection. Despite representing a larger percentage of the complete collection than initially planned (approximately 14%), the consolidated collection remains manageable in size, allowing for efficient characterization and utilization of the germplasm. Overall, the formation of the consolidated collection addresses the challenge of low coincidence in accession selection and provides a robust and representative resource for further research and breeding programs in cassava.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: <https://doi.org/10.6084/m9.figshare.23818875.v1>.

Author contributions

LD, EO designed the study. CS performed the field evaluations. CS, LD, and CC analyzed the data. CS, and CC wrote the paper. All authors contributed to the article and approved the submitted version.

References

- Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., and Yockteng, R. (2017). Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. *PLoS One* 12 (3), e0173039. doi: 10.1371/journal.pone.0173039
- Berends, M. S., Luz, C. F., Friedrich, A. W., Sinha, B. N., Albers, C. J., and Glasner, C. (2022). AMR: an R package for working with antimicrobial resistance data. *J. Stat. Software* 104, 1–31. doi: 10.18637/jss.v104.i03
- Beukelaer, H., and Davenport, G. (2018) *Corehunter: multi-purpose core subset selection*. Available at: <https://CRAN.R-project.org/package=corehunter>.
- Bhattacharjee, R., Dumet, D., Ilona, P., Folarin, S., and Franco, J. (2012). Establishment of a cassava (*Manihot esculenta* Crantz) core collection based on

Funding

CS: CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Grant number: 88887.492719/2020-00. LD: Empresa Brasileira de Pesquisa Agropecuária. Grant number: 20.18.01.012.00.00. EO: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Grant number: 409229/2018-0, 442050/2019-4 and 303912/2018-9. EO: FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia). Grant number: Pronem 15/2014. This work was partially funded by UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation. Grant INV-007637. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. The funder provided support in the form of fellowship and funds for the research, but did not have any additional role in the study design, data collection and analysis, decision to publish, nor preparation 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.2023.1250205/full#supplementary-material>

agromorphological descriptors. *Plant Genet. Resour.* 10, 119–127. doi: 10.1017/S1479262112000093

Blighe, K., and Lun, A. (2023) *PCAtools: PCAtools: Everything Principal Components Analysis*. Available at: <https://github.com/kevinblighe/PCAtools>.

Boczkowska, M., Lapiński, B., Kordulsińska, I., Dostatny, D. F., and Czembor, J. H. (2016). Promoting the use of common oat genetic resources through diversity analysis and core collection construction. *PLoS One* 11 (12), e0167855. doi: 10.1371/journal.pone.0167855

Bradbury, M. G., Egan, S. V., and Bradbury, J. H. (1999). Determination of all forms of cyanogens in cassava roots and cassava products using picrate paper kits. *J. Sci. Food Agric.* 79, 593–601. doi: 10.1002/(SICI)1097-0010(19990315)

- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308
- Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature* 34, 562–571. doi: 10.1038/nbt.3535
- Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management. *Genome* 31 (2), 818–824. doi: 10.1139/g89-144
- Brown, A. H. D., and Spillane, C. (1999). “Implement core collections – principles, procedures, progress, problems and promise,” in *Core collections for today and tomorrow*. Eds. R. C. Johnson and T. Hodgkin (Roma: IPGRI), 1–9.
- Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98 (1), 116–126. doi: 10.1016/j.ajhg.2015.11.020
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. psychol. measurement* 20 (1), 37–46. doi: 10.1177/001316446002000104
- Covarrubias-Pazarán, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11 (6), e0156744. doi: 10.1371/journal.pone.0156744
- Díez, M. J., de la Rosa, L., Martín, I., Guasch, L., Cartea, M. E., Mallor, C., et al. (2018). Plant genebanks: present situation and proposals for their improvement. *Case Spanish network. Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01794
- Doyle, J. J., and Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.
- 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
- Ferguson, M. E., Shah, T., Kulakow, P., and Ceballos, H. A. (2019). Global overview of cassava genetic diversity. *PLoS One* 14 (11), e0224763. doi: 10.1371/journal.pone.0224763
- Frankel, O. H., and Bennett, E. (1970). *Genetic resources in plants-their exploration and conservation* 1970. (Oxford: Blackwell Scientific Publications).
- Frankel, O. H., and Brown, A. H. D. (1984). “Plant genetic resources today: a critical appraisal,” in *Crop Genetic Resources: Conservation and Evaluation*. Eds. J. H. W. Holden and J. T. Williams (Winchester: Allen and Unwin), 249–257.
- Frankel, O. H., Frankel, O., and Soule, M. E. (1981). *Conservation and evolution* (New York: Cambridge University Press).
- Fukuda, W. M. G., Guevara, C. L., Kawuki, R., and Ferguson, M. E. (2010). *Selected morphological and agronomic descriptors for the characterization of cassava* (Ibadan, Nigeria: IITA).
- Fukuda, W. M. G., Oliveira, S., and Iglesias, C. (2002). Cassava breeding. *Crop Breed. Appl. Biotechnol.* 2 (4), 617–638. doi: 10.12702/1984-7033.v02n04a18
- Girma, G., Bhattacharjee, R., Lopez-Montes, A., Gueye, B., Ofole, S., Franco, J., et al. (2018). Re-defining the yam (*Dioscorea* spp.) core collection using morphological traits. *Plant Genet. Resour.* 16 (3), 193–200. doi: 10.1017/S1479262117000144
- Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5 (1), 184–186. doi: 10.1111/j.1471-8286.2004.00828.x
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27 (4), 857–871. doi: 10.2307/2528823
- Guo, Y., Li, Y., Hong, H., and Qiu, L. J. (2014). Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (*Glycine max*). *Crop J.* 2 (1), 38–45. doi: 10.1016/j.cj.2013.11.001
- Hamblin, M. T., and Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). *Crop Sci.* 54 (6), 2603–2608. doi: 10.2135/cropsci2014.02.0160
- Hershey, C. (1994). *Research for development: The CIAT cassava program*. [Centro Internacional de Agricultura Tropical (CIAT): Cali, CO.] 99 p.
- Hershey, C. (2017). “Ex situ conservation of cassava genetic material,” in *Achieving sustainable cultivation of cassava*. Ed. C. Hershey (Cambridge, UK: Burleigh Dodds Science Publishing), 59–68.
- Hu, J., Zhu, J., and Xu, H. M. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet.* 101, 264–268. doi: 10.1007/s001220051478
- Kawano, K., Fukuda, W. M. G., and Cenpukdee, U. (1987). Genetic and environmental effects on dry matter content of cassava root. *Crop Sci.* 26, 69–74. doi: 10.2135/cropsci1987.0011183X002700010018x
- Kumar, S., Ambreen, H., Variath, M. T., Rao, A. R., Agarwal, M., Kumar, A., et al. (2016). Utilization of molecular, phenotypic, and geographical diversity to develop compact composite core collection in the oilseed crop, safflower (*Carthamus tinctorius* L.) through maximization strategy. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01554
- Lebot, V. (2009). *Tropical root and tuber crops: cassava, sweet potato, yams and aroids* (CABI: Crop Production Science in Horticulture).
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, Y., Shi, Y., Cao, Y., and Wang, T. (2005). Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. *Genet. Resour. Crop Evol.* 51 (8), 845–852. doi: 10.1007/s10722-005-8313-8
- Mahmoodi, R., Dadpour, M. R., Hassani, D., Zeinalabedini, M., Vendramin, E., Micali, S., et al. (2019). Development of a core collection in Iranian walnut (*Juglans regia* L.) germplasm using the phenotypic diversity. *Scientia Hort.* 249, 439–448. doi: 10.1016/j.scienta.2019.02.017
- Martins, F. A., Silva, D. J. H., and Carneiro, P. C. S. (2015). Establishment of a core collection based on the integration of morphoagronomic, phytopathological and molecular data. *Rev. Ciec. Agron.* 46 (4), 836–845. doi: 10.5935/1806-6690.20150072
- Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C., and Georges, A. (2022). DartR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods Ecol. Evol.* 13 (10), 2150–2158. doi: 10.1111/2041-210X.13918
- Odong, T. L., Jansen, J., Van Eeuwijk, F. A., and Van Hintum, T. J. (2013). Quality of core collections for effective utilization of genetic resources review, discussion and interpretation. *Theor. Appl. Genet.* 126 (2), 289–305. doi: 10.1007/s00122-012-1971-y
- Oliveira, E. J., De Resende, M. D. V., Santos, V. S., Ferreira, C. F., Oliveira, G. A. F., Silva, M. S., et al. (2012). Genome-wide selection in cassava. *Euphytica* 187 (2), 263–276. doi: 10.1007/s10681-012-0722-0
- Oliveira, E. J., Ferreira, C. F., Santos, S., and Oliveira, G. A. (2014). Development of a cassava core collection based on single nucleotide polymorphism markers. *Genet. Mol. Res.* 13 (3), 6472–6485. doi: 10.4238/2014.august.25.11
- R Development Core Team (2023). *R: A language and environment for statistical computing, reference index version 4.3.1* (Vienna, Austria: R Foundation for Statistical Computing). Available at: <http://www.R-project.org>.
- Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., and Davenport, G. F. (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinf.* 10 (1), 1. doi: 10.1186/1471-2105-10-243
- Todorovska, E., Abumhadi, N., Kamenarova, K., Zheleva, D., Kostova, A., Christov, N., et al. (2005). Biotechnological approaches for cereal crops improvement: Part II: Use of molecular markers in cereal breeding. *Biotechnol. Biotechnol. Equip.* 19 (3), 91–104. doi: 10.1080/13102818.2005.10817289
- Upadhyaya, H. D., and Ortiz, R. (2001). A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. *Theor. Appl. Genet.* 102, 1292–1298. doi: 10.1007/s00122-001-0556-y
- Upadhyaya, H. D., Pundir, R. P. S., Dwivedi, S. L., Gowda, C. L. L., Reddy, G., and Singh, S. (2009). Developing a mini core collection of sorghum for diversified utilization of germplasm. *Crop Sci.* 49 (5), 1769–1780. doi: 10.2135/cropsci2009.01.0014
- Van Hintum, T. J. L., Brown, A. H. D., Spillane, C., and Hodgkin, T. (2000). *Core collection of plant genetic resources* (Roma, Italia: IPGRI Technical Bulletin: International Plant Genetic Resources), 75.
- Wang, C., Hu, S., Gardner, C., and Lübberstedt, T. (2017). Emerging avenues for utilization of exotic germplasm. *Trends Plant Sci.* 22 (7), 624–637. doi: 10.1016/j.tplants.2017.04.002
- Wang, Y., Wu, X., Li, Y., Feng, Z., Mu, Z., Wang, J., et al. (2021). Identification and validation of a core single-nucleotide polymorphism marker set for genetic diversity assessment, fingerprinting identification, and core collection development in bottle gourd. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.747940
- Wright, S. (1978). *Evolution and the genetics of populations: a treatise in four volumes Volume IV* (Chicago: University of Chicago Press).
- Xue, H., Yu, X., Fu, P., Liu, B., Zhang, S., Li, J., et al. (2021). Construction of the Core Collection of *Catalpa fargesii* f. *duclouxii* (Huangxinzimu) based on molecular markers and phenotypic traits. *Forests* 12 (11), 1518. doi: 10.3390/f12111518
- Yan, W. G., Li, Y., Agrama, H. A., Luo, D., Gao, F., Lu, X., et al. (2009). Association mapping of stigma and spikelet characteristics in rice (*Oryza sativa* L.). *Mol. Breed.* 24 (3), 277–292. doi: 10.1007/s11032-009-9290-y



OPEN ACCESS

EDITED BY

Manjusha Verma,
National Bureau of Plant Genetic
Resources (ICAR), India

REVIEWED BY

Dagmar Janovská,
Crop Research Institute (CRI),
Czechia
Mohammad Ehsan Dulloo,
Alliance Bioversity International and CIAT,
France

*CORRESPONDENCE

Stephan Weise
✉ weise@ipk-gatersleben.de

†PRESENT ADDRESS

Kim Jana Kutschan,
IDT Biologika GmbH, Dessau, Germany

RECEIVED 23 June 2023

ACCEPTED 06 September 2023

PUBLISHED 09 October 2023

CITATION

Weise S, Hoekstra R, Kutschan KJ,
Oppermann M, van Treuren R and
Lohwasser U (2023) Analysis of gaps in
rapeseed (*Brassica napus* L.) collections in
European genebanks.
Front. Plant Sci. 14:1244467.
doi: 10.3389/fpls.2023.1244467

COPYRIGHT

© 2023 Weise, Hoekstra, Kutschan,
Oppermann, van Treuren and Lohwasser.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). 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.

Analysis of gaps in rapeseed (*Brassica napus* L.) collections in European genebanks

Stephan Weise^{1*}, Roel Hoekstra², Kim Jana Kutschan^{1†},
Markus Oppermann¹, Rob van Treuren² and Ulrike Lohwasser¹

¹Genebank Department, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) OT Gatersleben, Seeland, Germany, ²Centre for Genetic Resources, the Netherlands (CGN), Wageningen University & Research, Wageningen, Netherlands

Rapeseed is one of the most important agricultural crops and is used in many ways. Due to the advancing climate crisis, the yield potential of rapeseed is increasingly impaired. In addition to changing environmental conditions, the expansion of cultivated areas also favours the infestation of rapeseed with various pests and pathogens. This results in the need for continuous further development of rapeseed varieties. To this end, the potential of the rapeseed gene pool should be exploited, as the various species included in it contain promising resistance alleles against pests and pathogens. In general, the biodiversity of crops and their wild relatives is increasingly endangered. In order to conserve them and to provide impulses for breeding activities as well, strategies for the conservation of plant genetic resources are necessary. In this study, we investigated to what extent the different species of the rapeseed gene pool are conserved in European genebanks and what gaps exist. In addition, a niche modelling approach was used to investigate how the natural distribution ranges of these species are expected to change by the end of the century, assuming different climate change scenarios. It was found that most species of the rapeseed gene pool are significantly underrepresented in European genebanks, especially regarding representation of the natural distribution areas. The situation is exacerbated by the fact that the natural distributions are expected to change, in some cases significantly, as a result of ongoing climate change. It is therefore necessary to further develop strategies to prevent the loss of wild relatives of rapeseed. Based on the results of the study, as a first step we have proposed a priority list of species that should be targeted for collecting in order to conserve the biodiversity of the rapeseed gene pool in the long term.

KEYWORDS

rapeseed, plant genetic resources, gap analysis, niche modelling, *Brassica*

1 Introduction

Crop plants are major components of human and animal nutrition (Grusak and DellaPenna, 1999), and play an important role as renewable resources or as basic ingredients for chemical or pharmaceutical industry (Metzger and Bornscheuer, 2006; Tilman et al., 2006). Rapeseed or canola (*Brassica napus* L.) is rated amongst the most important agricultural crops. It is used as high-quality edible oil, as high-protein feed for livestock breeding, as biofuel or as raw material for chemical industry like surfactants, softening agents or biodegradable varnishes (Bell, 1982; Piazza and Foglia, 2001; Link, 2008).

Amongst oil plants, rapeseed holds the second largest global market share after soybean (USDA, 2023). In addition, it is also one of the most important sources of protein for animal feed (Hu et al., 2021). Rapeseed provides an average of 40–50% of oil on a dry basis and a protein content of 17–26%. This was achieved by means of breeding programmes starting in the 1970s, which resulted in a significant reduction of harmful glucosinolates and erucic acid (Link, 2008; Jahreis and Schäfer, 2011; Barthet, 2016; Dhillon et al., 2016), in double-low (also called double-zero) cultivars. Rapeseed is the major crop for oilseed production in Europe (Carré and Pouzet, 2014).

Due to the progressing climate crisis, the yield potential of rapeseed is increasingly affected. In particular, rising temperatures, shifting rainfall patterns and increasing incidence of extreme weather events lead to unfavourable effects, such as decreasing yield (Qian et al., 2018). Changed environmental conditions and the extension of acreage foster the infestation of rapeseed with different pests and pathogens (Link, 2008; Williams, 2010). Amongst the most important pathogens are the beet western yellows virus, the *Phoma lingam* and *Verticillium* species, respectively (Gilligan et al., 1980; Heale and Karapapa, 1999; Fitt et al., 2006; Hwang et al., 2016).

In order to cope with the increasing demand for rapeseed and with climatic changes, it is necessary to improve rapeseed varieties continuously. Therefore, it is promising to exploit the potential of its gene pool for breeding through crossing with related species from the primary, secondary and tertiary gene pool, respectively (Chen and Heneen, 1990; Girke et al., 2012). The high potential of crop wild relative (CWR) species was first recognised in the 1920s by the Russian geneticist Nikolay Ivanovich Vavilov (Vavilov, 1926). CWRs usually have a broader genetic variability than crop plants domesticated over hundreds of years (Singh, 2001; McCouch, 2004; Vollbrecht and Sigmon, 2005). The importance of CWRs is continuously growing through scientific progress, e.g. by biotechnological methods enabling gene transfer between distantly related species (Hajjar and Hodgkin, 2007; Ford-Lloyd et al., 2011; Maxted et al., 2012).

In order to simplify CWR classification, Harlan and de Wet proposed to categorise the total available gene pool of a crop and its related species into three groups depending on their degree of relationship with the crop of interest (Harlan and de Wet, 1971). The primary gene pool contains the cultivated species and taxa with which it is completely inter-fertile, thus allowing easy inter-crossing. The secondary gene pool comprises taxa from different

species, which are nonetheless closely related. These species can be used for crossing with at least some fertile hybrids. Gene transfer is difficult and may require the use of biotechnological techniques. The tertiary gene pool consists of more distantly related species. Crossing is only possible through the use of biotechnological techniques, such as embryo rescue or bridge crossing, requiring considerable effort.

The gene pool of rapeseed contains species with a large variety of promising resistance alleles against pests and pathogens. Related species, such as *Brassica elongata* Ehrh., *Brassica nigra* (L.) W. D. J. Koch, *Brassica juncea* (L.) Czern., *Sinapis alba* L. or *Sinapis arvensis* L., are known to harbour resistance genes against *P. lingam* whereas *Brassica rapa* L. has been found to show resistances against *Verticillium* wilt (Gerdemann-Knörck et al., 1994; Diederichsen and Sacristan, 1996; Snowden et al., 2000; Chen et al., 2007; Rygulla et al., 2007; Wei et al., 2010). However, the biological diversity of crop wild relatives is increasingly threatened, not only by the changing climate but also by population expansion, urbanisation and environmental pollution, respectively (Bakarr et al., 2007; Jarvis et al., 2008; van Treuren et al., 2012).

In order to prevent the extinction of species on the one hand and to provide new impulses to breeding programmes on the other hand, it is indispensable to further develop strategies for preserving plant genetic resources for food and agriculture (PGRFA) (Maxted et al., 2012; Parra-Quijano et al., 2012). Important contributions to the preservation of PGRFA are the collection, the maintenance and the characterisation of crop plants and crop wild relatives. In particular, genebanks play an important role for the long-term conservation of PGRFA (Hoisington et al., 1999). There are about 1,800 collections conserving PGRFA around the world. Thereof, about 625 collections are maintained in Europe comprising more than 2 million accessions (Engels and Maggioni, 2012).

A widely accepted goal for the conservation of plant genetic resources is to conserve 95% of all alleles of a random locus that occur in a target population at a frequency of more than 5% (Nagel et al., 2022). However, for decades there has been controversy about what this means in terms of the minimum number of genebank accessions required. When collecting species, there is a tension between the goal of representing the greatest possible genetic diversity of the individual species and the simultaneous practical requirement of having to limit the size of the samples to a manageable level (Allard, 1970). Marshall and Brown (1975) have suggested sampling 50 populations within an ecogeographical region, collecting 50 individual plants per population. In contrast, Crossa et al. (1993) consider that a sample size of 160–210 plants is sufficient. Lawrence et al. (1995), in turn, conclude that 172 randomly sampled plants from a population of a species are sufficient to maintain genetic diversity. When collecting several populations, it is sufficient to take no more than 172 plants per population divided by the number of populations. Irrespective of this discussion, even in the case of a very intensive collection of individual species, the question arises to what extent genetic diversity is covered by a large number of accessions alone (Maxted et al., 2008). For this purpose, other evaluation criteria must also be taken into account, for example the taxonomic composition of the gene pool, the threat status or ecogeographical

aspects. It is assumed that sampling populations from distant sites and different habitats will give a more representative coverage of the genetic diversity of a taxon (Maxted et al., 1995). However, most of the existing collections of a target crop hardly contain a representation of the entire known population of the target crop.

In that regard, gap analysis is an important aspect of genetic resources management. In general, gap analysis is a technique to identify shortcomings in biodiversity conservation actions, e.g. missing biodiversity in plant genetic resources collections or in protected areas (Jennings, 2000; Margules and Pressey, 2000; Maxted et al., 2008). It comprises various steps (Burley, 1988): (1) to identify the biodiversity within a region; (2) to examine existing conservation approaches, e.g. protected areas; (3) to determine, which elements of the biodiversity are underrepresented by the existing conservation approaches; (4) to define additional conservation actions.

In principle, gap analysis is applicable to both *ex situ* and *in situ* conservation. The present paper focusses on the *ex situ* conservation of PGRFA in genebank collections. In this context, gap analysis helps to identify the geographical distribution of species of interest and allows comparing with existing genebank holdings. Detected gaps can then be closed, e.g. by organising collecting expeditions.

Comprehensive information about the composition of European germplasm collections is available from the European Search Catalogue for Plant Genetic Resources (EURISCO) (Weise et al., 2017; Kotni et al., 2023). EURISCO is an information system, which documents more than two million accessions maintained *ex situ* in more than 400 collections. It is maintained on behalf of the European Cooperative Programme for Plant Genetic Resources (ECPGR) and is based on a network of National Inventories of 43 member countries from Europe and beyond.

To prioritise species for the improvement of genebank collections, the expected effects of climate change on the distribution of species in their natural environment should be taken into account. In addition, other threat assessments should also be taken into consideration, such as the IUCN Red List of Threatened Species (IUCN, 2023).

Since the expected climate changes depend on a large number of factors, various scenarios have been developed that are referred to as Representative Concentration Pathways (RCPs). Four scenarios have become established (RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5), which predict possible changes in greenhouse gas emissions up to the year 2100 in relation to the pre-industrial age around the year 1750 (van Vuuren et al., 2011). This paper uses only the two most contrasting scenarios RCP 2.6 and RCP 8.5 to assess the impacts of climate change. RCP 2.6 is an optimistic scenario assuming that greenhouse gas emissions will decline after 2020. In this case, a global temperature increase of 0.3 to 1.7°C is expected between 2081 and 2100. In contrast, in the pessimistic scenario RCP 8.5, greenhouse gas emissions continue to rise resulting in a temperature increase of 2.6 to 4.8°C over the same period (Stocker et al., 2013). There's nowadays little doubt that climate change will have significant effects on the distribution of species in their natural habitats, as the current environmental conditions will very likely be affected by climate change. Such effects can be

estimated using Species Distribution Models (SDMs). In these models, the presence of a particular species at geographical locations is related to the local environmental conditions, such as temperature and precipitation parameters, after which these relationships can be used to predict species occurrence at other locations. When climate change scenarios are included in the modelling, predictions can be made of the future distribution of a species (Aguirre-Gutiérrez et al., 2017).

Here, we aim to analyse the representation of rapeseed and its wild relatives in European genebank collections and to identify gaps. In addition, ecological niche modelling was used to predict the effects of climate change on future species distributions in order to prioritise relevant species for conservation.

2 Materials and methods

2.1 Species data

To determine the composition of the primary, secondary and tertiary gene pool of rapeseed, the Crop Wild Relative Inventory (Vincent et al., 2013; CWR, 2023) of the Global Crop Diversity Trust was used. According to the Crop Wild Relative Inventory, the gene pool of rapeseed comprises various taxa from 16 genera. For improving the accuracy of further comparisons with data on germplasm holdings, these taxa were checked for synonym names based on The World Flora Online (WFO, 2023), which resulted in about 900 additional names (including subtaxa). Based on the identified composition of the rapeseed gene pool, germplasm collections across Europe were examined for gaps, i.e. countries in the distribution range of a species for which no accession are available in PGR collections. The emphasis was on European collections, because rapeseed originated from natural crossings of *B. rapa* and *Brassica oleracea* L (Neuffer, 2001), both occurring in the Mediterranean area. Also, the other species of the rapeseed gene pool mainly occur in European temperate areas. Therefore, the natural occurrence ranges of the species of the rapeseed gene pool were determined using the Euro+Med PlantBase (Euro+Med, 2023) and GRIN Taxonomy (GRIN, 2023). Information about the origin countries of genebank accessions, the countries maintaining the accessions as well as the numbers of available accessions was extracted from EURISCO (EURISCO, 2023). This data was then compared with the natural ranges of the species. In addition, the species of the rapeseed genepool were checked against the IUCN Red List of Threatened Species (IUCN, 2023).

2.2 Species distribution modelling

Species distribution modelling (or ecological niche modelling) was used to predict the effects of climate change on the future distribution of the wild relatives of *B. napus* in Europe and countries bordering the Mediterranean Sea. Modelling procedures followed the methods described by Aguirre-Gutiérrez et al. (2017) and van Treuren et al. (van Treuren et al., 2017; van Treuren et al., 2020). Table 1 lists 51 taxa related to rapeseed (*B. napus*). For the

modelling, geographic occurrence data of these species were downloaded from the Global Biodiversity Information Facility (GBIF) (GBIF, 2023), with the exception of the five cultivated species (*Brassica carinata* A. Braun, *B. juncea*, *B. napus*, *B. oleracea* and *Raphanus sativus* L.) where it is impossible to distinguish natural occurrences from escapes from cultivation. Five taxa (*Crambe hispanica* subsp. *abyssinica* (Hochst. ex R.E.Fr.) Prina, *Erucastrum abyssinicum* R. E. Fr., *Orychophragmus violaceus* (L.) O.E. Schulz, *Physaria fendleri* (A. Gray) OKane & Al-Shehbaz and *Rorippa indica* (L.) Hiern) had no occurrence data within the studied region. From five taxa (*Brassica dimorpha* Coss. & Durieu, *Brassica deserti* Danin & Hedge, *Brassica desnottesii* Emb. & Maire, *Brassica souliei* Batt. subsp. *souliei* Batt. and *Brassica hilarionis* Post) the number of georeferenced locality data was insufficient for distribution modelling. *Brassica souliei* Batt. (excluding subsp. *amplexicaulis*) was used instead of *Brassica souliei* Batt. subsp. *souliei* Batt. Occurrence data of *Hirschfeldia incana* (L.) Lagr.-Foss. were downloaded using its synonym *Sinapis incana* L. Records from outside the studied region as well as records with missing or incorrect geographic information were removed. For nine taxa having low numbers of occurrences some additional records could be georeferenced using the locality descriptions and Google Earth. A spatial resolution, corresponding to a grid size of 2.5 min of a degree of longitude and latitude in the WorldClim dataset (Hijmans et al., 2005), was used to process the occurrence data. Multiple occurrence data per grid cell were reduced to one observation. To avoid spatial autocorrelation, only records separated by at least one grid cell were used for the distribution modelling, using seven bioclimatic variables (related to temperature and precipitation) and two soil variables (van Treuren et al., 2020). In the supplementary data, the downloads from GBIF and the number of grid cells used are given for each taxon¹. The manually georeferenced records have been made available as well. The R programming language (R_Core_Team, 2019) was used for distribution modelling with the Biomod2 package (Thuiller et al., 2009). Details of the modelling procedures are provided by Aguirre-Gutiérrez et al. (2017) and van Treuren et al. (2017). Predicted occurrences are solely based on the expected suitability of geographic locations as a result of the examined bioclimatic and soil variables. Other factors that may influence species occurrence, such as dispersal ability or geographic barriers, are not taken into account.

3 Results and discussion

Europe and especially the Mediterranean area provide a great richness of species. Coincidentally, the most critical collection gaps are related to this area (Castañeda-Álvarez et al., 2016). The Cruciferous (Brassicaceae) family currently comprises 338 genera and 3,709 species (Al-Shehbaz et al., 2006). Thereof, 39 accepted species names belong to the genus *Brassica* (Warwick and Francis,

2006). This number considerably increases when taking into account the large number of existing synonyms. The origins of the *Brassica* species are the area of the Mediterranean and southwest Asia (Al-Shehbaz et al., 2006). Many of the species of the genus *Brassica* are economically important, especially *B. rapa*, *B. nigra*, *B. oleracea*, *B. juncea*, *B. napus* and *B. carinata* (Cheng et al., 2014). This is also reflected by the number of holdings in European germplasm collections. The majority of the *Brassica* species are wild relatives which are not economically significant and therefore collected to a much lesser extent. However, they are of great importance for resistance to abiotic and biotic stresses and have the potential to improve resilience in modern cultivars (Quezada-Martinez et al., 2021). In addition, some of the wild *Brassic*as could even serve as a source of a new crop (Razzaq et al., 2021).

3.1 Data basis used

The present study is based on freely accessible data. The main basis is passport data from the EURISCO system, which provides comprehensive data on the majority of European genebank collections. EURISCO is an aggregator database that is unique in terms of the quantity and quality of data available and the underlying network. Nevertheless, there are limitations that need to be considered. Collections of plant genetic resources are sometimes very old and documented to varying degrees. Therefore, it cannot be assumed that there is complete information about the countries of origin and the sites where the genebank samples were found. For this reason, only those accessions could be considered for which collecting information is available. Despite all limitations, this still represents the best possible data available.

Furthermore, data on natural occurrence countries of the different *Brassica* species were used. These data are not necessarily complete either, as they depend heavily on the available literature sources. The fact that a country, in contrast to its immediate neighbours, is not listed as a natural area of origin does not necessarily mean that the corresponding material does not exist there, but only that it has not been described there so far. As mentioned above, an attempt was made to complement data from several sources.

3.2 Gene pool inventory and representation in genebanks for rapeseed

Results of the inventory are shown in Table 1. The rapeseed gene pool comprises 51 species, of which a total of 34,777 accessions are included in EURISCO. Subtaxa were not considered, as corresponding information is only available for a part of the accessions. No accessions of *B. deserti* and *P. fendleri* were found in European genebank collections. If the inventory is further restricted to accessions originating from the native occurrence areas and marked as collected material, the total number of accessions is 7,001. In this context, only areas known for native occurrences (represented by countries) were considered. Areas, in which those species were introduced or are being cultivated, were ignored.

¹ The supplementary file "Niche modelling data.zip" is available from Zenodo by the DOI 10.5281/zenodo.8081795.

TABLE 1 Genepool of *Brassica napus* L. and its representation in European genebanks based on EURISCO data.

Taxon	Accs. in European collections	Countries of native occurrences	Collected from native countries	IUCN Red List category
Primary gene pool				
<i>Brassica napus</i> L.	5,922	cultivated	0	–
Secondary gene pool				
<i>Brassica cretica</i> Lam.	2,399	Greece, Turkey	97	Least concern
<i>Brassica juncea</i> (L.) Czern.	2,479	cultivated	0	–
<i>Brassica rapa</i> L.	4,941	Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Estonia, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Malta, Morocco, Netherlands, Norway, Poland, Romania, Slovakia, Slovenia, Spain, Switzerland, Ukraine, United Kingdom	1,231	Data deficient
<i>Erucastrum gallicum</i> (Willd.) O. E. Schulz	28	Albania, Austria, Croatia, France, Italy, Netherlands, Slovenia, Spain, Switzerland	4	–
Tertiary gene pool				
<i>Brassica bourgeau</i> (Webb ex Christ) Kuntze	4	Spain*	4	–
<i>Brassica carinata</i> A. Braun	386	naturalised in Ethiopia; cultivated in Africa and Northern America*	287	–
<i>Brassica deserti</i> Danin & Hedge	0	Egypt	0	–
<i>Brassica desnottesii</i> Emb. & Maire	2	Morocco	1	–
<i>Brassica dimorpha</i> Coss. & Durieu	1	Algeria, Tunisia	1	–
<i>Brassica elongata</i> Ehrh.	14	Armenia, Austria, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Hungary, Morocco, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, North Macedonia, Turkey, Ukraine	4	Least concern
<i>Brassica fruticulosa</i> Cirillo	42	Algeria, Bosnia and Herzegovina, France, Italy, Malta, Morocco, Spain, Tunisia	26	Least concern
<i>Brassica gravinae</i> Ten.	7	Algeria, Italy, Libya, Morocco, Tunisia	3	Data deficient
<i>Brassica hilarionis</i> Post	5	Cyprus	3	Endangered
<i>Brassica incana</i> Ten.	48	Albania, Bosnia and Herzegovina, Croatia, Greece, Italy, Malta	39	Data deficient
<i>Brassica insularis</i> Moris	31	Algeria, France, Italy, Malta, Tunisia	31	Near threatened
<i>Brassica maurorum</i> Durieu	7	Algeria, Morocco	5	–
<i>Brassica montana</i> Pourr.	59	France, Italy, Spain	46	Least concern
<i>Brassica nigra</i> (L.) W. D. J. Koch	415	Belgium, Croatia, Cyprus, Egypt, Israel, Italy, Lebanon, Luxembourg, Montenegro, Netherlands, Spain, Syria	94	Least concern
<i>Brassica oleracea</i> L.	11,663	France, Germany, Italy, Spain, United Kingdom	3,173	Data deficient
<i>Brassica repanda</i> (Willd.) DC.	29	Algeria, France, Italy, Morocco, Spain, Switzerland	27	Least concern
<i>Brassica souliei</i> Batt. subsp. <i>souliei</i> Batt.	4	Algeria, Morocco, Tunisia	1	Data deficient

(Continued)

TABLE 1 Continued

Taxon	Accs. in European collections	Countries of native occurrences	Collected from native countries	IUCN Red List category
<i>Brassica souliei</i> Batt. subsp. <i>amplexicaulis</i> (Desf.) Greuter & Burdet	4	Italy, Malta, Morocco	2	Data deficient
<i>Brassica tournefortii</i> Gouan	126	Algeria, Cyprus, Egypt, Greece, Israel, Italy, Lebanon, Libya, Portugal, Spain, Syria, Morocco, Malta, Tunisia, Turkey	111	Least concern
<i>Capsella bursa-pastoris</i> (L.) Medik.	106	Albania, Algeria, Andorra, Armenia, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Libya, Liechtenstein, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Morocco, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovenia, Spain, Sweden, Switzerland, Tunisia, Turkey, Ukraine, United Kingdom	93	Least concern
<i>Crambe hispanica</i> subsp. <i>abyssinica</i> (Hochst. ex R.E.Fr.) Prina	169	Ethiopia, Kenya, Rwanda, Tanzania, Uganda, Democratic Republic of the Congo	2	Least concern
<i>Descurainia sophia</i> (L.) Webb ex Prantl	22	Albania, Algeria, Andorra, Armenia, Austria, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Egypt, Estonia, France, North Macedonia, Greece, Hungary, Iceland, Israel, Italy, Latvia, Liechtenstein, Lithuania, Malta, Moldova, Montenegro, Morocco, Netherlands, Portugal, Russian Federation, Serbia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom	6	–
<i>Diplotaxis acris</i> (Forsk.) Boiss.	21	Egypt, Israel, Turkey	20	–
<i>Diplotaxis catholica</i> (L.) DC.	13	Morocco, Portugal, Spain	13	Least concern
<i>Diplotaxis eruroides</i> (L.) DC.	49	Algeria, Egypt, France, Israel, Italy, Lebanon, Malta, Morocco, Portugal, Romania, Spain, Syria, Tunisia, Turkey	48	Least concern
<i>Diplotaxis harra</i> (Forssk.) Boiss.	27	Algeria, Egypt, Israel, Italy, Lebanon, Libya, Morocco, Spain, Syria, Tunisia	23	Least concern
<i>Diplotaxis muralis</i> (L.) DC.	18	Albania, Algeria, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Egypt, France, Greece, Hungary, Italy, Libya, Luxembourg, Malta, Moldova, Montenegro, Morocco, Netherlands, North Macedonia, Portugal, Romania, Russian Federation, Serbia, Slovenia, Spain, Switzerland, Tunisia, Turkey, Ukraine	12	Least concern
<i>Diplotaxis sifolia</i> Kunze	14	Algeria, Morocco, Portugal, Spain	14	Near threatened
<i>Diplotaxis tenuifolia</i> (L.) DC.	27	Albania, Andorra, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, France, Greece, Hungary, Italy, Liechtenstein, Luxembourg, Malta, Moldova, Montenegro, Netherlands, North Macedonia, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Switzerland, Turkey, Ukraine	18	Least concern
<i>Diplotaxis viminea</i> (L.) DC.	4	Albania, Algeria, Bulgaria, Croatia, Cyprus, Egypt, France, Greece, Israel, Italy, Malta, Morocco, Portugal, Romania, Spain, Tunisia, Turkey, Ukraine	3	Least concern
<i>Enarthrocarpus lyratus</i> (Forssk.) DC.	2	Egypt, Jordan	0	–
<i>Eruca vesicaria</i> (L.) Cav.	169	Algeria, Bulgaria, Croatia, Egypt, France, Greece, Hungary, Israel, Italy, Lebanon, Libya, Malta, Moldova, Morocco, Portugal, Romania, Spain, Switzerland, Syria, Tunisia, Turkey, Ukraine	110	Least concern
<i>Erucastrum abyssinicum</i> R. E. Fr.	2	Eritrea, Ethiopia, Yemen	0	–
<i>Hirschfeldia incana</i> (L.) Lagr.-Foss.	131	Albania, Algeria, Andorra, Armenia, Croatia, Cyprus, France, Greece, Israel, Italy, Lebanon, Malta, Morocco, Portugal, Spain, Syria, Tunisia, Turkey, Ukraine	111	–

(Continued)

TABLE 1 Continued

Taxon	Accs. in European collections	Countries of native occurrences	Collected from native countries	IUCN Red List category
<i>Moricandia arvensis</i> (L.) DC.	15	Algeria, Croatia, France, Greece, Italy, Malta, Montenegro, Morocco, Portugal, Spain, Tunisia	13	–
<i>Moricandia nitens</i> (Viv.) E. A. Durand & Barratte	15	Egypt, Israel, Jordan, Libya, Morocco, Tunisia	15	–
<i>Orychophragmus violaceus</i> (L.) O.E. Schulz	1	China, Korea	0	–
<i>Physaria fendleri</i> (A. Gray) OKane & Al-Shehbaz	0	Mexico, USA	0	–
<i>Raphanus raphanistrum</i> L.	236	Albania, Algeria, Armenia, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Egypt, Estonia, France, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lebanon, Libya, Liechtenstein, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Morocco, Netherlands, North Macedonia, Portugal, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Syria, Tunisia, Turkey, Ukraine, United Kingdom	189	Least concern
<i>Raphanus sativus</i> L.	3,550	Cyprus, Israel, Portugal, Spain	240	–
<i>Rapistrum rugosum</i> (L.) All.	30	Albania, Algeria, Andorra, Armenia, Austria, Bulgaria, Croatia, Cyprus, Egypt, France, Greece, Israel, Italy, Lebanon, Libya, Malta, Montenegro, Morocco, North Macedonia, Portugal, Russian Federation, Slovenia, Spain, Syria, Tunisia, Turkey, Ukraine	27	–
<i>Rorippa indica</i> (L.) Hiern	5	Egypt	0	–
<i>Rorippa islandica</i> (Oeder) Borb	8	Armenia, Austria, Bosnia and Herzegovina, Croatia, France, Greece, Iceland, Ireland, Italy, Liechtenstein, Montenegro, North Macedonia, Norway, Russian Federation, Slovenia, Spain, Switzerland, Turkey, Ukraine, United Kingdom	7	Least concern
<i>Sinapis alba</i> L.	1,372	Albania, Algeria, Belgium, Bulgaria, Croatia, Cyprus, Denmark, Egypt, France, Germany, Greece, Hungary, Israel, Italy, Lebanon, Libya, Luxembourg, Malta, Montenegro, Moldova, Morocco, Netherlands, Norway, Poland, Portugal, Romania, Spain, Sweden, Switzerland, Syria, Tunisia, Turkey, Ukraine, United Kingdom	739	Least concern
<i>Sinapis arvensis</i> L.	144	Albania, Algeria, Austria, Armenia, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Egypt, Estonia, France, Greece, Hungary, Israel, Italy, Latvia, Libya, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Morocco, Netherlands, Portugal, Serbia, Slovenia, Spain, Russian Federation, Tunisia, Turkey, Ukraine	99	Least concern
<i>Sinapis pubescens</i> L.	11	Albania, France, Italy, Algeria, Tunisia	9	Least concern

The countries of native occurrences were manually checked using the Euro+Med PlantBase. Entries with an asterisk (*) were complemented with data from GRIN Taxonomy.

The primary gene pool of rapeseed only consists of the cultivated species *B. napus*, which is represented by 5,922 accessions in European genebanks. These numbers include not only oil types but also 564 swede accessions (*B. napus* var. *napobrassica* (L.) Rchb.) as well as 65 Siberian kale accessions (*B. napus* var. *pabularia* (DC.) Rchb.).

The secondary gene pool comprises four species, including 9,847 accessions in total. *B. juncea* is a cultivated species, for which 2,479 accessions are preserved. With 2,399 accessions, *Brassica cretica* Lam. is represented with high numbers. However, only 97 of them were collected in native occurrence countries, corresponding to 4% of the total number of accessions of this species. For *Erucastrum gallicum* (Willd.) O. E. Schulz, only 28

accessions are preserved in European genebanks with four of them collected from native occurrence countries. In the case of *B. rapa*, which is partially cultivated, 1,231 (25%) out of 4,941 accessions were collected from native countries.

The tertiary gene pool comprises 46 species, which are represented by 19,008 accessions. Despite some highly represented species, such as the economically important *B. oleracea* (11,663 accessions), *B. nigra* (415 acc.), *B. carinata* (386 acc.) as well as *R. sativus* (3,550 acc.) and *S. alba* (1,372 acc.), the majority of species are only maintained in relatively low numbers in European genebanks (Table 1). 35% of the species of the tertiary gene pool are represented by less than 10 accessions and 50% by less than 20 accessions (Figure 1).

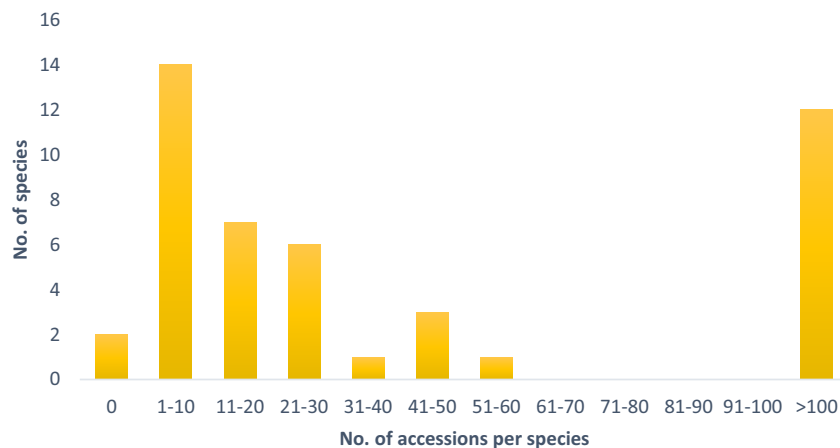


FIGURE 1

Number of accessions per species of the tertiary gene pool, which are maintained in European germplasm collections.

Even if the natural regions of origin are not considered, it is obvious that the majority of species of the *Brassica* gene pool are underrepresented in European genebanks although collecting missions were carried out in the 1970s (Razzaq et al., 2021). To solve this problem, a global strategy for the conservation of *Brassica* genetic resources was developed by the Global Crop Diversity Trust (Allender and Giovannini, 2023), confirming the importance of CWRs. Countries such as Italy have recognised the importance of *Brassica* CWRs and have established action plans (Ciancaleoni et al., 2021; Perrino and Wagensommer, 2022). Further collecting strategies for the whole of Europe need to be developed to improve *ex situ* conservation of at least some underrepresented species like *E. gallicum* from the secondary gene pool which has only four accessions from native countries but nine countries where the species naturally occurs. The CWRs are also important for other *Brassica* gene pools beside rapeseed because of the close evolutionary relationship. Rapeseed originated from an interspecific hybridisation of *B. rapa* and *B. oleracea* (Quezada-Martinez et al., 2021).

3.3 Conservation status

The Sampled Red List Index shows that approximately 22% of the plant species are threatened with extinction (Brummitt et al., 2015). Out of 1,350 European plant species, more than half are expected to be vulnerable or threatened by 2080 (Thuiller et al., 2005).

The species of the rapeseed gene pool were investigated for their conservation status. According to the IUCN Red List of Threatened Species, *B. hilarionis* is classified as endangered, while *Brassica insularis* Moris and *Diplotaxis siifolia* Kunze are classified as near threatened. With 5, 31 and 14 accessions, respectively, these species are maintained in low numbers in European collections. This refers only to the total accession numbers (34,777, see above); the natural occurrence countries are not taken into account here either. If these are

additionally included in the considerations, the underrepresentation is further aggravated.

In general, the numbers presented in Table 1 indicate that rapeseed gene pool accessions collected at native occurrence countries of the respective species are underrepresented in European genebanks. This situation is even worse when looking at the representation of individual countries belonging to the natural distribution area (Supplementary Table 1). For example, the near threatened species, *B. insularis* naturally occurs in five different countries, but was only collected in three of them. The near threatened species *D. siifolia* was collected in three out of four countries only, which is a cause for concern given the low number of accessions. *B. hilarionis* (endangered) is endemic to Cyprus and occurs nowhere else. Underrepresentation was also evident for species that currently are of least concern. For *B. elongata*, 17 countries were identified for native occurrences, but only in three of them accessions were collected. For *Descurainia sophia* (L.) Webb ex Prantl, 38 countries were identified, but accessions were only collected in six.

It should be noted here that the data situation regarding the endangerment status only allows limited statements to be made. Most species of the *Brassica* gene pool are not endangered or near threatened. 21 out of the 51 species are assigned to the IUCN Red List category of least concern. For six other species, data is indicated as deficient, while 21 species are not listed in the Red List at all. Only three species are endangered or near threatened, respectively. Nevertheless, it is important to improve the future conservation, as they play a major role in crop improvement (Raggi et al., 2022). In this context, of course, conservation under *in situ* conditions should not be ignored, especially for wild species. However, it must be taken into account that survival under *in situ* conditions is by no means guaranteed (e.g. due to climate change). In addition, access to *in situ* material is also difficult for users. A strong focus on *ex situ* conservation is therefore indispensable. Appropriate strategies need to be developed for this. Therefore, as a first step, we propose a priority list for the targeted collecting (see 3.5.)

3.4 Effects of climate change on species distribution

Figure 2 shows the results of the niche modelling for the wild species of the secondary gene pool. *B. juncea* was not considered because it is a cultivated species. *B. rapa* is not shown in Figure 2 since it is partly cultivated. In the case of *B. cretica* the simulation results show that the range remains almost constant at RCP 2.6 (expansion of 0.03%), but shrinks by 29.4% at RCP 8.5, assuming full migration potential. If there is no migration, however, the range will decrease by 33.9% (RCP 2.6) or 63.7% (RCP 8.5). Even more dramatic are the changes in *E. gallicum* (reduction of 29.5% and 71.5% respectively with migration; reduction of 40.6% and 89.4% respectively without migration) and *B. rapa* (reduction of 9.6% and 26.3% respectively with migration; reduction of 26.2% and 42.6% respectively without migration) (Table 2).

At this point, the three species listed in the IUCN Red List of Threatened Species as endangered or near threatened deserve closer examination. All three belong to the tertiary gene pool. For *B. hilarionis*, no niche modelling could be performed because the available occurrence data was insufficient. For *B. insularis*, the calculations indicate that the distribution area will decrease by 66.0% (RCP 2.6) or 78.0% (RCP 8.5) assuming full migration potential. Without migration, the decline will even be 66.5% (RCP 2.6) and 89.0% (RCP 8.5), respectively. In the case of *D. siifolia*, a significant reduction of the distribution area is also to be expected (reduction of 17.2% and 41.3% respectively with migration; reduction of 28.7% and 52.9% respectively without migration) (Figure 3). This is particularly dramatic against the background of the rather low number of germplasm accessions of these species in European collections. Modelling results of all species of the tertiary gene pool are shown in Table 2. In addition to the figures given there, the

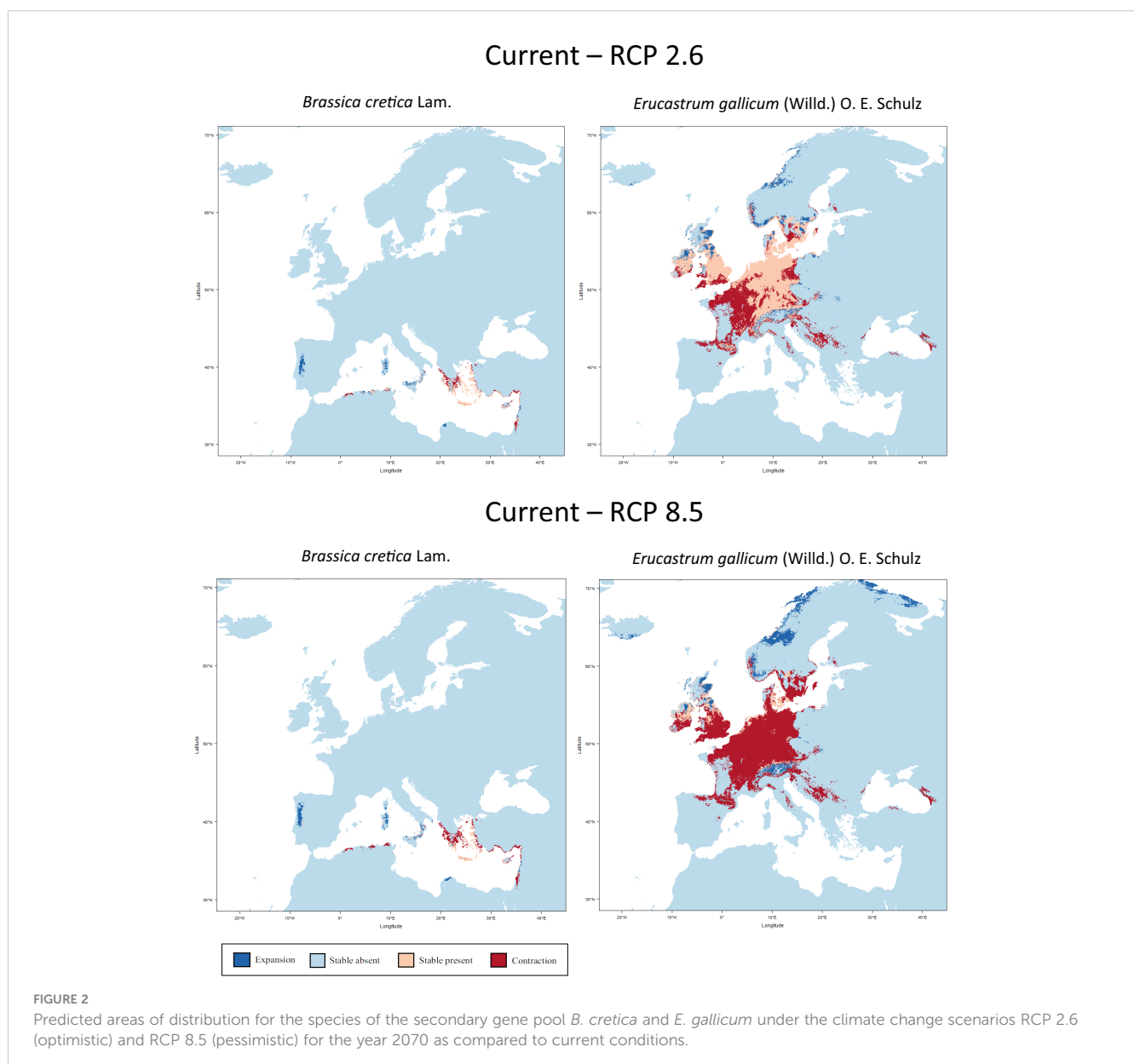


TABLE 2 Predicted changes of the distribution area of the species of the secondary and tertiary gene pool in 2070 under RCP 2.6 (optimistic scenario) and RCP 8.5 (pessimistic scenario), both with and without migration.

Taxon	Current range size	Range change with migration		Range change no migration	
		RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
Secondary gene pool					
<i>Brassica cretica</i> Lam.	6,769	0.03%	-29.4%	-33.9%	-63.7%
<i>Brassica juncea</i> (L.) Czern.		cultivated			
<i>Brassica rapa</i> L.	250,077	-9.6%	-26.3%	-26.2%	-42.6%
<i>Erucastrum gallicum</i> (Willd.) O. E. Schulz	103,289	-29.5%	-71.5%	-40.6%	-89.4%
Tertiary gene pool					
<i>Brassica bourgeau</i> i (Webb ex Christ) Kuntze	246	-5.3	-9.3	-5.3	-9.3
<i>Brassica carinata</i> A. Braun		Outside study area			
<i>Brassica deserti</i> Danin & Hedge		Only 3 GBIF records (2 after cleaning)			
<i>Brassica desnottesii</i> Emb. & Maire		Only 5 GBIF records (2 after cleaning)			
<i>Brassica dimorpha</i> Coss. & Durieu		Only 1 GBIF record (1 after cleaning)			
<i>Brassica elongata</i> Ehrh.	164,251	-15.4	-39.6	-38.4	-72.3
<i>Brassica fruticulosa</i> Cirillo	34,699	4.8	-4.6	-27.7	-60.4
<i>Brassica gravinae</i> Ten.	62,615	-10.3	-28.3	-31.4	-69.6
<i>Brassica hilarionis</i> Post		Only 5 GBIF records (3 after cleaning)			
<i>Brassica incana</i> Ten.	3,167	80.2	244.7	-19.1	-49.3
<i>Brassica insularis</i> Moris	9,277	-66.0	-78.0	-66.5	-89.0
<i>Brassica maurorum</i> Durieu	10,732	-36.9	-62.3	-41.1	-68.3
<i>Brassica montana</i> Pourr.	23,133	40.8	23.4	-38.7	-76.0
<i>Brassica nigra</i> (L.) W. D. J. Koch	159,081	-0.1	-5.8	-9.0	-26.1
<i>Brassica oleracea</i> L.		Cultivated			
<i>Brassica repanda</i> (Willd.) DC.	43,996	-46.2	-87.6	-50.8	-89.7
<i>Brassica souliei</i> Batt. subsp. <i>souliei</i> Batt.	32,540	-45.7	-80.8	-54.1	-85.8
<i>Brassica souliei</i> Batt. subsp. <i>amplexicaulis</i> (Desf.) Greuter & Burdet	5,685	-77.7	-96.7	-82.1	-97.5
<i>Brassica tournefortii</i> Gouan	73,804	-1.8	22.3	-21.9	-23.3
<i>Capsella bursa-pastoris</i> (L.) Medik.	280,915	-7.6	-24.4	-16.1	-32.9
<i>Crambe hispanica</i> subsp. <i>abyssinica</i> (Hochst. ex R.E.Fr.) Prina		Outside study area			
<i>Descurainia sophia</i> (L.) Webb ex Prantl	180,923	-17.5	-44.0	-34.6	-66.3
<i>Diplotaxis acris</i> (Forsk.) Boiss.	10,445	56.8	133.2	-23.2	-15.5
<i>Diplotaxis catholica</i> (L.) DC.	42,412	-6.1	-24.3	-17.7	-47.0
<i>Diplotaxis eruroides</i> (L.) DC.	68,478	49.2	37.7	-12.8	-36.1
<i>Diplotaxis harra</i> (Forssk.) Boiss.	47,471	39.1	84.0	-7.1	-4.9
<i>Diplotaxis muralis</i> (L.) DC.	132,933	2.3	-9.7	-12.9	-35.4
<i>Diplotaxis siifolia</i> Kunze	27,283	-17.2	-41.3	-28.7	-52.9
<i>Diplotaxis tenuifolia</i> (L.) DC.	127,278	-1.7	-14.5	-16.5	-42.9
<i>Diplotaxis viminea</i> (L.) DC.	77,775	29.3	19.1	-9.6	-28.6

(Continued)

TABLE 2 Continued

Taxon	Current range size	Range change with migration		Range change no migration	
		RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
<i>Enarthrocarpus lyratus</i> (Forssk.) DC.	2,013	-13.3	-16.7	-51.1	-52.9
<i>Eruca vesicaria</i> (L.) Cav.	133,801	-1.8	-19.4	-16.8	-38.5
<i>Erucastrum abyssinicum</i> R. E. Fr.		Outside study area			
<i>Hirschfeldia incana</i> (L.) Lagr.-Foss.	159,233	9.4	-1.6	-8.1	-24.3
<i>Moricandia arvensis</i> (L.) DC.	60,456	8.6	17.9	-19.8	-41.8
<i>Moricandia nitens</i> (Viv.) E. A. Durand & Barratte	25,965	-6.6	10.2	-29.9	-29.0
<i>Orychophragmus violaceus</i> (L.) O.E. Schulz		Outside study area			
<i>Physaria fendleri</i> (A. Gray) OKane & Al-Shehbaz		Outside study area			
<i>Raphanus raphanistrum</i> L.	255,467	2.6	1.3	-16.1	-24.6
<i>Raphanus sativus</i> L.		Cultivated			
<i>Rapistrum rugosum</i> (L.) All.	171,637	4.2	-5.9	-8.3	-24.5
<i>Rorippa indica</i> (L.) Hiern		Outside study area			
<i>Rorippa islandica</i> (Oeder) Borb	108,211	-31.8	-72.5	-41.1	-85.0
<i>Sinapis alba</i> L.	203,160	2.3	-0.3	-9.9	-24.0
<i>Sinapis arvensis</i> L.	299,489	-8.8	-25.3	-19.6	-36.1
<i>Sinapis pubescens</i> L.	38,339	-9.1	9.3	-24.1	-49.3

The current range sizes are given in numbers of cells (~4x4 km). Range changes are presented in percentage.

Supplementary Data also shows the predicted changes in the distribution areas for each species on maps.

Niche modelling shows the expected loss of distribution, but also the potential for new areas of occurrence. However, the dispersal ability of species or geographic barriers that restrict migration are not considered. For the studied region, it is not known whether the species are able to reach the areas where favourable climatic conditions prevail.

However, against the background of climate change and its likely effects in the whole of Europe, there is an urgent need for action, which was made evident by the niche modelling carried out. In this context, it should be noted that the modelling results of the optimistic scenario (RCP 2.6) are probably less likely compared to those of the pessimistic scenario (RCP 8.5). The long-term effects of global warming, which is taking place in the 21st century, will also have an impact on the following centuries (Schwalm et al., 2020; Lyon et al., 2022). The niche modelling used here offers the opportunity to react in advance to upcoming changes and to adjust the collection development of the European genebanks accordingly. Based on the results of niche modelling, the development of a collecting strategy for the endangered as well as underrepresented species is therefore urgently needed.

3.5 Implications for conservation

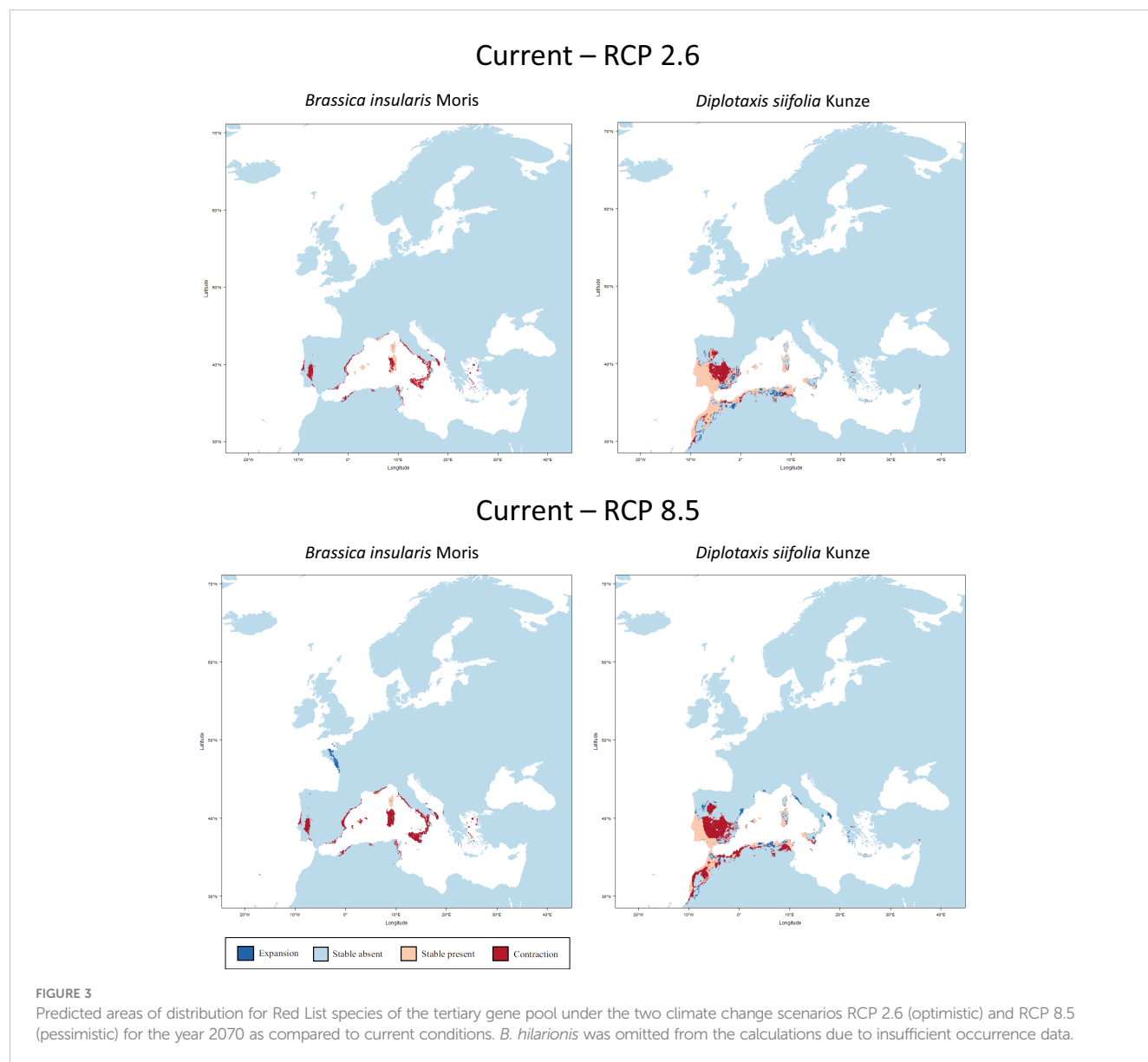
As described in the previous sections, the majority of species of the rapeseed gene pool are underrepresented in European

genebanks. This concerns in particular CWRs from the secondary and tertiary gene pool, which are also important for other *Brassica* species besides rapeseed due to the close evolutionary relationship between *B. napus*, *B. rapa* and *B. oleracea*. In addition, the CWRs have great potential for future crop improvement.

Based on the composition of the rapeseed gene pool, the information on the IUCN Red List status, the representation of the species in European genebanks (including consideration of the natural countries of origin) and the predicted effects of climate change on the future distribution ranges, we would like to propose a priority list for targeted collection. For this purpose, the species were assigned to three priority groups (high, medium, low). Cultivated species were excluded from this consideration. *B. rapa* is partly cultivated and was therefore also excluded from the priority list. All species listed as endangered or near threatened on the IUCN Red List were assigned to the highest priority.

For all other species, the considerations included how many accessions from the natural occurrence countries are maintained in European genebanks. As already noted in the introduction, a major difficulty in this context is that for decades there has been controversy about the minimum number of accessions required to maintain the natural diversity of a species. Maxted et al. (2020) summarises these discussions very well. In addition to the existing literature, many years of experience from practical genebank work were therefore also taken into account when drawing up the priority list.

Furthermore, the predicted future reduction of the natural distribution ranges was included in the considerations. Based on



the fact that human-induced global warming has increased at an unprecedented rate within the last ten years (Forster et al., 2023), we used the niche modelling results of the RCP 8.5 scenario and also assumed that no species migration takes place.

The procedure is described in detail in the [Supplementary File “Priority list creation.pdf”](#), the creation of the priority list in [Supplementary Table 2](#). The results are summarised in [Table 3](#), which lists species from the rapeseed gene pool that should be collected with priority in their natural occurrence countries.

As a result, 18 species of the rapeseed gene pool were assigned to the highest category of the priority list. Here, it is reasonable to target the natural areas of origin and to collect additional material to be maintained *ex situ* in genebanks.

Therefore, further conservation and collecting strategies need to be developed for Europe at large. It should be taken into account that, especially for wild species, *in situ* conservation can play an important additional role, but reliable *ex situ* conservation is

essential in any case. Data on the endangerment status can be used as a supplement in this context, but is not sufficient for a variety of species. An important role in the development of collecting strategies is taken by the expected effects of climate change on the natural distribution areas of the rapeseed gene pool species, as predicted by niche modelling. This makes it possible to react to future changes and to adapt the collection development of the genebanks accordingly.

4 Conclusion

In this study, we analysed the rapeseed gene pool and investigated to what extent the different species are conserved in European genebanks and which gaps exist. This also included the natural distribution ranges and it was found that most species of the rapeseed gene pool are significantly underrepresented in European

TABLE 3 Species from the rapeseed gene pool that should be collected with priority in their natural occurrence areas.

Taxon	Priority for collecting
<i>Brassica deserti</i> Danin & Hedge	high priority
<i>Brassica desnottesii</i> Emb. & Maire	high priority
<i>Brassica dimorpha</i> Coss. & Durieu	high priority
<i>Brassica elongata</i> Ehrh.	high priority
<i>Brassica fruticulosa</i> Cirillo	high priority
<i>Brassica gravinae</i> Ten.	high priority
<i>Brassica hilarionis</i> Post	high priority
<i>Brassica insularis</i> Moris	high priority
<i>Brassica maurorum</i> Durieu	high priority
<i>Brassica montana</i> Pourr.	high priority
<i>Brassica repanda</i> (Willd.) DC.	high priority
<i>Brassica souliei</i> Batt. subsp. <i>amplexicaulis</i> (Desf.) Greuter & Burdet	high priority
<i>Brassica souliei</i> Batt. subsp. <i>souliei</i> Batt.	high priority
<i>Descurainia sophia</i> (L.) Webb ex Prantl	high priority
<i>Diplotaxis siifolia</i> Kunze	high priority
<i>Enarthrocarpus lyratus</i> (Forssk.) DC.	high priority
<i>Erucastrum gallicum</i> (Willd.) O. E. Schulz	high priority
<i>Rorippa islandica</i> (Oeder) Borb	high priority
<i>Brassica bourgeau</i> (Webb ex Christ) Kuntze	medium priority
<i>Brassica cretica</i> Lam.	medium priority
<i>Brassica incana</i> Ten.	medium priority
<i>Brassica nigra</i> (L.) W. D. J. Koch	medium priority
<i>Capsella bursa-pastoris</i> (L.) Medik.	medium priority
<i>Diplotaxis acris</i> (Forsk.) Boiss.	medium priority
<i>Diplotaxis catholica</i> (L.) DC.	medium priority
<i>Diplotaxis eruroides</i> (L.) DC.	medium priority
<i>Diplotaxis harra</i> (Forssk.) Boiss.	medium priority
<i>Diplotaxis muralis</i> (L.) DC.	medium priority
<i>Diplotaxis tenuifolia</i> (L.) DC.	medium priority
<i>Diplotaxis viminea</i> (L.) DC.	medium priority
<i>Moricandia arvensis</i> (L.) DC.	medium priority
<i>Moricandia nitens</i> (Viv.) E. A. Durand & Barratte	medium priority
<i>Rapistrum rugosum</i> (L.) All.	medium priority
<i>Sinapis arvensis</i> L.	medium priority
<i>Sinapis pubescens</i> L.	medium priority
<i>Brassica tournefortii</i> Gouan	low priority
<i>Eruca vesicaria</i> (L.) Cav.	low priority
<i>Hirschfeldia incana</i> (L.) Lagr.-Foss.	low priority

genebanks. In addition, a niche modelling approach was used to investigate how the natural ranges of these species are likely to change by the end of the century under the assumption of various climate change scenarios. In some cases, considerable changes in the natural distribution areas were predicted. In order to close the existing gaps, a priority list was proposed. In addition to collecting trips, which are of course indispensable, sustainable conservation of CWRs requires a combination of *ex situ* and *in situ* efforts.

In general, various actions need to be taken to preserve CWRs, strategies are necessary to avoid loss of wild relatives of rapeseed: (1) Safeguarding the maintenance of all available CWR accessions in the genebanks in order to avoid further loss of material including regular regeneration and storage of a safety duplicate in Svalbard Global Seed Vault (2) Planning of collecting missions to increase the number of CWRs in genebanks for *ex situ* conservation. This should follow the priority list starting with the high priority species. (3) Undertaking also *in situ* conservation to increase the number of individuals in wild populations. (4) Protecting of natural habitats to prevent extinction. (5) Based on the niche modelling monitoring of the natural habitats. In case of loss due to climate change programmes for recolonisation or creation of new habitats.

Data availability statement

The datasets used for the niche modelling as well as the detailed results were published in an online repository. They are accessible by the DOI: 10.5281/zenodo.8081795.

Author contributions

SW coordinated the draft; SW, RH, KK, MO, RT and UL conceived and wrote the manuscript. All authors contributed to the article and approved the submitted version.

References

- Aguirre-Gutiérrez, J., van Treuren, R., Hoekstra, R., and van Hintum, T. J. L. (2017). Crop wild relatives range shifts and conservation in Europe under climate change. *Diversity Distribut.* 23 (7), 739–750. doi: 10.1111/ddi.12573
- Allard, R. W. (1970). "Population structure and sampling methods, in Genetic Resources in Plants," in *Their exploration and conservation*. Eds. O. H. Frankel and E. Bennet (Blackwell: Oxford and Edinburgh), 97–107.
- Allender, C., and Giovannini, P. (2023). *Global strategy for the conservation of brassica genetic resources* (Bonn, Germany: Global Crop Diversity Trust).
- Al-Shehbaz, I. A., Beilstein, M. A., and Kellogg, E. A. (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol.* 259 (2–4), 89–120. doi: 10.1007/s00606-006-0415-z
- Bakarr, M. I., Bennun, L. A., Brooks, T. M., Clay, R. P., Darwall, W. R. T., De Silva, N., et al. (2007). *Identification and gap analysis of key biodiversity areas: targets for comprehensive protected area systems. Best practice protected area guidelines series*. Ed. P. Valentine (Gland, Switzerland: IUCN), 116.
- Barthet, V. J. (2016). *Canola: overview, in encyclopedia of food grains. 2nd ed.* Ed. C. Wrigley, et al (Oxford: Academic Press), 237–241.
- Bell, J. M. (1982). From rapeseed to canola: A brief history of research for superior meal and edible oil. *Poultry Sci.* 61 (4), 613–622. doi: 10.3382/ps.0610613
- Brummitt, N., Bachman, S. P., Aletrari, E., Chadburn, H., Griffiths-Lee, J., Lutz, M., et al. (2015). The Sampled Red List Index for Plants, phase II: ground-truthing specimen-based conservation assessments. *Philos. Trans. R. Soc. B: Biol. Sci.* 370 (1662), 20140015. doi: 10.1098/rstb.2014.0015
- Burley, F. W. (1988). "Monitoring biological diversity for setting priorities in conservation," in *Biodiversity*. Eds. E. O. Wilson and F. M. Peter (Washington, DC: The National Academies Press), 227–230.
- Carré, P., and Pouzet, A. (2014). Rapeseed market, worldwide and in Europe. *OCL* 21 (1), D102. doi: 10.1051/ocl/2013054
- Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., et al. (2016). Global conservation priorities for crop wild relatives. *Nat. Plants* 2, 16022. doi: 10.1038/nplants.2016.22
- Chen, B.-Y., and Heneen, W. K. (1990). Resynthesized *Brassica napus* L.: A review of its potential in breeding and genetic analysis. *Hereditas* 111 (3), 255–263. doi: 10.1111/j.1601-5223.1990.tb00404.x
- Chen, H.-F., Wang, H., and Li, Z.-Y. (2007). Production and genetic analysis of partial hybrids in intertribal crosses between Brassica species (*B. rapa*, *B. napus*) and *Capsella bursa-pastoris*. *Plant Cell Rep.* 26 (10), 1791–1800. doi: 10.1007/s00299-007-0392-x
- Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of Brassica plants. *Horticul. Res.* 1, 14024. doi: 10.1038/hortres.2014.24
- Ciancaleoni, S., Raggi, L., Barone, G., Donnini, D., Gigante, D., Domina, G., et al. (2021). A new list and prioritization of wild plants of socioeconomic interest in Italy:

Funding

Costs for open access publishing were partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, grant 491250510).

Acknowledgments

We would like to thank Stefanie Kreide and Manuela Nagel for fruitful discussions.

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.2023.1244467/full#supplementary-material>

- toward a conservation strategy. *Agroecol. Sustain. Food Syst.* 45 (9), 1300–1326. doi: 10.1080/21683565.2021.1917469
- Crossa, J., Hernandez, C. M., Bretting, P., Eberhart, S. A., and Taba, S. (1993). Statistical genetic considerations for maintaining germ plasm collections. *Theor. Appl. Genet.* 86 (6), 673–678. doi: 10.1007/BF00222655
- CWR (2023). *Crop Wild Relative Inventory*. Available at: <https://www.cwrdiversity.org/checklist/> (Accessed Jan 2023).
- Dhillon, G. S., Kaur, S., Oberoi, H. S., Spier, M. R., and Brar, S. K. (2016). "Agricultural-based protein by-products: characterization and applications," in *Protein byproducts*. Ed. G. S. Dhillon (Amsterdam; Boston; Heidelberg; London; New York; Oxford; Paris; San Diego; San Francisco; Singapore; Sydney; Tokyo: Academic Press), 21–36. doi: 10.1016/B978-0-12-802391-4.00002-1
- Diederichsen, E., and Sacristan, M. D. (1996). Disease response of resynthesized *Brassica napus* L. lines carrying different combinations of resistance to *Plasmodiophora brassicae* Wor. *Plant Breed.* 115 (1), 5–10. doi: 10.1111/j.1439-0523.1996.tb00862.x
- Engels, J. M. M., and Maggioni, L. (2012). "AEGIS: a regionally based approach to PGR conservation," in *Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces*. Ed. N. Maxted, et al (Wallingford, UK; CAB), 321–326.
- EURISCO (2023). *European Search Catalogue for Plant Genetic Resources*. Available at: <http://eurisco.ecpgr.org/> (Accessed Jan 2023).
- Euro+Med (2023) *Euro+Med PlantBase - the information resource for Euro-Mediterranean plant diversity*. Available at: <https://www.europlusmed.org/> (Accessed Jan 2023).
- Fitt, B. D. L., Brun, H., Barbeti, M. J., and Rimmer, S. R. (2006). World-Wide Importance of Phoma Stem Canker (*Leptosphaeria maculans* and *L. biglobosa*) on Oilseed Rape (*Brassica napus*). *Eur. J. Plant Pathol.* 114 (1), 3–15. doi: 10.1007/s10658-005-2233-5
- Ford-Lloyd, B. V., Dulloo, E., and Toledo, A. (2011). Crop wild relatives - undervalued, underutilized and under threat? *BioScience* 61 (7), 559–565. doi: 10.1525/bio.2011.61.7.10
- Forster, P. M., Smith, C. J., Walsh, T., Lamb, W. F., Lamboll, R., Hauser, M., et al. (2023). Indicators of Global Climate Change 2022: annual update of large-scale indicators of the state of the climate system and human influence. *Earth Syst. Sci. Data* 15 (6), 2295–2327. doi: 10.5194/essd-15-2295-2023
- GBIF (2023) *Global Biodiversity Information Facility*. Available at: <https://www.gbif.org/> (Accessed Jan 2023).
- Gerdemann-Knörck, M., Sacristan, M. D., Braatz, C., and Schieder, O. (1994). Utilization of Asymmetric Somatic Hybridization for the Transfer of Disease Resistance from *Brassica nigra* to *Brassica napus*. *Plant Breed.* 113 (2), 106–113. doi: 10.1111/j.1439-0523.1994.tb00712.x
- Gilligan, C. A., Pechan, P. M., Day, R., and Hill, S. A. (1980). Beet western yellows virus on oilseed rape. *Plant Pathol.* 29 (1), 53. doi: 10.1111/j.1365-3059.1980.tb01138.x
- Girke, A., Schierholt, A., and Becker, H. C. (2012). Extending the rapeseed gene pool with resynthesized *Brassica napus* L. I: Genetic diversity. *Genet. Resour. Crop Evol.* 59 (7), 1441–1447. doi: 10.1007/s10722-011-9772-8
- GRIN (2023) *Taxonomic information on cultivated plants in GRIN-Global*. Available at: <https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearch/> (Accessed Jan 2023).
- Grusak, M. A., and DellaPenna, D. (1999). Improving the nutrient composition of plants to enhance human nutrition and health. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50 (1), 133–161. doi: 10.1146/annurev.arplant.50.1.133
- Hajjar, R., and Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* 156 (1–2), 1–13. doi: 10.1007/s10681-007-9363-0
- Harlan, J. R., and de Wet, J. M. J. (1971). Toward a rational classification of cultivated plants. *Taxon* 20 (4), 509–517. doi: 10.2307/1218252
- Heale, J. B., and Karapapa, V. K. (1999). The verticillium threat to Canada's major oilseed crop: canola. *Can. J. Plant Pathol.* 21 (1), 1–7. doi: 10.1080/07060661.1999.10600114
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25 (15), 1965–1978. doi: 10.1002/joc.1276
- Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S., et al. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? *Proc. Natl. Acad. Sci.* 96 (11), 5937–5943. doi: 10.1073/pnas.96.11.5937
- Hu, D., Jing, J., Snowdon, R. J., Mason, A. S., Shen, J., Meng, J., et al. (2021). Exploring the gene pool of *Brassica napus* by genomics-based approaches. *Plant Biotechnol. J.* 19 (9), 1693–1712. doi: 10.1111/pbi.13636
- Hwang, S.-F., Strelkov, S. E., Peng, G., Ahmed, H., Zhou, Q., Turnbull, G., et al. (2016). Blackleg (*Leptosphaeria maculans*) severity and yield loss in canola in Alberta, Canada. *Plants* 5 (3), 31. doi: 10.3390/plants5030031
- IUCN (2023) *The IUCN Red List of threatened species*. Available at: <https://www.iucnredlist.org/> (Accessed Jan 2023).
- Jahreis, G., and Schäfer, U. (2011). "Rapeseed (*Brassica napus*) Oil and its Benefits for Human Health," in *Nuts and seeds in health and disease prevention*. Eds. V. R. Preedy, R. R. Watson and V. B. Patel (San Diego: Academic Press), 967–974.
- Jarvis, A., Lane, A., and Hijmans, R. J. (2008). The effect of climate change on crop wild relatives. *Agricul. Ecosyst. Environ.* 126 (1–2), 13–23. doi: 10.1016/j.agee.2008.01.013
- Jennings, M. D. (2000). Gap analysis: concepts, methods, and recent results. *Landscape Ecol.* 15 (1), 5–20. doi: 10.1023/A:1008184408300
- Kotni, P., vanÂ Hintum, T., Maggioni, L., Oppermann, M., and Weise, S. (2023). EURISCO update 2023: the European Search Catalogue for Plant Genetic Resources, a pillar for documentation of genebank material. *Nucleic Acids Res.* 51 (D1), D1465–D1469. doi: 10.1093/nar/gkac852
- Lawrence, M. J., Marshall, D. F., and Davies, P. (1995). Genetics of genetic conservation. I. Sample size when collecting germplasm. *Euphytica* 84 (2), 89–99. doi: 10.1007/BF01677945
- Link, K. (2008). "Nutzung genetischer Diversität in Raps (*Brassica napus*) für Assoziationsstudien zur Resistenz gegen die Wurzelhals- und Stängelfäule (*Leptosphaeria maculans*)," in *Institut für Pflanzenbau und Pflanzenzüchtung I.* (Germany: Giessen University).
- Lyon, C., Saupe, E. E., Smith, C. J., Hill, D. J., Beckerman, A. P., Stringer, L. C., et al. (2022). Climate change research and action must look beyond 2100. *Global Change Biol.* 28 (2), 349–361. doi: 10.1111/gcb.15871
- Margules, C. R., and Pressey, R. L. (2000). Systematic conservation planning. *Nature* 405, 243–253. doi: 10.1038/35012251
- Marshall, D. R., and Brown, H. D. (1975). "Optimum sampling strategies in genetic conservation," in *Crop genetic resources for today and tomorrow*. Eds. O. H. Frankel and J. G. Hawkes (Cambridge, UK: Cambridge University Press), 53–80.
- Maxted, N., Dulloo, E., Ford-Lloyd, B. V., Iriondo, J. M., and Jarvis, A. (2008). Gap analysis: a tool for complementary genetic conservation assessment. *Diversity Distribut.* 14 (6), 1018–1030. doi: 10.1111/j.1472-4642.2008.00512.x
- Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E., and Toledo, A. (2012). Toward the systematic conservation of global crop wild relative diversity. *Crop Sci.* 52 (2), 774–785. doi: 10.2135/cropsci2011.08.0415
- Maxted, N., Hunter, D., and Ortiz Rios, R. (2020). *Plant genetic conservation* (Cambridge: Cambridge University Press).
- Maxted, N., van Slageren, M. W., and Rihan, J. (1995). *Ecogeographic surveys, in Collecting plant genetic diversity: technical guidelines*. Eds. L. Guarino, V. R. Rao and R. Reid (Wallingford: CAB International), 255–286.
- McCouch, S. (2004). Diversifying selection in plant breeding. *PLoS Biol.* 2 (10), e347. doi: 10.1371/journal.pbio.0020347
- Metzger, J. O., and Bornscheuer, U. (2006). Lipids as renewable resources: current state of chemical and biotechnological conversion and diversification. *Appl. Microbiol. Biotechnol.* 71 (1), 13–22. doi: 10.1007/s00253-006-0335-4
- Nagel, M., Dulloo, M. E., Bissessur, P., Gavrilenko, T., Bamberg, J., Ellis, D., et al. (2022). *Global strategy for the conservation of potato* (Bonn, Germany: Global Crop Diversity Trust).
- Neuffer, B. (2001). *Vom Wildkohl zum Kohlrabi: Nutzpflanzen der Kreuzblütler in der Antike und heute. Schriftenreihe des Botanischen Gartens Osnabrück* (Osnabrueck, Germany: Botanischer Garten der Universität Osnabrück).
- Parra-Quijano, M., Iriondo, J. M., and Torres, E. (2012). Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodiversity Conserv.* 21 (1), 79–96. doi: 10.1007/s10531-011-0167-0
- Perrino, E. V., and Wagensommer, R. P. (2022). Crop wild relatives (CWRs) threatened and endemic to Italy: urgent actions for protection and use. *Biology* 11 (2), 193. doi: 10.3390/biology11020193
- Piazza, G. J., and Foglia, T. A. (2001). Rapeseed oil for oleochemical usage. *Eur. J. Lipid Sci. Technol.* 103 (7), 450–454. doi: 10.1002/1438-9312(200107)103:7<450::AID-EJLT450>3.0.CO;2-D
- Qian, B., Jing, Q., Bélanger, G., Shang, J., Huffman, T., Liu, J., et al. (2018). Simulated canola yield responses to climate change and adaptation in Canada. *Agron. J.* 110 (1), 133–146. doi: 10.2134/agronj2017.02.0076
- Quezada-Martinez, D., Addo Nyarko, C. P., Schiessl, S. V., and Mason, A. S. (2021). Using wild relatives and related species to build climate resilience in *Brassica* crops. *Theor. Appl. Genet.* 134 (6), 1711–1728. doi: 10.1007/s00122-021-03793-3
- R_Core_Team (2019). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).
- Raggi, L., Zucchini, C., Gigante, D., and Negri, V. (2022). *In situ* occurrence and protection of crop wild relatives in Italian sites of natura 2000 network: Insights from a data-driven approach. *Front. Plant Sci.* 13, 1080615. doi: 10.3389/fpls.2022.1080615
- Razzaq, H., Armstrong, S. J., and Saleem, H. (2021). "Brassicas: A complete guide to the potential of their wild relatives," in *Wild germplasm for genetic improvement in crop plants*. Eds. M. T. Azhar and S. H. Wani (Elsevier, Academic Press), 187–199. doi: 10.1016/B978-0-12-822137-2.00010-2
- Rygulla, W., Snowdon, R. J., Eynck, C., Koopmann, B., von Tiedemann, A., Lühs, W., et al. (2007). Broadening the Genetic Basis of Verticillium longisporium Resistance in *Brassica napus* by Interspecific Hybridization. *Phytopathology* 97 (11), 1391–1396. doi: 10.1094/PHYTO-97-11-1391
- Schwalm, C. R., Glendon, S., and Duffy, P. B. (2020). RCP8.5 tracks cumulative CO₂ emissions. *Proc. Natl. Acad. Sci.* 117 (33), 19656–19657. doi: 10.1073/pnas.2007117117
- Singh, S. P. (2001). Broadening the genetic base of common bean cultivars: A review. *Crop Sci.* 41 (6), 1659–1675. doi: 10.2135/cropsci2001.1659
- Snowdon, R. J., Winter, H., Diestel, A., and Sacristan, M. D. (2000). Development and characterisation of *Brassica napus*-*Sinapis arvensis* addition lines exhibiting resistance to *Leptosphaeria maculans*. *Theor. Appl. Genet.* 101 (7), 1008–1014. doi: 10.1007/s001220051574

- Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J., et al. (2013). *Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (United Kingdom and New York, NY, USA: Cambridge University Press).
- Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., and Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proc. Natl. Acad. Sci. U.S.A.* 102 (23), 8245–8250. doi: 10.1073/pnas.0409902102
- Thuiller, W., Lafourcade, B., Engler, R., and Araújo, M. B. (2009). BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* 32 (3), 369–373. doi: 10.1111/j.1600-0587.2008.05742.x
- Tilman, D., Hill, J., and Lehman, C. (2006). Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science* 314 (5805), 1598. doi: 10.1126/science.1133306
- USDA (2023). *Oilseeds: world markets and trade* (Washington DC: United States Department of Agriculture, Foreign Agricultural Service).
- van Treuren, R., Coquin, P., and Lohwasser, U. (2012). Genetic resources collections of leafy vegetables (lettuce, spinach, chicory, artichoke, asparagus, lamb's lettuce, rhubarb and rocket salad): composition and gaps. *Genet. Resour. Crop Evol.* 59 (6), 981–997. doi: 10.1007/s10722-011-9738-x
- van Treuren, R., Hoekstra, R., Wehrens, R., and van Hintum, T. (2020). Effects of climate change on the distribution of crop wild relatives in the Netherlands in relation to conservation status and ecotope variation. *Global Ecol. Conserv.* 23, e01054. doi: 10.1016/j.gecco.2020.e01054
- van Treuren, R., Hoekstra, R., and van Hintum, T. J. L. (2017). Inventory and prioritization for the conservation of crop wild relatives in The Netherlands under climate change. *Biol. Conserv.* 216, 123–139. doi: 10.1016/j.biocon.2017.10.003
- van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., and Hibbard, K. (2011). The representative concentration pathways: an overview. *Climatic Change* 109 (1), 5. doi: 10.1007/s10584-011-0148-z
- Vavilov, N. I. (1926). *Studies in the origin of cultivated plants* (Leningrad: Institute of Applied Botany and Plant Breeding), 248.
- Vincent, H., Wiersema, J., Kell, S., Fielder, H., Dobbie, S., Castañeda-Álvarez, N. P., et al. (2013). A prioritized crop wild relative inventory to help underpin global food security. *Biol. Conserv.* 167, 265–275. doi: 10.1016/j.biocon.2013.08.011
- Vollbrecht, E., and Sigmon, B. (2005). Amazing grass: developmental genetics of maize domestication. *Biochem. Soc. Trans.* 33 (6), 1502–1506. doi: 10.1042/BST0331502
- Warwick, S. I., and Francis, A. (2006). I.A. Al-Shehbaz, Brassicaceae: Species checklist and database on CD-Rom. *Plant Syst. Evol.* 259 (2-4), 249–258. doi: 10.1007/s00606-006-0422-0
- Wei, W., Li, Y., Wang, L., Liu, S., Yan, X., Mei, D., et al. (2010). Development of a novel *Sinapis arvensis* disomic addition line in *Brassica napus* containing the restorer gene for Ns_a CMS and improved resistance to *Sclerotinia sclerotiorum* and pod shattering. *Theor. Appl. Genet.* 120 (6), 1089–1097. doi: 10.1007/s00122-009-1236-6
- Weise, S., Oppermann, M., Maggioni, L., van Hintum, T., and Knüpfer, H. (2017). EURISCO: The European search catalogue for plant genetic resources. *Nucleic Acids Res.* 45 (D1), D1003–D1008. doi: 10.1093/nar/gkw755
- WFO (2023). *World Flora Online*. Available at: <http://www.worldfloraonline.org/> (Accessed Jan 2023).
- Williams, I. H. (2010). “The major insect pests of oilseed rape in Europe and their management: an overview,” in *Biocontrol-based integrated management of oilseed rape pests*. Ed. I. H. Williams (Dordrecht: Springer Netherlands), 1–43.



OPEN ACCESS

EDITED BY

Maarten Van Zonneveld,
World Vegetable Center, Taiwan

REVIEWED BY

Toi J. Tsilo,
Agricultural Research Council of South
Africa (ARC-SA), South Africa
Sebastian Beier,
Forschungszentrum Jülich GmbH,
Germany
Hakan Ozkan,
Çukurova University, Türkiye

*CORRESPONDENCE

Jesse Poland
✉ jesse.poland@kaust.edu.sa

RECEIVED 27 July 2023

ACCEPTED 25 September 2023

PUBLISHED 17 October 2023

CITATION

Adhikari L, Raupp J, Wu S, Koo D-H,
Friebe B and Poland J (2023) Genomic
characterization and gene bank curation of
Aegilops: the wild relatives of wheat.
Front. Plant Sci. 14:1268370.
doi: 10.3389/fpls.2023.1268370

COPYRIGHT

© 2023 Adhikari, Raupp, Wu, Koo, Friebe and
Poland. 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.

Genomic characterization and gene bank curation of *Aegilops*: the wild relatives of wheat

Laxman Adhikari^{1,2}, John Raupp², Shuangye Wu²,
Dal-Hoe Koo², Bernd Friebe² and Jesse Poland^{1,2,3*}

¹Plant Breeding and Genetics Lab, Center for Desert Agriculture, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, ²Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS, United States, ³Plant Science Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Genetic diversity found in crop wild relatives is critical to preserve and utilize for crop improvement to achieve sustainable food production amid climate change and increased demand. We genetically characterized a large collection of 1,041 *Aegilops* accessions distributed among 23 different species using more than 45K single nucleotide polymorphisms identified by genotyping-by-sequencing. The Wheat Genetics Resource Center (WGRC) *Aegilops* germplasm collection was curated through the identification of misclassified and redundant accessions. There were 49 misclassified and 28 sets of redundant accessions within the four diploid species. The curated germplasm sets now have improved utility for genetic studies and wheat improvement. We constructed a phylogenetic tree and principal component analysis cluster for all *Aegilops* species together, giving one of the most comprehensive views of *Aegilops*. The *Sitopsis* section and the U genome *Aegilops* clade were further scrutinized with in-depth population analysis. The genetic relatedness among the pair of *Aegilops* species provided strong evidence for the species evolution, speciation, and diversification. We inferred genome symbols for two species *Ae. neglecta* and *Ae. columnaris* based on the sequence read mapping and the presence of segregating loci on the pertinent genomes as well as genetic clustering. The high genetic diversity observed among *Aegilops* species indicated that the genus could play an even greater role in providing the critical need for untapped genetic diversity for future wheat breeding and improvement. To fully characterize these *Aegilops* species, there is an urgent need to generate reference assemblies for these wild wheats, especially for the polyploid *Aegilops*.

KEYWORDS

Aegilops, genotyping-by-sequencing (GBS), gene bank curation, genetic diversity, phylogenetic analysis, population structure, wheat wild relatives

1 Introduction

Global climate change with increasingly variable weather, declining soil quality, and increased biotic and abiotic stresses impede crop production. For instance from crop modeling, an increase in a global mean temperature of a degree Celsius reduces the global wheat yield by 6% (Asseng et al., 2015; Zhao et al., 2017). In this context, the continual genetic improvement of commercial cultivars is needed, including incorporating novel alleles for improved stress tolerance and disease resistance. However, the domestication bottleneck and variety selection practices are major drivers that limit the genetic diversity currently available in the primary gene pool for wheat (*Triticum aestivum* L.) improvement (Haudry et al., 2007). Several studies have indicated that wild wheat relatives are reliable sources for increasing the genetic diversity in wheat breeding (Lopes et al., 2015; Leigh et al., 2022; Ahmed et al., 2023).

The genus *Aegilops* encompasses the secondary and tertiary gene pool of bread wheat with a central role in wheat evolution and domestication being the donors of B and D subgenomes. The *Aegilops* species are critically important in providing biotic resistance and abiotic tolerance as well as yield-related genetic loci to wheat (Kishii, 2019; Rakszegi et al., 2020). For instance, *Ae. speltoides* harbors agronomically important genes, such as *Sr32* which is effective against the devastating wheat stem rust pathogen Ug99 (Friebe et al., 1996). Similarly, *Ae. kotschyi* has been shown to confer leaf and stripe rust resistance with genes *Lr54* and *Yr37* (Marais et al., 2005), and *Ae. biuncialis* possesses a wheat powdery mildew resistance gene (Li et al., 2019). Likewise, the 2NS translocation from *Ae. ventricosa* provided multiple disease resistance including root-knot nematode, stripe rust, stem rust, leaf rust, and the wheat blast caused by *Magnaporthe oryzae* (Cruz et al., 2016; Gao et al., 2021). Finally, *Ae. tauschii* has been frequently used in wheat breeding as the genetic resource for various wheat disease resistance and abiotic-stress tolerance (Suneja et al., 2019).

Although *Aegilops* species hold great potential as genetic resources, limited information is available on the genomic characterization of the genus as a whole. Most of the work to date has focused on a limited number of *Aegilops* species and has been based on cytology, traditional molecular markers, and a limited number of loci. Genomic characterization is complex, because *Aegilops* species have various ploidy levels and unique genomic compositions and some polyploids have multiple copies of the same sub-genome [e.g., DDM, 6X *Ae. crassa*]. Also, reference genomes for only a few *Aegilops* species have been released to date. Therefore, the complicated genomic features and inadequate resources are major challenges for *Aegilops* population studies and more focused, targeted mining of the genetic resources.

These limitations are quickly changing with the recently available genome assemblies of some diploid *Aegilops* such as *Ae. tauschii* (Luo et al., 2017), *Ae. speltoides* and *Ae. longissima* (Avni et al., 2022), *Ae. sharonensis* (Yu et al., 2022), *Ae. bicornis*, and *Ae. searsii* (Li et al., 2022). These genome assemblies are shedding light on *Aegilops*' evolutionary and population genetic analysis. Additionally, the high-throughput sequencing method such as genotyping-by-sequencing (GBS), which can generate *de-novo*

genomics variants for complex genome species (Poland et al., 2012), has also been proven as an efficient genotyping tool for gene bank collections (Adhikari et al., 2022a).

The Wheat Genetics Resource Center (WGRC) gene bank at Kansas State University has been maintaining myriads of wild wheat accessions under the *Triticum* and *Aegilops* genera. We previously curated the collections of A-genome diploid wheat (Adhikari et al., 2022a) and *Ae. tauschii* (Singh et al., 2019a). Thus, the focus of this current study was to characterize the genetic diversity, population structure, and genomic composition of the *Aegilops* collection in the WGRC with the curation of the germplasm. Throughout this study, we followed the *Aegilops* species nomenclature by Van Slageren (1994) except for *Ae. mutica*, and genome symbols were followed as described by Waines and Barnhart (1992). Utilizing variants from GBS, we dissected the genetic and genomic relationships among the 23 *Aegilops* species through phylogenetic clustering, principal component analysis (PCA), population structure analysis, and diversity analysis. We also examined *Aegilops* and wheat genomes relationships through *Aegilops* sequence mapping to the wheat genome and genetic clustering.

2 Materials and methods

2.1 Plant resources

This study primarily included 1,041 accessions of the *Aegilops* species preserved and maintained in the WGRC gene bank (Supplementary Material Table S1; Figure 1). The accessions were originally collected from various sources and sites including the Middle East, Anatolia, East Asia, and northern Africa (Figure 1; Supplementary Material Table S1). Accessions comprise 22 different *Aegilops* species under five sections (*Aegilops*, *Comopyrum*, *Cylindricum*, *Sitopsis*, and *Vertebrata*) (Van Slageren, 1994) and *Ae. mutica*, which is synonymously known as *Amblopyrum muticum*. For gene bank curation and most part of the population analysis, only those *Ae. tauschii* accessions that were not in the previous gene bank curation experiment (Singh et al., 2019a) were used. We also used CIMMYT wheat lines and already curated *Ae. tauschii* lines (Supplementary Material Table S1) for genotyping together with the diploid *Aegilops* to dissect the genetic relationships among wheat and *Aegilops* genomes.

Most of these species are self-pollinated and were primarily maintained by single seed descent, with exceptions described below. *Ae. speltoides* and *Ae. mutica* are partially out-crossing and were maintained through sib-mating multiple plants. Specifically, *Ae. mutica* accessions consisted of 54 samples from five out-crossing plants bulked together.

2.2 Genotyping and marker identification

The DNA extraction, GBS library preparation, and sequencing were performed as we described in our earlier studies (Adhikari et al., 2022a) using two enzyme-based GBS (Poland et al., 2012). Only a single plant per accession was sequenced for all species except *Ae. mutica*, where we sequenced 54 individuals obtained

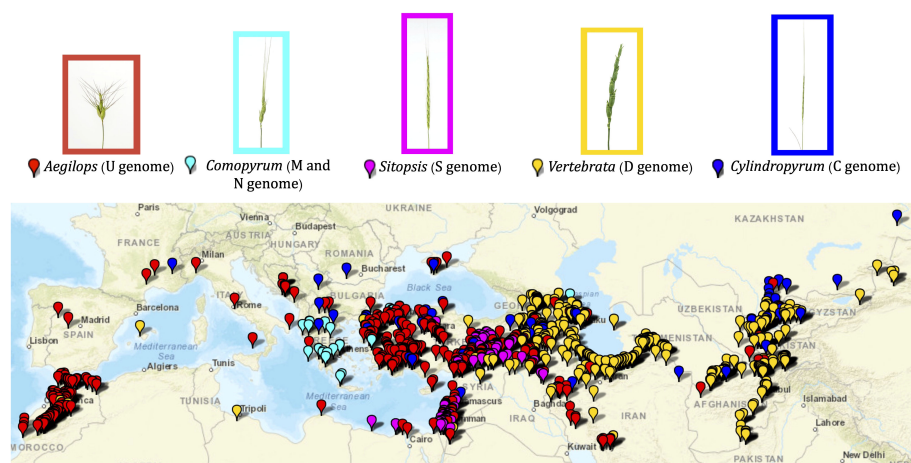


FIGURE 1

Geographic distribution of the *Aegilops* accessions maintained in the WGRC gene bank. Spike morphologies of representative accessions for the five *Aegilops* sections are shown with the enclosed rectangles. Each section is designated by corresponding color.

from randomly crossing five plants, because the species is cross-pollinating and it has a low germination rate.

For the *de-novo* single nucleotide polymorphism (SNP) calling, reads were demultiplexed using *sabre* (<https://github.com/najoshi/sabre>) and adapters were trimmed using *fastp* (Chen et al., 2018). The variants were called using the available reference assemblies of diploid *Aegilops* and wheat and using mock references generated as described (Melo et al., 2016; Adhikari et al., 2018). For mock references, the raw GBS reads of selected accessions with higher sequence data were used as the reference source. We also ensured that the mock reference represents the sequences of relevant *Aegilops* species or the genomes [C, D, M, N, S, U, T] for the population to be genotyped. The *de-novo* variants were called using *BCFtools* (Li, 2011) and used for initial gene bank curation and population clustering of the whole collection. Then the *de-novo* variants were also called for some species independently depending on the objectives of the specific analysis (Supplementary Material Table S2). For some species in polyploid lineages, we called variants on a diploid ancestor and, later, the same variants were called in the polyploids using *BCFtools* (Li, 2011). After calling variants, unless otherwise stated, we filtered loci to keep any variants passing these conditions: minor allele frequency (MAF) >0.01, missing <30%, and heterozygous <10%.

The TASSEL5 GBSv2 pipeline was used for reference-based SNP calling (Glaubitz et al., 2014). For this method, *Ae. tauschii* reference genome Aet v5.0 (Wang et al., 2021) or *Ae. sharonensis* (Yu et al., 2022), *Ae. speltoides* (Avni et al., 2022), *Ae. searsii*, and *Ae. bicornis* (Li et al., 2022) genomes were used. We also called variants in all these diploids species to the wheat reference using the “Chinese Spring” wheat reference (IWGSC CS RefSeq v2.1) (Zhu et al., 2021) to observe the relationship between *Aegilops* and wheat.

2.3 Gene bank curation

In the first step, the germplasm curation identified misclassified accessions and corrected the taxonomy of these accessions in the

database (Singh et al., 2019a). We identified misclassified accessions by constructing a phylogenetic cluster colored with the recorded species. These were further verified using PCA clustering followed by a visual assessment of seeds and spikes. The misclassified accessions were identified and confirmed with multiple genotyping sets *viz.* entire collection, species alone, and same genome accessions together.

In the second step, the genetically identical accessions were determined using allele matching (Singh et al., 2019a; Adhikari et al., 2022a). However, this assessment was done only for the accessions of the species whose reference genome is available, for example, *Ae. tauschii* and the *Sitopsis* section *Aegilops*. The allele matching (>99% identity by state) was used as a threshold to confirm genetically identical accessions. Allele matching used homozygous and non-missing sites between two given accessions, and the raw markers were filtered using MAF >0.01, missing <50%, and heterozygous <20% parameters before allele matching. We conducted further examinations of the sets of genetic duplicates to assess their phenotypic similarities, collection sites, and sources of collection.

2.4 Genetic clustering, population analysis, and diversity

The genotyping matrices were analyzed for the genetic distances among the *Aegilops* populations, which were then used for exploring the population structure and ancestry. For phylogenetic clustering, the genetic distance was computed using the “dist” function in R (R Core Team, 2020), and the R packages *ape* (Paradis and Schliep, 2019) and *phyclust* (Chen, 2011) were then used to generate unrooted neighbor-joining (NJ) tree with the default parameters (Singh et al., 2019b; Adhikari et al., 2022a).

The genetic relationships among the *Aegilops* accessions were further examined via PCA, which was performed in two steps. The A matrix was derived from *A.mat()* function within the R package

rrBLUP (Endelman, 2011), and the eigenvalues and eigenvectors were derived using the “e” function (Adhikari et al., 2022a). Furthermore, the population structure of the *Sitopsis* group of *Aegilops* was also performed with the reference-based genotyping profile using fastStructure software (Raj et al., 2014) as explained (Adhikari et al., 2022a). We computed Nei's diversity index (Nei, 1987) and total segregating loci for each of the *Aegilops* species to assess the relative diversity of the species.

2.5 *Ae. columnaris* and *Ae. neglecta* genome symbols

We investigated the traditional genome symbols of *Ae. columnaris* (UM) and *Ae. neglecta* (UM, UMN) for the presence/absence of the M genome. There are recent cytology-based findings that have questioned the traditional genome symbols of these species (Badaeva et al., 2018). To test this, we computed the sequence read mapping and segregating loci on the M and U mock reference genomes for the *Ae. columnaris* and *Ae. neglecta* accessions as well as two other tetraploids (*Ae. neglecta* and *biuncialis*) whose genomic compositions are unequivocally recognized as MU or UM. The *de-novo* variants were first identified for the diploid M genome (*Ae. comosa*) and U genome (*Ae. umbellulata*) populations separately, and then the same variants were called on these four tetraploid species. We also constructed the phylogenetic clustering among *Ae. columnaris*, *Ae. neglecta*, *Ae. geniculata*, *Ae. biuncialis*, and a tetraploid that shares only the U genome (*Ae. triuncialis*) to see their relative positions in the tree.

2.6 The *Aegilops* genome relation to the wheat genome

We mapped diploid *Aegilops* GBS reads to the wheat genome (CS.Ref.v1) (Appels et al., 2018) and computed sequence read mapping coverage. The reads mapped per Mb wheat subgenome and the total variants mapped for each wheat subgenome (A, B, D) were recorded. We did not further evaluate *Ae. tauschii* whose close genetic relationship as the wheat D subgenome donor has been clearly established. We also generated an unrooted NJ phylogenetic tree among diploid *Aegilops* and wheat using the variants called on wheat B and D reference subgenomes independently.

3 Results

3.1 *Aegilops* distributions

Aegilops species characterized in this study were primarily collected around the Fertile Crescent, Anatolia, central Asia, northern Africa, and southern Europe (Figure 1; Supplementary Material Table S1). Of the five sections, the *Aegilops* section [*Ae. umbellulata* (U), *Ae. kotschy* (US), *Ae. peregrina* (US), *Ae. triuncialis* (CU), *Ae. columnaris* (UM), *Ae. biuncialis* (UM), *Ae.*

neglecta (UM, UMN), *Ae. geniculata* (MU)] exhibited a much wider distribution from central Asia to northern Africa (Figure 1). The species of *Cylindropyrum* [*Ae. markgraffii* (C), *Ae. caudata* (C), and *Ae. cylindrica* (CD)] were primarily collected from Uzbekistan, Tajikistan, Kazakhstan, Azerbaijan, and Turkey. The species of the *Comopyrum* [*Ae. comosa* (M), *Ae. uniaristata* (N)] mainly come from Greece, Turkey, and Russia. The *Sitopsis* (S genome) species [*Ae. bicornis*, *Ae. searsii*, *Ae. sharonesis*, *Ae. longissima*, and *Ae. speltoides*] were predominantly collected in Turkey, Israel, Syria, Iraq, and Jordan. The *Vertebrata* section species [*Ae. tauschii* (D), *Ae. crassa* (DM, DDM), *Ae. ventricosa* (DN), *Ae. juvenalis* (DMU), and *Ae. vavilovii* (DMS)] were obtained from central Asia to southern Europe (Figure 1; Supplementary Material Table S1). The *Ae. mutica* tested here originated from Turkey and Armenia (Supplementary Material Table S1).

3.2 Marker discovery

We identified 54,667 *de novo* called SNPs for the entire *Aegilops* collections genotyped together. After filtering (MAF >0.01, missing <30%, and heterozygosity <10%), we retained 46,879 SNPs (Table 1). We removed 10 accessions (TA2674, TA2633, TA1733, TA11097, TA1740, TA2178, TA2042, TA1739, TA2316, and TA2296) with high rate of missing call (>80%). When we separated the genotyping information per species, we identified filtered segregating SNPs in the range of 1,483 for *Ae. searsii* to 14,322 for *Ae. speltoides* (Table 1). We also generated other SNP-genotyping matrices for analysis-specific purposes, such as for particular species' genetic relations and for genetically identical accession determination (Supplementary Material Table S2).

3.3 Gene bank curation

3.3.1 Misclassified accessions

The phylogenetic clustering and PCA enabled us to identify and correct the classification of 49 accessions (Figure 2; Supplementary Material Table S3). Most of the misclassified accessions were observed within tetraploid *Aegilops*. Twelve accessions that were previously considered as *Ae. triuncialis* were now identified as different *Aegilops*, whereas nine accessions that were classified as different *Aegilops* species are now re-identified as *Ae. triuncialis* (Supplementary Material Table S3). Similarly, 11 accessions identified as *Ae. neglecta* were now genetically identified as different *Aegilops*. The other misclassified example includes four accessions of each of *Ae. geniculata* and *Ae. vavilovii* (Supplementary Material Table S3). A few misclassified accessions of diploid *Aegilops* included *Ae. umbellulata* (2), *Ae. markgraffii* (2), and *Ae. searsii* (1) (Figure 2). The classes of all misclassified accessions were updated prior to the downstream population genomic analysis.

3.3.2 Genetically identical accessions

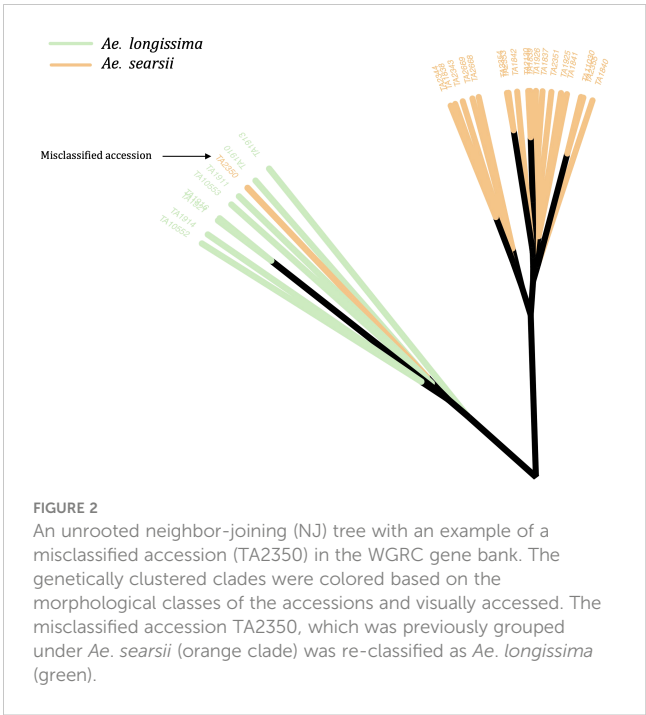
The gene bank curation discovered total 28 genetically identical accessions in *Ae. tauschii* and four members of the *Sitopsis* section (Supplementary Material Table S3). There were no pairs of *Ae.*

TABLE 1 *Aegilops* species with number of accessions, number of segregating loci, and the Nei's diversity indices.

Species	# Acces-sions	Segregating loci	Nei's index
All collection	1041	54667	0.104
<i>Ae. tauschii</i>	47	3369	0.024
<i>Ae. vavilovii</i>	6	9955	0.093
<i>Ae. mutica</i>	54*	8094	0.053
<i>Ae. ventricosa</i>	17	5828	0.05
<i>Ae. uniariistata</i>	24	5416	0.019
<i>Ae. umbellulata</i>	58	3391	0.015
<i>Ae. triuncialis</i>	199	8601	0.032
<i>Ae. speltoides</i>	97	14322	0.072
<i>Ae. sharonensis</i>	9	2224	0.019
<i>Ae. searsii</i>	18	1483	0.013
<i>Ae. peregrina</i>	33	7981	0.053
<i>Ae. neglecta</i>	71	11931	0.062
<i>Ae. markgrafii</i>	16	3474	0.022
<i>Ae. longissima</i>	14	3043	0.023
<i>Ae. kotschyi</i>	24	6876	0.053
<i>Ae. juvenalis</i>	9	8796	0.081
<i>Ae. geniculata</i>	143	8248	0.038
<i>Ae. cylindrica</i>	79	6173	0.046
<i>Ae. crassa</i>	32	8999	0.074
<i>Ae. comosa</i>	17	3388	0.025
<i>Ae. columnaris</i>	12	5382	0.041
<i>Ae. biuncialis</i>	52	7819	0.042
<i>Ae. bicornis</i>	13	1493	0.012

(*) The *Ae. mutica* being cross-pollinated we used many different samples from a single accession (s), so total of 54 plants rather than accessions.

speltoides accessions that have allele matching above 95%. Of 28 duplicated accessions, 17 were from *Ae. tauschii*, even though we only had a total of 47 *Ae. tauschii* accession for this experiment (Supplementary Material Table S3). In our previous study, we also reported many genetically identical accessions in *Ae. tauschii* collection (Singh et al., 2019a). The gene bank curator's observations also confirmed the phenotypic similarities among these genetically proven duplicate *Aegilops* accessions. As we examined the sources of these duplicate accessions, we found that most of them come from various institutes rather than from direct collectors. For instance, the *Ae. bicornis* genetically identical accessions TA1952, TA1956, and TA11023 were obtained from Kyoto University, the University of Manitoba, and the University of Missouri, respectively (Supplementary Material Table S1).



3.4 Phylogenetic clustering, PCA, and population structure

The unrooted NJ phylogenetic tree with all tested *Aegilops* accessions gave clear separation of species as the branches of clades and sub-clades differentiated all 23 species and the relevant groups (Figure 3). We observed the species sharing genomes as closely related clades, such as *Ae. kotschyi* and *Ae. peregrina* (SU) and *Ae. geniculata* and *Ae. biuncialis* (UM), clustered into respective primary clades. Overall, there were three primary clades: (i) the first clade consisted of *Ae. speltoides* and *Ae. mutica*; (ii) the second clade has four diploids of *Sitopsis* (except *Ae. speltoides*), *Ae. tauschii*, and D genome polyploids (except *Ae. cylindrica*); (iii) the third primary clade has all other species, including M, N, C, and U genome diploids and polyploids.

The hexaploid (6X) and tetraploid (4X) species within a clade, such as *Ae. neglecta* and *Ae. crassa*, were grouped separately by ploidy. The ploidy levels of these genetically clustered sub-groups (6X and 4X) were also verified using chromosome counting (Supplementary Material Figure S1) following Koo et al. (2017). The chromosome numbers of some accessions of *Ae. crassa* (Supplementary Material Figure S2) were also confirmed with the published data (Badaeva et al., 1998).

PCA also grouped the *Aegilops* species commensurate with the phylogenetic analysis. The first and second principal components (PC1 and PC2) explained about 17% and 14% of the variations among the *Aegilops*, respectively. PC1 separated *Ae. speltoides* from other polyploids and diploids (Figure 4), while the PC2 primarily differentiated *Ae. tauschii* and *Ae. speltoides*, the D genome donor to wheat and the potential sister group of the wheat B genome donor, respectively. As in phylogenetic analysis, PCA grouping also divided the 4X and 6X accessions of the *Ae. neglecta* and *Ae. crassa* (Figure 4).

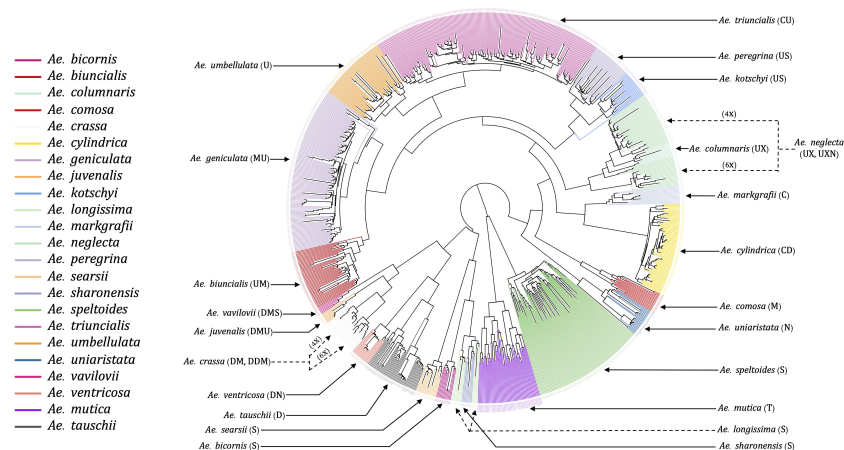


FIGURE 3

An unrooted neighbor-joining (NJ) tree of 23 different *Aegilops* species. The tree branches were colored based on the accessions genetic grouping after adjusting the misclassified accessions classes. The genome signs of each of the species were annotated along with their names as indicated by solid and dotted arrowheads.

3.5 Population genomics of *Sitopsis* and *Ae. mutica*

As we observed the separation of four *Sitopsis* members with *Ae. speltoides* and *Ae. mutica*, we separately examined the population of these species using reference-based variants from the *Ae. speltoides* genome assembly. The constructed phylogenetic tree distinctly divided the S-genome diploids into two large clades, one representing *Ae. speltoides* and the other with the remaining four *Sitopsis* (Figure 5). The genetic clustering corresponded to the historical sub-section division of the section is *Truncata* (*Ae. speltoides*) and the *Emarginata*. We also observed that the *Ae. mutica* (T genome) clustered closer to *Ae. speltoides* both in PCA and phylogenetic analysis (Figure 5). The relationships among *Sitopsis* group and *Ae. mutica* were further verified by computing pairwise Nei's F_{ST} (Nei, 1987), where we observed *Ae. mutica* has

the closest genetic relationship [lowest F_{ST} (0.65)] with *Ae. speltoides*, closer than any other members of the *Sitopsis* (Supplementary Material Table S4). Hence, all these analyses support that *Ae. mutica* as the sister taxon to *Ae. speltoides* and it is an *Aegilops* species.

Furthermore, within the S-genome diploids, the *Ae. speltoides* and *Ae. searsii* had the most genetic differentiation with the highest F_{ST} value 0.88 (Supplementary Material Table S4). However, the pairwise F_{ST} indicated that *speltoides* is genetically almost equally and highly differentiated from all other S-genome diploids (*Emarginata*) (Supplementary Material Table S4).

Population structure analysis of S-genome diploids matched with the phylogenetic tree and pairwise F_{ST} analysis. At $K = 2$, there was a differentiation between *Ae. speltoides* and the rest of the *Sitopsis*, while at $K = 3$, *Ae. searsii* also differentiated from the rest of the *Sitopsis* (Figure 6). At $K = 7$, *Ae. bicornis* accessions separated

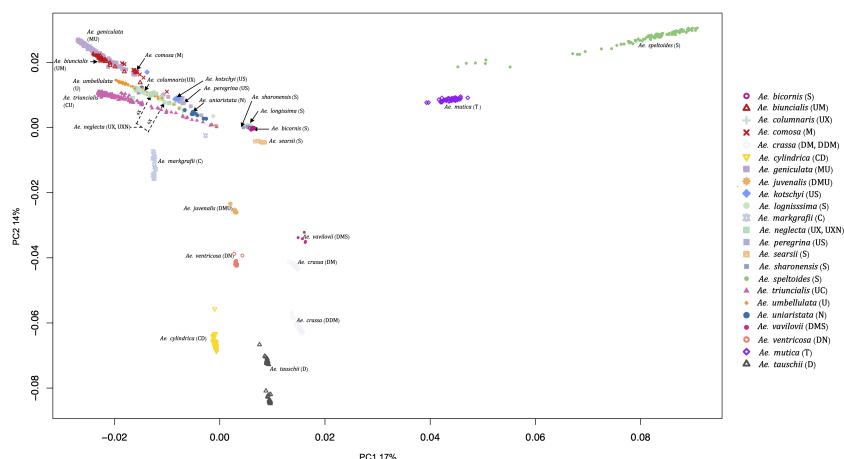


FIGURE 4

Principal component analysis (PCA) plot for all 23 *Aegilops* species with the first PCs. The 23 *Aegilops* species were grouped and colored based on their species and genome compositions.

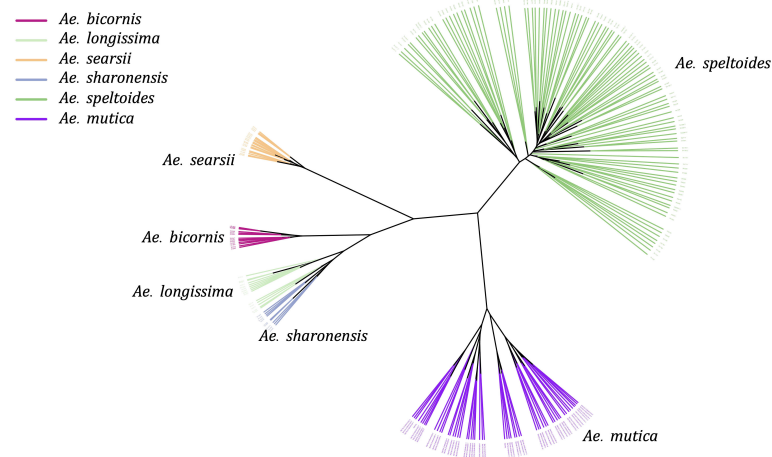


FIGURE 5

An unrooted Neighbor-Joining tree of five *Aegilops* species including *Sitopsis* section members (S genome) and *Ae. mutica* (T genome).

from others and then no new differentiation was observed until $K = 12$. Both in the phylogenetic tree and in population structure analysis, the *Ae. longissima* and *Ae. sharonensis* appeared as highly genetically similar groups (Figures 5, 6). In fact, there was no population differentiation between these two species at any level of K . The pairwise F_{ST} values also confirmed that these two species have the lowest pairwise $F_{ST} = 0.006$ (Supplementary Material Table S4), and the population differentiation is very low. Furthermore, two sub-groups within *Ae. speltoides*, var. *speltoides*, and var. *ligustica* also did not differentiate at any levels of K in the population structure analysis (Figure 6) and the PCA (Supplementary Material Figure S3). However, within *Ae. speltoides*, a few admixtures were observed and were differentiated for their geographical origins (Figure 6).

3.6 *Ae. umbellulata* and U-genome tetraploids

Most of the tetraploid *Aegilops* have the U genome; therefore, understanding the genetic relationship among members of the U-genome clade gives insight into a large set of taxa in the genus. Phylogenetic clustering of these species only showed two larger clades, where one was represented by *Ae. triuncialis* (UC) and the other had all remaining tetraploids (Figure 7). The diploid *Ae. umbellulata* sits on the intermediate position between the larger clades. Although the variants were only called on U-genome (*Ae. umbellulata*) *de-novo* reference, the tetraploids distinctly grouped for their genomic compositions. The tetraploid species *Ae. pregerina* and *Ae. kotschy* (US genome), *Ae. neglecta* and *Ae. columnaris*

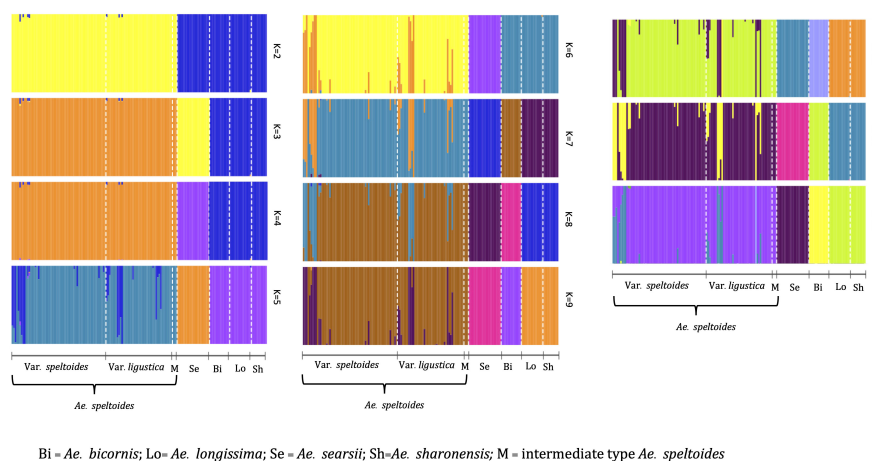


FIGURE 6

The population structure of S-genome diploids *Aegilops*, where the value of K and colors of the bars indicate the description of the groups. Each color represents a population and each bar with more than one color indicates the admixtures with the admixture proportions as represented by the proportion of each color.

(traditionally assigned as UM), and the UM genome tetraploids *Ae. biuncialis* and *Ae. geniculata* formed a separate clade and sub-clades (Figure 7). Also, we observed the splitting of *Ae. umbellulata* accessions into smaller clades. With a few exceptions as noted below, these phylogenies largely agree with previous genome designations.

3.7 Genome symbols of *Ae. columnaris* and *Ae. neglecta*

Ae. columnaris and *Ae. neglecta* formed a different clade than the other tetraploids with U and M genomes such as *Ae. geniculata* (UM) and *Ae. biuncialis* (MU) in both phylogenetic clustering and PCA (Figures 3, 4, 7; Supplementary Material Figure S4). The comparative positions of these tetraploids with other tetraploids in the genetic cluster indicated that these two tetraploids must be given unique genome symbols than the *Ae. geniculata* and *Ae. biuncialis* (Supplementary Material Figure S4). Thus, we hypothesized that *Ae. columnaris* and *Ae. neglecta* do not carry the M genome. The absence of M genome in *Ae. columnaris* and *Ae. neglecta* accessions was further confirmed by computing total reads mapped and total variants called on M-genome (*Ae. comosa* mock reference) and U genome (*Ae. umbellulata* mock reference) (Supplementary Material Figure S5, Supplementary Material Table S5). All four tetraploid species, namely, *Ae. columnaris* and *Ae. neglecta* along with *Ae. geniculata* and *Ae. biuncialis* exhibited an equal percentage of overall reads alignment (~38%) on the U genome, whereas the percentage read alignment of *Ae. columnaris* and *Ae. neglecta* on M genome was low (~21%) as compared to the alignment of *Ae. geniculata* and *Ae. biuncialis* reads (~38%). We also noticed that a few *Ae. comosa* segregating loci were mapped for *Ae. columnaris* (10%) and *Ae. neglecta* (24%) on the M genome. In contrast, *Ae. biuncialis* had 50% and *Ae. geniculata* had 46% M-genome loci. Hence, the proportion of mapped reads and loci also suggested that the *Ae. neglecta* and *Ae. columnaris* must have the U

genome, but a different second sub-genome than M. Thus, we proposed that *Ae. columnaris* and *Ae. neglecta* genome formulas are most likely UX (X, the unknown genome) or UXN in hexaploid form as proposed based on the cytology (Dvorak, 1998; Badaeva et al., 2018).

3.8 *Aegilops* species diversity

For the entire collection, we obtained 54,667 SNPs, which were skewed to low MAF as expected for a diverse population like this (Supplementary Material Figure S6). Despite the differences in population size, the total segregating loci for the species or groups were mostly dependent on the ploidy levels and the reproductive biology (inbred vs. outcrossing) (Table 1). The polyploids and outcrossing species had a higher number of segregating loci compared to other diploids (Table 1). Notably, the MAF of the loci in partially cross-pollinated species, such as *Ae. speltoides*, had a higher frequency (Supplementary Material Figure S7) than that of the MAF of the loci for the entire *Aegilops* collection (Supplementary Material Figure S6).

The Nei's diversity indices also followed the pattern of segregating loci which were greater in polyploid and cross-pollinated species. We computed Nei's diversity index for the entire collection as 0.10 (Table 1). Of all 23 species, *Ae. bicornis* had the lowest Nei's diversity index (0.012) followed by *Ae. searsii* (0.013) and *Ae. umbellulata* (0.015). Among the diploids, the *Ae. speltoides* had the highest Nei's diversity (0.072), which was followed by *Ae. mutica* (0.053). Among the tetraploids, the *Ae. triuncialis* had the lowest diversity index (0.032) while the *Ae. neglecta* had the highest diversity index (0.062). The hexaploid species *Ae. vavilovii* has the highest Nei's diversity index value among all 23 species analyzed in the experiment (Table 1). This increased diversity can be attributed to various factors such as multiple gene copies, hybridization during speciation, increased mutation rates, and more opportunities for recombination due to the presence of multiple genomes.

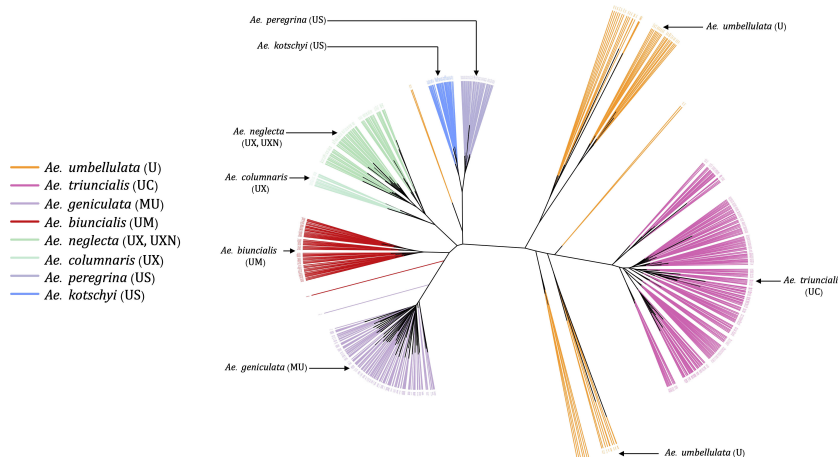


FIGURE 7

An unrooted neighbor-joining (NJ) tree for *Ae. umbellulata* and U genome containing tetraploids within the genus *Aegilops*.

3.9 Wheat and *Aegilops* genomes

The genetic clustering between wheat and all diploid *Aegilops* showed that *Ae. tauschii* is the nearest extant *Aegilops* to the bread wheat (Supplementary Material Figure S8). The genetic cluster clearly showed that *Ae. speltoides* is not closer to wheat as *Ae. tauschii* and even other diploids, and supporting that, *Ae. speltoides* is likely not the direct progenitor of the wheat subgenome B (Supplementary Figure S8). However, the *Ae. speltoides* read depth mapping and SNP detection occurred at its maximal on the wheat subgenome B (Figure 8), indicating the species as the sister group of wheat B genome progenitor. Furthermore, the other members of the *Sitopsis* group clustered between *Ae. speltoides* clade and the clade with *Ae. tauschii* and the wheat subclades in the phylogenetic tree (Supplementary Material Figure S8). Consistent with the genetic clustering, their maximum read mapping and SNP detection also occurred at subgenome D and B chromosomes (Supplementary Material Figures S8–S10), suggesting that the four members of *Sitopsis*, except *Ae. speltoides*, have very strong genomic relationships with both D and B subgenomes.

Similarly, in the U genome diploid (*Ae. umbellulata*), the highest proportion of sequence reads was mapped onto wheat

chromosomes of the D subgenome, followed by those of the A and B subgenomes (Supplementary Material Figure S11). Exceptionally, a slightly higher proportion of reads were mapped on 2A than the 2D. The pattern of SNP detection was exactly the same as read mapping, indicating that wheat subgenome D is the closest to the U genome of the *Aegilops*. However, relations between the wheat A genome and the *Aegilops* U genome cannot be overlooked, as reasonably higher reads and loci were mapped on the A genome as compared to the wheat B genome (Supplementary Material Figure S11). Likewise, the highest number of reads and SNPs were mapped onto wheat subgenome D for the N genome diploid (*Ae. uniaristata*) (Supplementary Material Figure S12), for the M genome diploid (*Ae. comosa*) (Supplementary Material Figure S13), and C genome diploid (*Ae. markgraffii*) (Supplementary Material Figure S14). These observations suggest that the N, M, and C genomes of *Aegilops* are also genetically closer to the D subgenome than A and B.

Interestingly, the *Ae. mutica* accessions when mapped onto the wheat subgenomes showed higher sequence read and loci mapped on the wheat D subgenome (Supplementary Material Figure S15). The read and loci mapping pattern was unchanged even when we

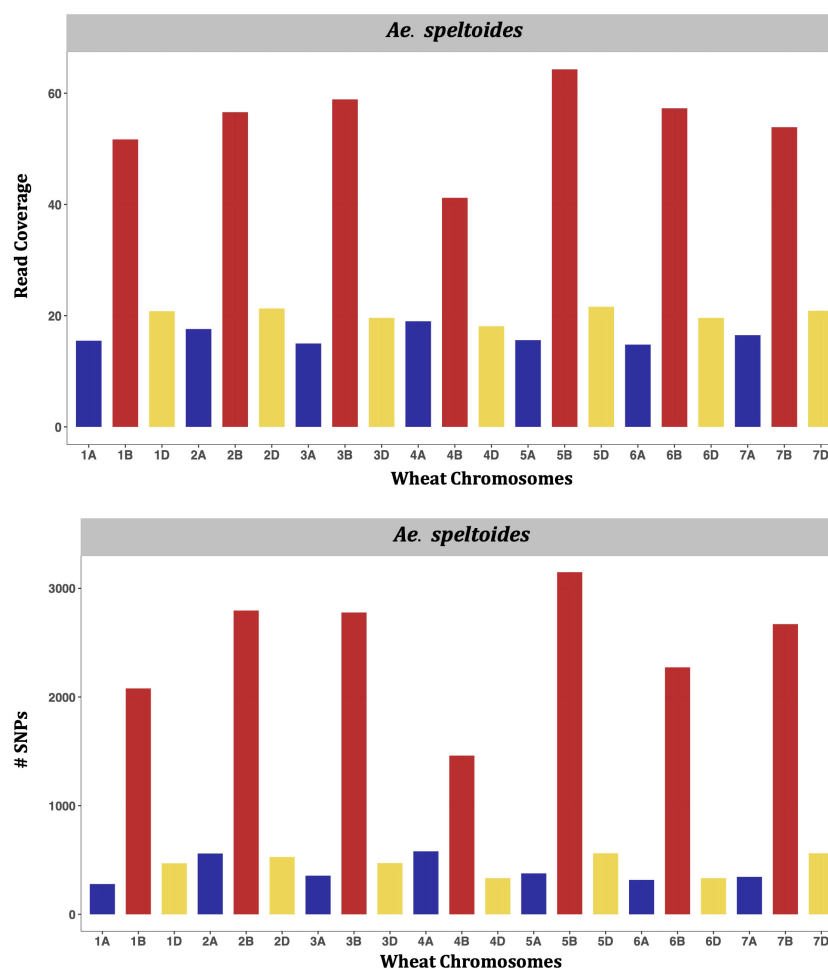


FIGURE 8

Bar charts showing genomic relations between *Ae. speltoides* and wheat. The average number of *Ae. speltoides* sequence reads mapped per Mb of the wheat genome (upper panel), and numbers of *Ae. speltoides* variants mapped on the respective wheat chromosomes (lower panel).

replaced wheat D subgenome chromosomes with *Ae. tauschii* chromosomes. Nevertheless, all types of population grouping within *Aegilops* (Figures 3–5; Supplementary Material Figure S8) evidently showed that *Ae. mutica* is a sister group of *Ae. speltoides* and still a member of B lineage. Some recent studies based on whole genome sequencing data have also reported a higher sequence read and loci mapping of *Ae. mutica* on the wheat D subgenome compared to others (Grewal et al., 2022; Li et al., 2022).

4 Discussions

4.1 Multi-species diverse *Aegilops* collection and gene bank curation

In this study, we genotyped over a thousand accessions representing almost all species of the *Aegilops* genus, covering the full range of their natural distributions under the Van Slageren (1994) nomenclature, with missing only *Ae. caudata*. We curated the WGRG gene bank *Aegilops* collection, giving curated germplasm sets that are ready to screen for the novel alleles and utilize in the breeding program. The misclassified accession were confirmed with multiple analyses including phylogenetic clustering of the whole population, species or genome-specific populations and PCA, therefore there is strong support for the genotype-based identification of these misclassified accessions (Supplementary Material Table S3). Since the genotype-based clustering evidently differentiated the hexaploid and tetraploid accessions within the species such as *Ae. crassa* and *Ae. neglecta*, we can also provide the ploidy levels information as a means of within-species classification and update the gene bank database.

Here, we identified the redundant accessions in the species with variants called directly on reference genome assemblies. This gives increased power and accuracy in variant calling. Therefore, we suggest the re-assessment of genetically redundant accessions for other *Aegilops* species in the future when reference assemblies are available. For the polyploid *Aegilops*, reference variant calling can be done whenever the component species reference genomes are available using a combined reference genome or independent variant calling to each genome. As we examined the origins of these genetically verified and visually confirmed duplicates, we discovered that many of them originated from various research institutes rather than directly from collectors. Therefore, we here recommend the need for curating the global collection of these naturally collected germplasms, as the same genetic materials can be preserved under different plant IDs or accession numbers. In our previous studies, we also observed several duplicates originating from the exact same collection sites (Singh et al., 2019a; Adhikari et al., 2022a). This is because these self-pollinated species have already reached genomic saturation, and the progeny of the same mother parents are genetically identical inbred. Although we do not suggest discarding the duplicated accessions identified here, we strongly suggest for considering these results when utilizing the collection, such as screening the accessions for disease resistance or developing introgression populations. Overall, gene bank curation

helps in the management, preservation, and utilization of the germplasms (Singh et al., 2019a; Volk et al., 2021).

4.2 *Aegilops* population analysis

This is the most comprehensive *Aegilops* population genetic study reported so far with over 45 thousand *de-novo* filtered SNPs and reference-based variants. In the study, we took advantage of recently completed chromosome-scale genome assemblies of diploid *Aegilops* (Wang et al., 2021; Avni et al., 2022; Li et al., 2022; Yu et al., 2022). Until now, the lack of genomic resources including reference assemblies has been a major issue hindering the species population genomic analysis. Therefore, future genomic studies on *Aegilops* must focus on generating more genomic resources for other diploids and polyploids. With a larger population and thousands of genomic variants, the population grouping that we observed here was at the finest level, enabling us to differentiate the 4X and 6X accessions within a species (Supplementary Material Figure S1).

4.3 *Ae. speltoides*, other *Sitopsis* and *Ae. mutica*

Our genetic analysis supports that the *Ae. mutica* requires no genus-level separation from other *Aegilops* as Van Slageren (1994) suggested. It is genetically an *Aegilops* taxon closer to *Ae. speltoides* (Figures 4, 5). This is in agreement with recent reports (Bernhardt et al., 2020; Li et al., 2022). Further genomic analysis may require high coverage genomic data and a greater number of samples to better understand the relationship among *Ae. mutica* and other diploid *Aegilops*. Additionally, the genetic differences that we observed here between the *Truncata* (*Ae. speltoides*) and *Emarginata* (four other) *Sitopsis* were greater; therefore, the redefinition of the section *Sitopsis* could be desirable. One of the ideas could be the separation of *Ae. speltoides* from the rest of the four *Sitopsis* members and regrouping the *Ae. speltoides* with *Ae. mutica* (Figures 3–5; Supplementary Material Figure S8).

We also showed that the *Ae. sharonensis* and *Ae. longissima* have very high genetic similarities or a low genetic differentiation ($F_{ST} = 0.006$) and are most likely the sub-species of the same species. Also, both of these species are equally distant from *Ae. speltoides*. The finding is also supported by the latest study, where Avni et al. (2022) reported that the genomes of these two species are highly similar with identical genome sizes and also share 292 orthogroups.

In this study, we observed a little genetic difference between the two sub-taxa of *Ae. speltoides*; var. *speltoides* and *ligustica* with no population differentiation (Figure 6; Supplementary Material Figure S3), in accordance with several past studies. These two sub-groups of *speltoides* not only have distinct spike morphology and mode of seed dispersal but also exhibit similar karyotype structure, producing fully fertile hybrid and mixed stands of two types naturally exhibits (Zohary and Imber, 1963). A single locus *Lig* on chromosome 3S governs the spike morphology of these two sub-

groups (Luo et al., 2005); otherwise, they are highly genetically similar.

4.4 U-genome species, some tetraploid genome symbols and polyploid *Aegilops*

The U genome tetraploids and its progenitor *Ae. umbellulata* genetic clustering revealed the unique relationships among the species. We observed the *Ae. umbellulata* accessions split into sub-groups in such a way that some accessions were clustered closer to *Ae. triuncialis* clade whereas some other accessions reposed near the other tetraploid clades (Figure 7), suggesting the potential unique *Ae. umbellulata* ancestries for the two groups.

In this study, we found further evidence that the *Ae. columnaris* and *Ae. neglecta* genome symbols should not include the M genome designation (Supplementary Material Figures S4, S5 and Supplementary Table S5), based on sequence read and loci mapping data, and phylogenetic clustering (Supplementary Material Figure S4). Cytology-based approaches (Resta et al., 1996; Dvorak, 1998; Badaeva et al., 2004; Badaeva et al., 2018) have previously discussed this issue and suggested the symbol “X” (Resta et al., 1996). Several lines of evidence, including low chromosome pairing in hybrids of *Ae. columnaris* x *Ae. comosa* (the M genome progenitor), variation in repetitive nucleotide sequences, and differences in the karyotype structure C-banding pattern, have been used to confirm the absence of the M genome in *Ae. neglecta* and *Ae. columnaris* (Badaeva et al., 2018). This study has provided further verification with thousands of loci. Therefore, we suggest research communities for the consistent use of genome symbols for *Ae. columnaris* (UX) and *Ae. neglecta* (UX or UXN). Furthermore, cytological and genomic evaluation of the X genome is certainly warranted.

4.5 *Aegilops* genetic diversity

Ploidy level and the mode of fertilization appeared as major determinants of *Aegilops* accessions diversity (Table 1). Interestingly, we did not observe the direct impact of population size on Nei's diversity index (Nei, 1987) at any ploidy levels (Table 1). For example, the diploid *Ae. sharonensis* (nine accessions) exhibited a higher diversity index (0.019) compared to *Ae. umbellulata* (58 accessions), and the tetraploid *Ae. ventricosa* (17 accessions) had a higher diversity index than another tetraploid, *Ae. triuncialis* (199 accessions) (Table 1). Additionally, we noted that *Ae. speltoides*, as the diploid species, displayed the greatest diversity, and relatively higher diversity indices were observed in the S genome polyploids such as *Ae. kotschyi*, *Ae. peregrina*, and *Ae. vavilovii* (Table 1). In summary, most of the *Aegilops* species exhibited a wider and more variable diversity and had greater potential to be utilized in wheat breeding. Therefore, it is crucial to make serious efforts toward the *in-situ* conservation of these germplasms and enhance *ex-situ* *Aegilops* germplasm collections. Kilian et al. (2011) also emphasized the urgency of protecting these *Aegilops* germplasms, highlighting the importance of understanding

Aegilops genetic diversity, *Aegilops-Triticum* molecular biological relationships, and identifying and preserving suitable *Aegilops* alleles for wheat breeding.

4.6 *Aegilops* and wheat genomes

This study represents, perhaps, the first comprehensive report on genomic relationships between all *Aegilops* genomes and wheat sub-genomes, based on high-throughput sequence-based markers and robust phylogeny of these wild wheat species. Consistent with some earlier reports, our findings indicate that most of the *Aegilops* genomes (U, M, N, C) are genetically closer to the wheat D subgenome (Supplementary Material Figures S9-S15), with the exception of *Ae. speltoides* (Figure 8). Several studies have reported that the speciation event of the B genome donor occurred earlier than the speciation of *Ae. tauschii* (the D-genome lineage), resulting in stronger evolutionary relationships of the U, M, N, and C diploid *Aegilops* within the D-genome lineage (Glémin et al., 2019; Tanaka et al., 2020; Said et al., 2021).

In our study, we observed unique relationships between certain genomes within the *Aegilops-Triticum* complex that had not been clearly described in earlier studies. One of the most important observations is that four *Sitopsis* species exhibit relationships with both the B and D subgenomes of wheat. These relationships were evident in the phylogenetic tree and supported by statistic on sequence read and mapped loci (Supplementary Material Figures S8-S10). Interestingly, recent reports have also considered these four *Sitopsis* members as part of the D lineage, and are closer to the wheat D subgenome (Li, 2011; Avni et al., 2022; Li et al., 2022).

4.7 *Ae. mutica*, wheat genomes, and homoploid hybridization

In this study, we observed unique genetic characteristics of *Ae. mutica* as it was phylogenetically closer to the *Ae. speltoides* (Figures 3–5 and Supplementary Material Figure S8); however, it showed genetic similarities with the wheat D subgenome (Supplementary Material Figure S15). Interestingly, similar observations have been reported in recent studies. Li et al. (2022) reported lower genetic similarities between *Ae. mutica* and wheat B subgenome computed as genetic relatedness. Likewise, Grewal et al. (2022) reported a similar relationship between *Ae. mutica* and wheat subgenomes, with the highest number of *Ae. mutica* loci mapped on the D subgenome, rather than the A and B subgenomes (Supplementary Material Figure S15). Therefore, the genetic similarities and phylogenetic relationship between the *Ae. mutica* and the *Aegilops-Triticum* complex are exclusive and warrant further investigation in a larger population with high-depth sequencing. Furthermore, these analyses indicate that *Ae. mutica* genome may have undergone independent evolution or played a role in the evolution of polyploid genomes following its divergence from *Ae. speltoides*. Some recent studies also argued that *Ae. mutica* and the D lineage underwent homoploid hybridization followed by introgression (Bernhardt et al., 2020; Li et al., 2022). Bernhardt et al.

(2020) reported that most of the members of the *Aegilops* genus, except *Ae. speltoides*, likely evolved through ancient primordial hybrid speciation events involving the ancestral *Triticum* and *Ae. mutica*. Earlier studies also indicated a higher degree of homology between *Ae. mutica* and the wheat D subgenome (Jones and Majisu, 1968).

4.8 Utilizing *Aegilops* novel alleles in high-throughput genotyping era

This study establishes a solid foundation for the future utilization of *Aegilops* germplasm within the WGRC gene bank. The development of introgression populations, combined with new genomic tools, has the potential to accelerate the selection and advancement of novel alleles in wheat breeding. In an ongoing investigation, we have successfully created wheat—*Ae. speltoides* introgression lines and have achieved the mapping of introgression segments using a skim-sequencing approach (Adhikari et al., 2022b). Likewise, association genomics approaches can be leveraged to identify novel *Aegilops* alleles directly within the wild germplasm collections (Gaurav et al., 2022). As an example, candidate genes associated with various agronomic traits in another wild wheat relative, einkorn, were identified using the cost-effective skim-sequencing technique (Saripalli et al., 2023). Within this context, the importance of these highly diverse *Aegilops* accessions is further enhanced. Finding trait-related alleles through genome-wide association studies, generating reference assemblies, and resequencing diverse panels represent some of the future steps in harnessing the potential of these valuable *Aegilops* genetic resources for enhancing wheat.

In conclusion, this study has unveiled the genomic and genetic relationships among all *Aegilops* species and demonstrated the efficient use of the GBS approach for curating gene bank accessions and investigating the genetic diversity and population structure of the entire *Aegilops* collection. Most likely this is the first genomic analysis of a nearly complete set of the genus *Aegilops* encompassing 23 species. We dissected a larger population (1,041) using over 45K SNPs and constructed a robust phylogenetic tree and the PCA clusters. The population grouping and structuring of this valuable wild wheat species largely align with the traditional nomenclatures at the species level. Moreover, using these high-throughput genome-wide markers, we have confirmed the genome symbols of two tetraploid species that were previously under debate in the literature.

Our findings also reveal that each *Aegilops* subgenome and wheat subgenomes exhibit unique relationships at the genomic level, warranting further investigation. Notably, *Ae. mutica* showed unique characteristics, appearing as a sister group of *Ae. speltoides*, yet displaying a higher number of sequences and variants mapped onto the wheat subgenome D. The genetic and evolutionary relationships among *Aegilops* and with wheat will become clearer when we have more genomic resources, such as genome assemblies and resequencing data for each *Aegilops* species. This study offers a comprehensive view of the relative genetic

diversities of all 23 species together for the first time. The substantial genetic diversity observed, along with its relative extent in each *Aegilops* species, presents an opportunity to select species and germplasms as sources of novel alleles for wheat breeding and improvement.

Data availability statement

The Raw GBS data, the fastq files, are available in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under the BioProject accession PRJNA985892. The key file and necessary SNP matrices and the R script files (.rmd) are provided in the dryad public repository which are available with the unique DOI: 10.5061/dryad.mgqnk994n. All data are available in the article or the supplementary files and at the Dryad digital repositories <https://datadryad.org/stash/dataset/doi:10.5061/dryad.mgqnk994n>.

Author contributions

LA: Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. JR: Data curation, Formal Analysis, Methodology, Resources, Validation, Writing – review & editing. SW: Investigation, Methodology, Writing – review & editing. D-HK: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – review & editing. BF: Conceptualization, Methodology, Resources, Supervision, Validation, Writing – review & editing. JP: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This material is based upon work supported by the US National Science Foundation and Industry Partners under Award No. (1822162) “Phase II Industry/University research consortium (IUCRC) at Kansas State University (KSU) Center for Wheat Genetic Resources” and from King Abdullah University of Science and Technology. This research was also partly supported by the U.S. Department of Agriculture, National Institute of Food and Agriculture (Grant No. 2020-67103-31455).

Acknowledgments

We would like to acknowledge Kansas high-performance computing cluster “beocat” for data storage and the Linux environment for data analysis. We are thankful to everyone who contributed to WGRC gene bank *Aegilops* collection.

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.

Author disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or industry partners.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2023.1268370/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

The GBS SNP-based unrooted neighbor-joining (NJ) tree separating tetraploid and hexaploid accessions of *Ae. neglecta* (blue clade) and the chromosome counts of two representative individuals from each 4X and 6X sub-clade of the *Ae. neglecta*.

SUPPLEMENTARY FIGURE 2

An unrooted neighbor-joining (NJ) tree of *Ae. juvenalis*, *Ae. crassa*, and *Ae. vavilovii*. The tree branches were colored based on the accession's taxon.

SUPPLEMENTARY FIGURE 3

Principal component analysis (PCA) plot showing two forms of *Ae. speltoides*: var. *speltoides* and *ligustica*.

SUPPLEMENTARY FIGURE 4

An unrooted neighbor-joining (NJ) tree separating some tetraploid *Aegilops* accessions containing two species whose genome formula is controversial, the *Ae. neglecta* and *Ae. columnaris*.

SUPPLEMENTARY FIGURE 5

The bar chart showing the overall sequence read alignment of four tetraploid *Aegilops* species: *Ae. biuncialis*, *Ae. geniculata*, *Ae. columnaris*, and *Ae. neglecta* when aligned on M and U genome *de-novo* mock reference.

SUPPLEMENTARY FIGURE 6

Minor allele frequency (MAF) distribution within the loci for the entire *Aegilops* collection.

SUPPLEMENTARY FIGURE 7

Distribution of minor alleles frequency (MAF) for segregating variants in *Ae. speltoides*.

SUPPLEMENTARY FIGURE 8

An unrooted neighbor-joining (NJ) tree constructed using the genotyping information generated by using wheat B genome as a reference (left); and the unrooted NJ tree constructed using genotyping profile generated using the wheat D genome as a reference (right).

SUPPLEMENTARY FIGURE 9

Bar charts showing genomic relations between the *Sitopsis* section *Aegilops* (except *Ae. speltoides*) and the wheat.

SUPPLEMENTARY FIGURE 10

Bar charts showing genomic relations between the *Sitopsis* section *Aegilops* (except *Ae. speltoides*) and the wheat.

SUPPLEMENTARY FIGURE 11

Bar chart showing genomic relation between U genome diploid *Ae. umbellulata* and wheat.

SUPPLEMENTARY FIGURE 12

Bar chart showing genomic relation between N genome diploid *Ae. uniaristata* and wheat.

SUPPLEMENTARY FIGURE 13

Bar chart showing genomic relation between M genome diploid *Ae. comosa* and wheat.

SUPPLEMENTARY FIGURE 14

Bar chart showing genomic relation between C genome diploid *Ae. markgrafii* and wheat.

SUPPLEMENTARY FIGURE 15

Bar charts showing genomic relations between *Ae. mutica* and wheat.

SUPPLEMENTARY TABLE 1

List of *Aegilops* germplasms in the WGRC gene bank collection with the taxa and origins of the accessions (separate excel file).

SUPPLEMENTARY TABLE 2

Different SNP matrices, population genotyped, the reference sequence used and the application which used the SNP matrix.

SUPPLEMENTARY TABLE 3

Misclassified and genetically identical (redundant) *Aegilops* accessions (separate excel file).

SUPPLEMENTARY TABLE 4

Sitopsis section *Aegilops* and *Ae. mutica* pairwise F_{ST} values.

SUPPLEMENTARY TABLE 5

Total segregating loci in UM and UX genome species when called variants on the M genome and U genome mock references independently.

References

- Adhikari, L., Lindstrom, O. M., Markham, J., and Missaoui, A. M. (2018). Dissecting key adaptation traits in the polyploid perennial medicago sativa using GBS-SNP mapping. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00934
- Adhikari, L., Raupp, J., Wu, S., Wilson, D., Evers, B., Koo, D. H., et al. (2022a). Genetic characterization and curation of diploid A-genome wheat species. *Plant Physiol.* 188, 2101–2114. doi: 10.1093/plphys/kiac006
- Adhikari, L., Shrestha, S., Wu, S., Crain, J., Gao, L., Evers, B., et al. (2022b). A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. *Sci. Rep.* 12, 17583. doi: 10.1038/s41598-022-19858-2
- Ahmed, H. I., Heuberger, M., Schoen, A., Koo, D.-H., Quiroz-Chavez, J., Adhikari, L., et al. (2023). Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature* 620, 830–838. doi: 10.1038/s41586-023-06389-7

- Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361. doi: 10.1126/science.aar7191
- Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nat. Climate Change* 5, 143–147. doi: 10.1038/nclimate2470
- Avni, R., Lux, T., Minz-Dub, A., Millet, E., Sela, H., Distelfeld, A., et al. (2022). Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. *Plant J.* 110, 179–192. doi: 10.1111/tpj.15664
- Badaeva, E., Amosova, A., Samatadze, T., Zoshchuk, S., Shostak, N., Chikida, N., et al. (2004). Genome differentiation in Aegilops. 4. Evolution of the U-genome cluster. *Plant Systematics Evol.* 246, 45–76. doi: 10.1007/s00606-003-0072-4
- Badaeva, E. D., Friebe, B., Zoshchuk, S. A., Zelenin, A. V., and Gill, B. S. (1998). Molecular cytogenetic analysis of tetraploid and hexaploid aegilops crassa. *Chromosome Res.* 6, 629–637. doi: 10.1023/A:1009257527391
- Badaeva, E. D., Ruban, A. S., Shishkina, A. A., Sibikeev, S. N., Druzhin, A. E., Surzhikov, S. A., et al. (2018). Genetic classification of Aegilops columnaris Zhuk. (2n=4x=28, UuUcXcXc) chromosomes based on FISH analysis and substitution patterns in common wheat × Ae. columnaris introgressive lines. *Genome* 61, 131–143. doi: 10.1139/gen-2017-0186gM29216443
- Bernhardt, N., Brassac, J., Dong, X., Willing, E.-M., Poskar, C. H., Kilian, B., et al. (2020). Genome-wide sequence information reveals recurrent hybridization among diploid wheat wild relatives. *Plant J.* 102, 493–506. doi: 10.1111/tpj.14641
- Chen, W.-C. (2011). *Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm* (IA, United States: Iowa State University).
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560
- Cruz, C. D., Peterson, G. L., Bockus, W. W., Kankana, P., Dubcovsky, J., Jordan, K. W., et al. (2016). The 2NS Translocation from Aegilops ventricosa Confers Resistance to the Triticum Pathotype of Magnaporthe oryzae. *Crop Sci.* 56, 990–1000. doi: 10.2135/cropsci2015.07.0410
- Dvorak, J. (1998). “Genome analysis in the Triticum-Aegilops alliance,” in *Proceedings of the 9th international wheat genetics symposium*. A. E. Slinkard (Ed.). (Saskatoon, Saskatchewan, Canada: University Extension Press, University of Saskatchewan) 1, 8–11.
- Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024
- Friebe, B., Jiang, J., Raupp, W. J., McIntosh, R. A., and Gill, B. S. (1996). Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. *Euphytica* 91, 59–87. doi: 10.1007/BF00035277
- Gao, L., Koo, D.-H., Juliana, P., Rife, T., Singh, D., Lemes Da Silva, C., et al. (2021). The Aegilops ventricosa 2Nvs segment in bread wheat: cytology, genomics and breeding. *Theor. Appl. Genet.* 134, 529–542. doi: 10.1007/s00122-020-03712-y
- Gaurav, K., Arora, S., Silva, P., Sánchez-Martín, J., Horsnell, R., Gao, L., et al. (2022). Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat. Biotechnol.* 40, 422–431. doi: 10.1038/s41587-021-01058-4
- Glaubit, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9, e90346. doi: 10.1371/journal.pone.0090346
- Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., et al. (2019). Pervasive hybridizations in the history of wheat relatives. *Sci. Adv.* 5, eaav9188. doi: 10.1126/sciadv.aav9188
- Grewal, S., Coombes, B., Joynson, R., Hall, A., Fellers, J., Yang, C. Y., et al. (2022). Chromosome-specific KASP markers for detecting Amblyopyrum muticum segments in wheat introgression lines. *Plant Genome* 15, e20193. doi: 10.1002/tpg2.20193
- Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., et al. (2007). Grinding up wheat: A massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* 24, 1506–1517. doi: 10.1093/molbev/msm077
- Jones, J. K., and Majisu, B. N. (1968). THE HOMOEOLGY OF AEGILOPS MUTICA CHROMOSOMES. *Can. J. Genet. Cytology* 10, 620–626. doi: 10.1139/g68-080
- Kilian, B., Mammen, K., Millet, E., Sharma, R., Graner, A., Salamini, F., et al. (2011). “Aegilops,” in *Wild Crop Relatives: Genomic and Breeding Resources: Cereals*. Ed. C. Kole (Berlin, Heidelberg: Springer Berlin Heidelberg).
- Kishii, M. (2019). An update of recent use of aegilops species in wheat breeding. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00585
- Koo, D.-H., Liu, W., Friebe, B., and Gill, B. S. (2017). Homoeologous recombination in the presence of Ph1 gene in wheat. *Chromosoma* 126, 531–540. doi: 10.1007/s00412-016-0622-5
- Leigh, F. J., Wright, T. I. C., Horsnell, R. A., Dyer, S., and Bentley, A. R. (2022). Progenitor species hold untapped diversity for potential climate-responsive traits for use in wheat breeding and crop improvement. *Heredity* 128, 291–303. doi: 10.1038/s41437-022-00527-z
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509
- Li, H., Dong, Z., Ma, C., Tian, X., Xiang, Z., Xia, Q., et al. (2019). Discovery of powdery mildew resistance gene candidates from Aegilops biuncialis chromosome 2Mb based on transcriptome sequencing. *PLoS One* 14, e0220089. doi: 10.1371/journal.pone.0220089
- Li, L. F., Zhang, Z. B., Wang, Z. H., Li, N., Sha, Y., Wang, X. F., et al. (2022). Genome sequences of five Sitopsis species of Aegilops and the origin of polyploid wheat B subgenome. *Mol. Plant* 15, 488–503. doi: 10.1016/j.molp.2021.12.019
- Lopes, M. S., El-Basyoni, I., Baenziger, P. S., Singh, S., Royo, C., Ozbek, K., et al. (2015). Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J. Exp. Bot.* 66, 3477–3486. doi: 10.1093/jxb/erv122
- Luo, M.-C., Deal, K. R., Yang, Z.-L., and Dvorak, J. (2005). Comparative genetic maps reveal extreme crossover localization in the Aegilops speltoides chromosomes. *Theor. Appl. Genet.* 111, 1098–1106. doi: 10.1007/s00122-005-0035-y
- Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., et al. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* 551, 498–502. doi: 10.1038/nature24486
- Marais, G. F., Mccallum, B., Snyman, J. E., Pretorius, Z. A., and Marais, A. S. (2005). Leaf rust and stripe rust resistance genes Lr54 and Yr37 transferred to wheat from Aegilops kotschy. *Plant Breed.* 124, 538–541. doi: 10.1111/j.1439-0523.2005.01116.x
- Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinf.* 17, 29. doi: 10.1186/s12859-016-0879-y
- Nei, M. (1987). *Molecular Evolutionary Genetics* (New York: Columbia University Press), 512.
- Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633
- Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7, e32253. doi: 10.1371/journal.pone.0032253
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.r-project.org/>.
- 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
- Rakszegi, M., Molnár, I., Darkó, É., Tiwari, V. K., and Shewry, P. (2020). Editorial: aegilops: promising genesources to improve agronomical and quality traits of wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01060
- Resta, P., Zhang, G.-B., Dubcovsky, J., and Dvořák, J. (1996). The origins of the genomes of Triticum biunciale, t. ovatum, t. neglectum, t. columnare, and t. rectum (poaceae) based on variation in repeated nucleotide sequences. *Am. J. Bot.* 83, 1556–1565. doi: 10.1002/j.1537-2197.1996.tb12813.x
- Said, M., Holušová, K., Farkas, A., Ivanizs, L., Gaál, E., Cápál, P., et al. (2021). Development of DNA Markers From Physically Mapped Loci in Aegilops comosa and Aegilops umbellulata Using Single-Gene FISH and Chromosome Sequences. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.689031
- Saripalli, G., Adhikari, L., Amos, C., Kibriya, A., Ahmed, H. I., Heuberger, M., et al. (2023). Integration of genetic and genomics resources in einkorn wheat enables precision mapping of important traits. *Commun. Biol.* 6, 835. doi: 10.1038/s42003-023-05189-z
- Singh, N., Wu, S., Raupp, W. J., Sehgal, S., Arora, S., Tiwari, V., et al. (2019a). Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep.* 9, 650. doi: 10.1038/s41598-018-37269-0
- Singh, N., Wu, S., Tiwari, V., Sehgal, S., Raupp, J., Wilson, D., et al. (2019b). Genomic analysis confirms population structure and identifies inter-lineage hybrids in aegilops tauschii. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00009
- Suneja, Y., Gupta, A. K., and Bains, N. S. (2019). Stress Adaptive Plasticity: Aegilops tauschii and Triticum dicoccoides as Potential Donors of Drought Associated Morpho-Physiological Traits in Wheat. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00211
- Tanaka, S., Yoshida, K., Sato, K., and Takumi, S. (2020). Diploid genome differentiation conferred by RNA sequencing-based survey of genome-wide polymorphisms throughout homoeologous loci in Triticum and Aegilops. *BMC Genomics* 21, 246. doi: 10.1186/s12864-020-6664-3
- Van Slageren, M. W. (1994). *Wild wheats: a monograph of Aegilops L. and Amblyopyrum (Jaub. & Spach) Eig (Poaceae)* (the Netherlands: Wageningen Agricultural University Papers).
- Volk, G. M., Byrne, P. F., Coyne, C. J., Flint-Garcia, S., Reeves, P. A., and Richards, C. (2021). Integrating genomic and phenomic approaches to support plant genetic resources conservation and use. *Plants* 10, 2260. doi: 10.3390/plants10112260
- Waines, J. G., and Barnhart, D. (1992). Biosystematic research in aegilops and triticum. *Hereditas* 116, 207–212. doi: 10.1111/j.1601-5223.1992.tb00825.x
- Wang, L., Zhu, T., Rodriguez, J. C., Deal, K. R., Dubcovsky, J., Mcguire, P. E., et al. (2021). Aegilops tauschii genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. *G3 (Bethesda)* 11. doi: 10.1093/g3journal/jkab325

Yu, G., Matny, O., Champouret, N., Steuernagel, B., Moscou, M. J., Hernández-Pinzón, I., et al. (2022). *Aegilops sharonensis* genome-assisted identification of stem rust resistance gene Sr62. *Nat. Commun.* 13, 1607. doi: 10.1038/s41467-022-29132-8

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci.* 114, 9326–9331. doi: 10.1073/pnas.1701762114

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., et al. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *Plant J. Cell Mol. Biol.* 107, 303–314. doi: 10.1111/tpj.15289

Zohary, D., and Imber, D. (1963). Genetic dimorphism in fruit types in *Aegilops speltoides*. *Heredity* 18, 223–231. doi: 10.1038/hdy.1963.24



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

REVIEWED BY

Bernadette Julier,
INRAE Nouvelle Aquitaine Poitiers, France
Rakesh Kumar,
University of California, Berkeley,
United States

*CORRESPONDENCE

Subhash Chand

✉ subhashchand5415@gmail.com

Ajoy Kumar Roy

✉ royak3333@gmail.com

RECEIVED 17 July 2023

ACCEPTED 09 October 2023

PUBLISHED 26 October 2023

CITATION

Chand S, Roy AK, Singh T, Agrawal RK,
Yadav VK, Kumar S, Malaviya DR,
Chandra A and Yadava DK (2023) Twenty-
four years lucerne (*Medicago sativa* L.)
breeder seed production in India: a
retrospective study.
Front. Plant Sci. 14:1259967.
doi: 10.3389/fpls.2023.1259967

COPYRIGHT

© 2023 Chand, Roy, Singh, Agrawal, Yadav,
Kumar, Malaviya, Chandra and Yadava. 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.

Twenty-four years lucerne (*Medicago sativa* L.) breeder seed production in India: a retrospective study

Subhash Chand ^{1*}, Ajoy Kumar Roy ^{1*}, Tejveer Singh¹,
Rajiv Kumar Agrawal¹, Vijay Kumar Yadav¹, Sanjay Kumar²,
Devendra Ram Malaviya¹, Amaresh Chandra¹
and Devendra Kumar Yadava³

¹ICAR-Indian Grassland and Fodder Research Institute, Jhansi, India, ²ICAR-National Research Centre on Seed Spices, Ajmer, India, ³ICAR-Indian Agricultural Research Institute, New Delhi, India

Lucerne (*Medicago sativa* L.) is the second most significant winter leguminous fodder crop after berseem in India. Breeder seed (BS) is the first stage of the seed production chain, as it is the base material for producing foundation and certified seeds. In India, lucerne BS demand has been reduced by 85.58% during the last 24 years (1998–1999 to 2021–2022), declining from 2150 kg to 310 kg. Out of 14 varieties released and notified so far, only nine varieties entered the seed chain since 1998–1999. It shows narrow varietal diversification and, hence, needs robust breeding programs towards enriching genetic variability and varietal development. The present study also highlights the disparity in BS demand and production over the years and puts forth the possible reasons behind the reduction in BS demand and production in the country. Out of the nine varieties, the BS demand of Anand-2 (53.11%) was highest, followed by Type-9 (19.44%) and RL-88 (13.60%). Varietal replacement rate (VRR) was found to be moderate, i.e., 23.67% for the varieties having <5 years old age in the last 3 years (2019–2020 to 2021–2022). It has also been estimated that BS produced (233 kg) during 2021–2022 can cover the approximate area of 6,300 ha at farmers' fields in 2024–2025 if the seed chain functions 100%, effectively. The present study provides a holistic overview of lucerne BS demand and production, challenges in BS production, and the way forward to develop more varieties and surplus BS production in the country.

KEYWORDS

breeder seed, lucerne, quality seed, varietal replacement rate, fodder crop

1 Introduction

Lucerne (*Medicago sativa* L.), also known as alfalfa or *rijika* in Hindi, is an important crop in the temperate agro-climatic regions globally and has the highest productivity of feed protein per unit area (Annicchiarico et al., 2010; 2015). It is an autotetraploid crop with $2n=4x=32$ originating in north-western Iran and north-eastern Turkey (Irwin et al., 2001; Wang and Şakiroğlu, 2021). It is mainly preferred for hay production and quality pasture for livestock due to its high protein content (Gawel, 2008; Babu et al., 2014). Lucerne, as a fodder crop, has better adaptability over the other grasses and legumes due to its nutritional superiority, having a high content of proteins, vitamins, and minerals, high green fodder, and good atmospheric nitrogen fixation (Lamb et al., 2006; Bouton, 2012). However, the genetic gain of forage yield in lucerne is low (0.2%–0.3% per year) as compared to maize [*Zea mays* L., 2%] and white clover [*Trifolium repens* L., 1%] (Woodfield and Brummer, 2001; Lamb et al., 2006). The genetic gain is influenced by various factors such as autotetraploid nature, perennial growth habit, highly cross-pollinating system, high level of non-additive genetic variance due to gene interaction, and genotype-by-environment interaction (GEI) (Woodfield, 1999; Kapadia, 2019).

In India, lucerne is predominantly cultivated in subtropical and tropical climatic conditions as a major *Rabi* fodder crop and is estimated to cover 1.0 Mha area (Chauhan et al., 2017). *Rabi* is the winter season in India, where crops are sown in October–November and are harvested in April–June. The states having the highest area under cultivation are Gujarat, Rajasthan, Maharashtra, Punjab, Haryana, Madhya Pradesh, Uttar Pradesh, Tamil Nadu, and Karnataka (Roy et al., 2020). In the northern and central states like Uttar Pradesh, Madhya Pradesh, Haryana, and Punjab, it is taken as a multicut winter annual crop during the period October–June, whereas in western states like Gujarat, Rajasthan, and Maharashtra, it is cultivated as a perennial multicut crop with 3 years rotation.

There is a well-organized and carefully defined system of variety release and notification in India. All India Coordinated Research Projects (AICRPs) in different crops arrange multilocation and multiyear testing of all the entries contributed from different public and private sector institutions along with national, zonal, and local checks in coded form (Chand et al., 2020). In annual crops, three cycles of evaluation are followed: initial varietal trial (IVT) and advanced varietal trials (AVT-1, AVT-2) with defined promotion criteria at each stage. However, the perennial crops require 4 years—1 year for crop establishment and 3 years for evaluation under coded form in the same field under multicut system. The data of 3 years are pooled and analyzed. Various agromorphological parameters, green and dry matter yield, nutritive parameters, tolerance to major biotic stress, etc. are recorded along with seed production potential and agronomic responses such as phosphorus use efficiency, cutting and irrigation schedule. Based on cumulative 3 years' result, a decision is taken by a duly constituted Varietal Identification Committee based on merit. If identified, it is put before the Central Varietal Release Committee, a statutory body

of the Government of India. If approved, it is notified in the gazette for cultivation.

The seed chain in any crop could be sustainable and effective only when breeder seed (BS) production meets the BS demand. BS is the progeny of nucleus seed and is produced by the concerned institutions that have developed the variety like ICAR institutes or State Agricultural Universities (SAUs)/agencies with the help of Project Coordinators (PCs)/Project Directors (PDs) of AICRPs in the different crops (<https://seednet.gov.in/>). In lucerne, only public sector-bred varieties come into the seed chain system for seed multiplication; however, the varieties developed by the private sector and notified through the Central Variety Release Committee (CVRC) do not proceed into the seed chain and sell their seeds directly to the farmers. In India, BS production is the mandate and responsibility of the ICAR and DAC (Department of Agriculture, Cooperation and Farmers Welfare, Government of India) compiles the BS indents of states, union territories (UTs), public sector units (PSUs), and private seed companies in different crops and provide them to the concerned authorities of ICAR, viz., PCs/PDs for the production of the BS in each crop. The indents are then allocated to the concerned breeder/parent institute for BS production after considering factors like the availability of nucleus seed and other facilities at the center. Monitoring is done by the duly constituted committee that includes the breeder of the variety, the concerned PC/PD or his or her nominee, and one member of the National Seed Corporation (NSC) at regular intervals as per ICAR guidelines (<https://seednet.gov.in/>). AICRP on Forage Crop and Utilization (AICRP FC&U) plays a vital role in the supervision and coordination of maintenance and production of nucleus and breeder seed and their supply network, thereby indirectly helping in the production of the required quantity of foundation and certified seed in the country. Varietal replacement rate (VRR) affects crop productivity and resilience to climate-driven factors. The availability of good-quality seeds of high-yielding varieties with superior genetic purity is essential for high production under different agro-climatic conditions in any crop. Farm productivity has a significant positive connection with farmer's prosperity and livelihood, and the timely availability of high-quality seeds of high-yielding varieties plays a vital role in it (Singh, 2015; Chand et al., 2022b). Timely availability of quality seed alone can increase the yield by 15%–20%, and it may go up to 45% with proper management practices (<https://seednet.gov.in/>). Several challenges have been reported in the past, mainly related to quality seed production, which adversely affected the expansion of cultivated areas under the particular crop (Chauhan et al., 2017; 2021).

In the present study, we have analyzed the BS demand and production trend during the last 24 years (1998–1999 to 2021–2022) in India. We explained the factors affecting BS demand and production and the possible reasons for the shortfall in BS production. The VRR is also calculated in the lucerne for the first time. In addition, we have predicted the potential production of foundation and certified seed from the produced BS. The present study highlights the existing challenges in BS production. It also points towards the need to reframe our breeding programs to develop new high-yielding varieties.

2 Materials and methods

The BS indent and production data of different lucerne varieties under the seed chain in India were collected from the AICRP FC&U (Anonymous, 1999; 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019; 2020; 2021; 2022) located in ICAR–Indian Grassland and Fodder Research Institute, Jhansi (India). The raw data were compiled, analyzed, and interpreted to express vividly the status of BS demand and production of varieties in the seed chain since 1998–1999. Microsoft Excel (2013 version) was used for preparing different figures, such as BS demand and production trends, and the contribution of major institutions to BS allocation and production. In addition, VRR for the last 3 years (2019–2020 to 2021–2022) was also calculated using the following formula: $VRR = (A/B) \times 100$; where A = indent of given varieties (kg) for the calculated years; B = total indent of all varieties of the given crop (kg) for the calculated years. It expressed the contribution of recently released varieties (varietal age <5 years) in BS demand. Likewise, foundation seed, certified seed, and area under certified seed of lucerne varieties at farmer's fields were also predicted based on available 2021–2022 BS production.

3 Results

3.1 Aggregate breeder seed demand and production status

In lucerne, during the last 24 years, the BS demand gradually decreased with few fluctuations in trend, and only three to five notified varieties were in the seed chain for any particular year (Figure 1). For ease of calculation and interpretation, the 24 years were divided into six blocks of 4 years each. For instance, from 1998–1999 to 2001–2002, DAC indented 9,085 kg of lucerne varieties for BS production to AICRP FC&U and was considered base year block (Supplementary Table S1). During 2002–2003 to 2005–2006, the BS indent was 5,101 kg, a reduction of 43.85% over

the base year block. Likewise, the BS indent was reduced by 27.19%, 64.56%, 79.97%, and 79.06% over the base year block from 2006–2007 to 2009–2010, 2010–2011 to 2013–2014, 2014–2015 to 2017–2018, and 2018–2019 to 2021–2022, respectively.

In 1998–1999, BS production was 1,460 kg against the indent 2,150 kg, and the deficit was 690 kg (32.09%). However, BS indent was reduced by 85.58% in 2021–2022 compared to 1998–1999 (Figure 1). The BS production was 233 kg for the indent 310 kg and was a deficit of 77 kg (24.84%) in 2021–2022. The BS production of lucerne varieties could not match their respective allocations during the last 24 years except for a few years like 1999–2000 (+295 kg), 2001–2002 (+144 kg), 2005–2006 (+288 kg), 2009–2010 (+80 kg), 2010–2011 (+8 kg), 2016–2017 (equal), 2018–2019 (+155 kg), and 2019–2020 (+16 kg) (Figure 1). It is worth mentioning that institutions take up the BS production only after getting the demand; hence, BS production directly connects with the BS indent in any particular year. Less the BS indent would be BS production. BS production was reduced by 67.26%, 41.00%, 69.36%, 83.02%, and 78.93% during 2002–2003 to 2005–2006, 2006–2007 to 2009–2010, 2010–2011 to 2013–2014, 2014–2015 to 2017–2018 and 2018–2019 to 2021–2022, respectively, in lucerne over the base year block (Supplementary Table S1).

3.2 Varietal diversification, their BS indent, and production status

In India, only 14 varieties have been released and notified in lucerne since 1978 under the National Agricultural Research System (NARS); however, only nine varieties have been incorporated in the seed chain since 1998–1999 (Roy et al., 2020). Detailed information, viz., breeding method, mother institute, and adoption area, related to indented varieties is presented in Table 1. However, only three to five lucerne varieties were in the seed chain in any particular calendar year during the last 24 years (Figure 1), and varietal BS demand and production over the years are presented in Supplementary Table S2. Varietal diversification is essential for increasing crop production, suitability of a variety in a specific

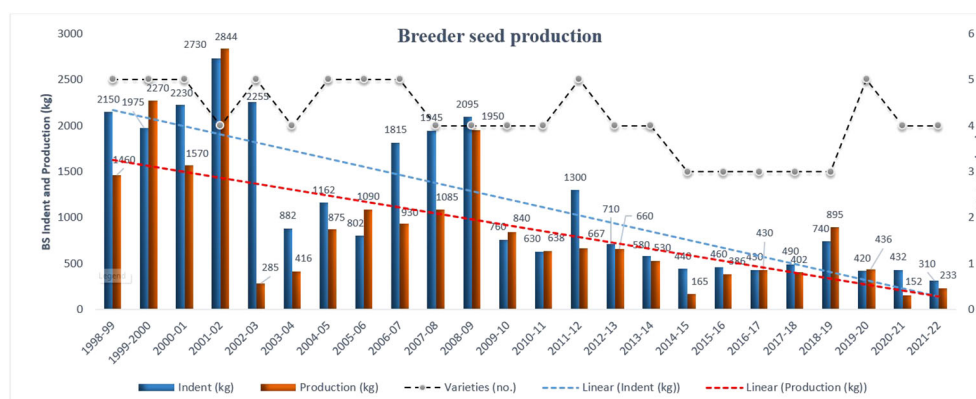


FIGURE 1

Lucerne breeder seed indent, production, and number of varieties over the years in seed chain during the last 24 years in India. Dotted blue and red lines are linear regression and express the declining trend of both BS indent and production, respectively.

TABLE 1 Detailed description of the lucerne varieties indented during the last 24 years in India (Data source: Roy et al., 2020).

S.N.	Variety*	Year of notification	Breeding method/ source	Parent institute	GFY (000 kg/ ha)	Area of adoption	Specific features
1.	Type-9	1978	Mass selection from the lucerne germplasm	CCSHAU, Hisar	45–50	Haryana, Rajasthan, Gujarat, Himachal Pradesh and Delhi under irrigated conditions	Perennial
2.	CO-1	1982	Mass selection from local material of Coimbatore	TNAU, Coimbatore	100–120	Tamil Nadu	Perennial, free from Cuscuta
3.	LLC-5	1984	Selected clones from Kutchh area of Gujarat after three cycles of recurrent selection	PAU, Ludhiana	70–75	Gujarat, Maharashtra, Andhra Pradesh, Rajasthan, Himachal Pradesh, Punjab and Haryana	Annual, moderately resistant to downy mildew
4.	Anand-2	1984	Pure line selection from the material collected from Kutchh areas of Gujarat	GAU, Banaskantha	70–75	Gujarat, Rajasthan and Maharashtra	Annual type; vigorous growth, suitable for seasonal and annual cultivation
5.	Anand-3	1995	Introduction from GAU, Anand (Gujarat)	CSKHPKV, Palampur	45–50	Cold dry zone of Kinnaur and Lahul and Spiti valley of Himachal Pradesh	Perennial, resistant to logging and frost and highly responsive to Phosphate fertilizer
6.	RL-88	1996	Selection from Ahmednagar local Lucerne	MPKV, Rahuri	100–120	Lucerne growing irrigated areas in the country	Perennial, quick re-growth, more vigorous than any other varieties
7.	AL-3	2009	Pure line selection from the material collected from Kutchh areas of Gujarat	AAU, Anand	100–120	Sub-tropical areas of Gujarat and Maharashtra	Perennial, oblong dark green leaves; free from major diseases
8.	RBB 07-01	2016	Composite of seven cultivars	SKRAU, Bikaner	160–180	North West zone of India	Perennial, high crude protein
9.	TNLC-14	2019	Polycross derivative involving CO 1	TNAU, Coimbatore	45–50	Telangana, Andhra Pradesh, Tamil Nadu, Karnataka	Perennial

*Type-9 is also known as T-9, LLC-5 as LL composite-5, RBB 07-01 as Krishna; Anand-2 as GAUL-1, RL-88 as RLS-88, CO-3 as TNLC-14; GFY: potential green fodder yield (000 kg/ha).

cropping system, more buffering capacity to biotic and abiotic stresses, and more choice for the farmers based on their available resources.

Five lucerne varieties were in the seed chain during 1998–1999 to 2001–2002, and BS indent was highest for Anand-2, followed by Type-9 and RL-88 contributing 42.71%, 35.17%, and 10.40%, respectively (Table 2). Likewise, BS indent was maximum for Anand-2 followed by Type-9 and contributed 40.07% and 23.88%, respectively, during 2002–2003 to 2005–2006. Variety Anand-2 had maximum BS indent and shared more than 50% contribution to the total BS indent after each 4-year interval since 2006–2007. Overall, Anand-2 (53.11%) had contributed the highest share in BS indent, followed by Type-9 (19.44%) and RL-88 (13.60%) during the last 24 years.

As far as varietal BS production is concerned, Anand-2 contributed maximum, followed by Type-9 from 1998–1999 to 2005–2006 (Table 2). Likewise, Anand-2 contributed maximum, followed by RL-88 to the total BS production during 2006–2007 to 2009–2010, 2010–2011 to 2013–2014, and 2018–2019 to 2021–2022, respectively. Anand-3 (7.59%) contributed second highest after Anand-2 (86.04%) during 2014–2015 to 2017–2018. Overall, Anand-2 had the highest share with the value of 68.67%, followed by RL-88 (11.96%) and Type-9 (10.20%) in the total BS production during the last 24 years.

3.3 Disparity in varietal BS indent and production

Allocated production centers failed to meet the BS production targets against their indent in lucerne since 1998–1999 (Table 3). For instance, allocated centers produced 8,144 kg BS against the indent 9,085 kg with a net deficit of 941 kg (10.36%) during 1998–1999 to 2001–2002 (Supplementary Table S1). Likewise, BS production was net deficit of 2,435 kg (47.74%) as against the total BS indent, i.e., 5,101 kg from 2002–2003 to 2005–2006. However, after that, the gap between BS production and BS indent was narrowed down by extensive efforts of the ICAR, AICRP FC&U, concerned breeders, and parent institutes. For example, BS production was a net deficit of 27.36%, 22.52%, 24.01%, and (9.78%) against the allocated quantity from 2006–2007 to 2009–2010, 2010–2011 to 2013–2014, 2014–2015 to 2017–2018 and 2018–2019 to 2021–2022, respectively. The varietal BS production scenario indicates that only Anand-2 was produced in surplus (+1,469 kg) against the indent (3,880 kg), whereas other indented varieties could not meet the BS indent during 1998–1999 to 2001–2002 (Table 2). For instance, Type-9 had a deficit of 1,755 kg against the BS indent of 3,195 kg. From 2002–2003 to 2005–2006, BS production was less than its demand in all indented lucerne varieties; for instance, there was a net deficit of 719 kg and 702 kg against the indent of 2,044 kg and 1,218 kg in Anand-2 and Type-9, respectively.

TABLE 2 Varietal breeder seed indent and production status of lucerne varieties and their contribution after every 4 years during the last 24 years in India.

Years	Status*	Type-9	CO-1	LLC-5	Anand-2	Anand-3	RL-88	AL-3	RBB 07-01	CO-3
1998–1999 to 2001–2002	Indent	3,195 (35.17)	770 (8.48)	295 (3.25)	3,880 (42.71)		945 (10.40)			
	Production	1,440 (17.68)	460 (5.65)	90 (1.11)	5,349 (65.68)		805 (9.88)			
	+/-	-1,755	-310	-205	+1,469		-140			
2002–2003 to 2005–2006	Indent	1,218 (23.88)	646 (12.66)	123 (2.41)	2,044 (40.07)		1070 (20.98)			
	Production	516 (19.35)	285 (10.69)	65 (2.44)	1,325 (49.70)		475 (17.82)			
	+/-	-702	-361	-58	-719		-595			
2006–2007 to 2009–10	Indent	780 (11.79)	115 (1.74)		4,355 (65.84)	70 (1.06)	1,195 (18.07)	100 (1.51)		
	Production	90 (1.87)	110 (2.29)	20 (0.42)	3,895 (81.06)		590 (12.28)	100 (2.08)		
	+/-	-690	-5	20	-460	-70	-605			
2010–2011 to 2013–14	Indent	200 (6.21)	300 (9.32)		1,750 (54.35)	50 (1.55)	410 (12.73)	510 (15.84)		
	Production	118 (4.73)	100 (4.01)		1750 (70.14)		310 (12.42)	217 (8.70)		
	+/-	-82	-200		-	-50	-100	-293		
2014–2015 to 2017–2018	Indent				1,400 (76.92)	280 (15.38)	70 (3.85)	70 (3.85)		
	Production				1,190 (86.04)	105 (7.59)	73 (5.28)	15 (1.08)		
	+/-				-210	-175	+3	-55		
2018–2019 to 2021–2222	Indent				1,305 (68.61)	10 (0.53)	82 (4.31)	230 (12.09)	185 (9.73)	90 (4.73)
	Production				1,055 (61.48)	10 (0.58)	283 (16.49)	230 (13.40)	73 (4.25)	65 (3.79)
	+/-				-250	-	+201	-	-112	-25
Total	Indent	5,393 (19.44)	1831 (6.60)	418 (1.51)	14,734 (53.11)	410 (1.48)	3,772 (13.60)	910 (3.28)	185 (0.67)	90 (0.32)
	Production	2,164 (10.20)	955 (4.50)	175 (0.83)	14,564 (68.67)	115 (0.54)	2,536 (11.96)	562 (2.65)	73 (0.34)	65 (0.31)
	+/-	-3,229	-876	-243	-170	-295	-1,236	-348	-112	-25
	% change	-59.87	-47.84	-58.13	-1.15	-71.95	-32.77	-38.24	-60.54	-27.78

*+/- indicates surplus or deficit BS production (kg) compared to allocation; values in parentheses represents the percent contribution of each variety to the total BS indent and production.

Except for LLC-5, BS demand for other indented varieties could not be met from 2006–2007 to 2009–2010. For example, Type-9 and RL-88 had a net deficit of 690 kg and 605 kg against their BS indent, respectively. Similarly, BS production was less than BS indent for all indented varieties except Anand-2, where production was equal to indent (1,750 kg) during 2010–2011 to 2013–2014. From 2014–2015 to 2017–2018, allocated centers could not meet the BS demand for indented varieties except RL-88, which was in surplus (+3.0 kg) against the allocation. BS demand was fulfilled for Anand-3 and AL-3, whereas RL-88 was produced in surplus (+201 kg) during 2018–2019 to 2021–2022. Overall, the BS demand for lucerne varieties

could not be fulfilled during the last 24 years. However, the total BS production of the most popular variety, i.e., Anand-2, was 14,564 kg against the indent (14,734 kg), and only 170 kg was a deficit.

3.4 BS producing centers, indent, and production status

Only three SAUs, Anand Agriculture University (AAU)–Anand, Mahatma Phule Krishi Vidyapeeth (MPKV)–Rahuri, and Tamil Nadu Agriculture University (TNAU)–Coimbatore, have

TABLE 3 The breeder seed allocation and production status of indented lucerne varieties to the different production centres and their net BS balance during the last 24 years in India (data source: Anonymous, 1999; 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019; 2020; 2021; 2022).

Variety	Allocated center*	Allocation (kg)	Production (kg)	Net deficit (kg)	Deficit (%)
Anand-2	AAU, Anand	14,734	14,564	-170	-1.15
CO-1	TNAU, Coimbatore	1,831	955	-876	-47.84
LLC-5	PAU, Ludhiana	418	175	-243	-58.13
RL-88	MPKV, Rahuri	3,772	2,536	-1,236	-32.77
Type-9	CCSHAU, Hisar	380	208	-172	-45.26
	IGFRI, Jhansi	30	0	-30	-100.00
	MPKV Rahuri	400	0	-400	-100.00
	UAS Bangalore	170	0	-1.70	-100.00
	AAU, Anand	1,218	516	-702	-57.64
	Gandhinagar, RSFPD	2,175	1,040	-1,135	-52.18
	NDDB, Anand	1,002	400	-602	-60.08
Anand-3	AAU, Anand	400	175	-225	-56.25
	CSKHPKV, Palampur	70	0	-70	-100.00
AL-3	AAU, Anand	850	502	-348	-40.94
RBB-07-01	SKRAU, Bikaner	185	73	-112	-60.54
CO-3	TNAU, Coimbatore	90	65	-25	-27.78
	Total	27,743	21,209	-6,534	-23.55

*AAU, Anand Agricultural University, Anand; TNAU, Tamil Nadu Agricultural University, Coimbatore; PAU, Punjab Agricultural University, Ludhiana; MPKV, Mahatma Phule Krishi Vidyapeeth, Rahuri; Chaudhary Charan Singh Haryana Agricultural University, Hisar; IGFRI, Indian Grassland and Fodder Research Institute, Jhansi; UAS, University of Agricultural Sciences, Bangalore; RSFPD, Regional Station for Forage Production and Demonstration, Gandhinagar; NDDB, National Dairy Development Board, Anand; CSKHPKV, Chaudhary Sarwan Kumar Himachal Pradesh Krishi Vishwavidyalaya, Palampur, SKRAU, Swami Keshwanand Rajasthan Agricultural university, Bikaner.

consistently participated and also shared 91.06% to the total BS indent in lucerne from 1998–1999 to 2021–2022 (Supplementary Figure S1). The AAU–Anand has maximum contribution (74.29%) to the BS indent, followed by MPKV–Rahuri (11.96%) and TNAU–Coimbatore (4.81%) from 1998–1999 to 2021–2022. Overall, allocating centers, viz., AAU–Anand, MPKV–Rahuri, TNAU–Coimbatore, RSFPD–Gandhinagar, and NDDB–Anand could not meet the BS demand in lucerne from 1998–1999 to 2021–2022.

The AAU–Anand could not meet the BS demand of lucerne varieties; for example, Anand-2 had a deficit of 170 kg, Type-9 of 702 kg, Anand-3 of 225 kg, and AL-3 of 348 kg against the BS indent during the last 24 years (Table 3). Similarly, TNAU–Coimbatore's production figures also indicated less CO-1 (876 kg) and CO-3 (25 kg) against their allocation. There was a deficit production of 243 kg against the indent (418 kg) for variety LLC-5 by PAU–Ludhiana and a deficit of 1,236 kg against the indent (3772 kg) for variety RL-88 by MPKV–Rahuri. Similarly, other centers could not produce BS in sufficient amounts to meet the BS demand of the allocated varieties (Table 3).

3.5 Varietal replacement rate

Since 2017–2018, the percent share of old varieties (>5 years) has declined substantially and *vice versa* for <5-year-old varieties (Figure 2).

Two varieties, having less than 5 years of varietal age, were indented (275 kg) and shared 23.67% of the total indent (Table 4). Likewise, the varieties having <15 years of age (only three varieties) shared 38.30%; however, the varieties having more than 15 years of age (only three varieties) still had a 61.70% contribution to the total BS indent.

3.6 Prediction of foundation and certified seeds

Foundation seed is the progeny of BS, whereas certified seed is derived from foundation seed. Certified seed, also known as commercial seed, is available to farmers with 99% genetic purity. The quantity of the foundation seed and certified seed produced was calculated based on conversion of total breeder seeds at 1:26 seed multiplication ratio (Chauhan et al., 2017). In this study, the total BS production in lucerne crop was 233 kg in four different varieties, viz., Anand-2 (175 kg), RBB-07-01 (23 kg), AL-3 (20 kg), and CO-3 (15 kg) in Rabi 2021–2022 (Supplementary Table S2). The foundation seed production would be 6058 kg in 2022–2023 if the seed chain functions 100% effectively and all the practices are followed correctly (Table 5). Likewise, certified seed production would be 157,508 kg in the subsequent year (2023–2024) and will cover 6.3 thousand hectares of area in the farmer's field under fodder lucerne.

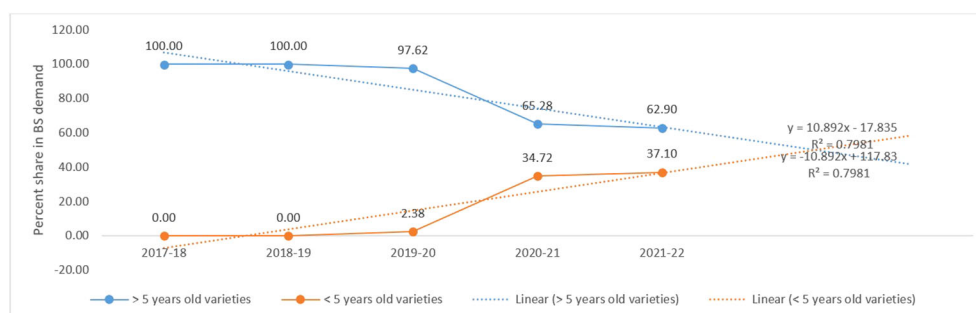


FIGURE 2

Percent contribution of lucerne varieties (<5 years and >5 years old) to the total BS indent during the last 5 years (2017–2018 to 2021–2022).

4 Discussion

Breeder seed production is vital for maintaining the long-term seed chain and is the prime responsibility of public sector institutions and agencies (Vishnu et al., 2013; Prasad et al., 2022). The Indian seed program recognizes three generation system: breeder, foundation, and certified seeds. It assures adequate quality standard in the seed multiplication chain and also maintains the genetic purity of a variety as it moves from the breeder to the last stakeholder, i.e., farmers (Mishra et al., 2022; Yadav et al., 2022). Certified seeds are commercialized and sold in the market to farmers/agencies for raising the crop and its utilization.

In lucerne, the BS demand for indented varieties has declined substantially since 1998–1999. The present study observed that BS indent has reduced by 85.58% from 1998–1999 to 2021–2022. However, increasing trends of BS demand have been reported in food crops like wheat (Krishna et al., 2016), rice (Prasad et al., 2022), barley (Vishnu et al., 2013), and pulses (Parihar and Dixit, 2016). There might be few probable but imperative reasons that could explain the declining lucerne BS demand in the country such as (1) shifting of cultivated lucerne area into different fodder crops,

(2) disparity in BS production and demand, (3) unrealistic and extremely high BS demand by the indenters, (4) extensive efforts of private seed companies in seed production and direct selling to the farmers, and (5) presence of the unorganized seed sector.

First, lucerne cultivated area has been occupied by other fodder crops, more specifically berseem, ryegrass, and multicut sorghum over time; however, berseem BS demand during the last 15 years indicated that there is not much area expansion in the berseem (Figure 3). Both lucerne and berseem crops are important leguminous fodder crops of Rabi season in India; however, lucerne is superior over berseem in nutritional quality and dry matter production (Nasrullah et al., 2015; Kaithwas et al., 2020; Roy et al., 2020). In addition, the country's berseem BS demand and production is almost constant from 2013–2014 to 2020–2021. Both the crops have different cultivable agro-ecological areas; for instance, berseem is mainly grown in the states of Punjab, Haryana, Uttar Pradesh, and parts of Rajasthan and Madhya Pradesh; however, lucerne cultivation areas are Gujrat, Rajasthan, Maharashtra, Karnataka, and Tamil Nadu. Lucerne crop is mostly preferred in those areas where the water supply is inadequate and winter period is short (<https://aicrponforagecrops.icar.gov.in/>). Furthermore, it has been observed that a few areas of lucerne

TABLE 4 Varietal replacement rate (VRR) in lucerne during last 3 years (2019–2020 to 2021–2022) in India under public sector.

Number of total notified varieties	No. of varieties in seed chain	Total BS indent (kg)	Varieties < 5 years old			Varieties < 15 years old			Varieties > 15 years old		
			No.	Indent (kg)	% share in total indent	No.	Indent (kg)	% share in total indent	No.	Indent (kg)	% share in total indent
14	6	1,162	2	275	23.67	3	445	38.30	3	717	61.70

TABLE 5 Breeder, foundation, and certified seed demand and prediction of foundation and certified seed in lucerne from the available breeder seed in India.

Crop	Seed rate (kg/ha) for		SMR*	Approx. area* (Mha)	Seed demand (kg)			Seed production (kg)			Estimated area covered (000 ha) (2024–2025)
	Fodder production	Seed production			BS (2021–2022)	FS (2022–2023)	CS (2023–2024)	BS (2021–2022)	FS (2022–2023)	CS (2023–2024)	
Lucerne	25	15	26	1.0	36,500	950,000	24,750,000	233	6,058	157,508	6.30

*According to Chauhan et al., 2017.

SMR, seed multiplication ratio; BS, breeder seed; FS, foundation seed; CS, certified seed.

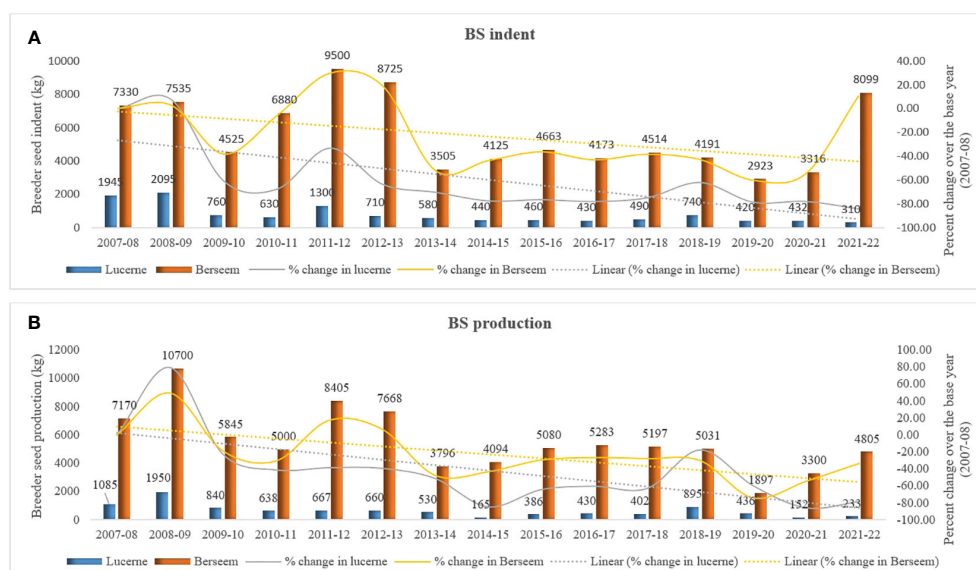


FIGURE 3
Comparison between BS demand and production of lucerne (A) and berseem (B) crops during the last 15 years (2007–2008 to 2021–2022).

(Rajasthan, Uttar Pradesh and *Malwa* region of Madhya Pradesh) have been converted into berseem cultivation due to the availability of seed with affordable lower prices. In addition, ryegrass and multicut sorghum are also encroaching on the traditional acreages of lucerne in Gujarat and Madhya Pradesh (personal observation).

Second, production centers could not produce sufficient BS to meet the demand; therefore, indenters indented less BS demand or changed the fodder crop for the same use in the subsequent years since 1998–1999. However, the production center almost fulfilled the BS demand of the most indented variety, i.e., Anand-2, in the last 24 years (Table 2). In addition, when a production center could not meet the BS demand of a particular variety, the indent gradually shifted to another notified variety. For example, BS production of Type-9 and Anand-3 was too low to meet the BS demand; therefore, the indent was shifted to Anand-2 and RL-88 in the subsequent years (Supplementary Table S2). Anand-2, being an annual cultivar, is mostly preferred in the northern parts of India such as western Punjab and Haryana where it provides high green acreage in five to six cuts spread over October to June. On the other side, RL-88—being a perennial variety—is regularly preferred by farmers in central and southern states of India like Maharashtra, Gujarat, and Tamil Nadu and could be maintained for up to 3 years in farmer's field, providing high production of green forage.

Third, indenters placed unrealistic and unwarranted BS demand, and production centers produced substantial amounts of BS against the indent; however, indenters could not uplift the BS from the centers due to various reasons for many years (personal observation). This matter was discussed at various platforms of ICAR and DAC, and thereafter, indenters started to place realistic BS demand to the DAC. Therefore, BS indent declined substantially after 2008–2009, and proper seed chain was followed for foundation and certified seed production.

Fourth, private seed companies' contribution to lucerne seed production has increased substantially over the last 15 years. More than 500 private sector companies actively participate in seed production in various crops and have a share of >80% in India's seed sale and are involved mostly in low-volume and high-value crops (Agrawal, 2012; HanChinal, 2012; Chauhan et al., 2016). However, only 10–12 private seed companies, viz., Alamdar Seeds, Foragen Seeds, and Kisan Kutch Seeds, are actively involved in lucerne seed production and marketing of un-certified seeds to the farmers. Their robust extension programs, promotion strategies, and direct participation of farmers provide an edge over government programs to disseminate their technologies to the farmer's field. Private companies contract with farmers to produce seed at a large scale, where farmers get quality seed and other inputs from seed companies (Chauhan et al., 2016). These companies purchase the farm produce at reasonable rates and sell it to the markets after processing or grading at competitive rates. Fifth, the informal seed sector, particularly farmers who produce seed at their farms, distribute it among their relatives and sell it to other growers in nearby villages (Hiremath et al., 2020).

Crop production is adversely affected by climatic conditions such as abiotic (rainfall pattern and intensity, low and high temperature, drought, salinity, alkalinity, etc.) and biotic (disease and pest infestation) factors (Joshi et al., 2007; Roy et al., 2016; Chand et al., 2022a). In India, lucerne BS productivity is low (180–250 kg/ha), and several factors such as unprecedented and erratic rainfall patterns, high isolation distance, severe inbreeding depression on selfing, increased dependency on bee visits for tripping, and high sensitivity to abiotic and biotic stresses.

The development and deployment of high-yielding and stable varieties are the need of the hour to increase the production and productivity of any crop, and VRR is an indicator of the

dissemination of genetic progress. (Singh et al., 2020; Chauhan et al., 2021; Prasad et al., 2022). In the present study, the contribution of recently released and notified lucerne varieties (<5 years old) is moderate, and the contribution of old varieties (>15 years old) is very high in the last 3 years (2019–2020 to 2021–2022). Wheat has the highest VRR among the crops, followed by mung bean and chickpea in India. In wheat, varieties notified during the last 5 and 10 years shared 45.3% and 74.0%, respectively, during 3 years (2017–2018 to 2019–2020) (Singh et al., 2020). In addition, the percentage share of varieties over 5 years of age has been declining gradually since 2017–2018, and the contribution might be more than 50% in the coming years at this pace. India requires 36,500 kg BS per annum, if 100% seed replacement rate is followed, to cover the existing 1 Mha area (Chauhan et al., 2017). However, available BS can meet only 0.6% requirement of the commercial seed, and 0.4% would be met by informal seed supply chain including private seed companies, farm saved seeds, and non-certified seeds from local markets. Therefore, there is an urgent need to develop breeding programs to improve genetic gain using conventional and non-conventional approaches and develop more genotypes with high green forage yield and nutritional superiority. Conventional approaches mean classical breeding tools like introduction, selection, hybridization, and pedigree. Non-conventional approaches include biotechnological tools such as marker-assisted selection (MAS), genetic engineering, tissue culture, embryo rescue, and genetic transformation.

5 Conclusion

In India, low productivity of forage crops is a major concern in which timely availability of sufficient quality seed is a foremost factor. The BS is a vital component of the seed chain and decides the time-bound availability of certified seed to the farmers. The BS demand for improved varieties for different agro-climatic conditions and cropping systems needs to be improved in lucerne. The BS demand will increase with the certified seed demand. The government would act to promote lucerne growing and the use of certified (public) seed. Genetic progress is another, but important, subject. In addition, it must be ensured that the production center should produce an adequate amount of BS to meet the demand.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://aicrponforagecrops.icar.gov.in>.

Author contributions

SC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AKR: Conceptualization, Data

curation, Formal analysis, Resources, Validation, Writing – review & editing, Investigation, Project administration, Supervision, Writing – original draft. RKA: Conceptualization, Formal analysis, Resources, Supervision, Writing – review & editing. TS: Conceptualization, Data curation, Validation, Visualization, Writing – review & editing. DKY: Data curation, Investigation, Project administration, Supervision, Validation, Visualization, Writing – review & editing. VKY: Conceptualization, Data curation, Writing – review & editing. AC: Conceptualization, Formal analysis, Validation, Writing – review & editing. SK: Software, Writing – review & editing. DRM: Supervision, Formal analysis, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors are thankful to the Indian Council of Agricultural Research (ICAR), ICAR–AICRP on Forage Crops and Utilization, and ICAR–Indian Grassland and Fodder Research Institute, Jhansi (UP) India, for providing all the necessary information. The authors are also very thankful to Dr. K. Ganeshan, Dr. D.P. Gohil, and Dr. Digvijay Singh for giving valuable time and guidance during the preparation 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.2023.1259967/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Histogram highlighting the total BS production against the indent at major centers and their percent contribution to the lucerne BS production centers during the last 24 years in India.

References

- Agrawal, P. K. (2012). "Indian Seed Industry: Today and its potential in next five years," in *National Seed Congress on Welfare and economic prosperity of the Indian farmers through seeds* (Raipur, Chhattisgarh), 60–69.
- 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, 1020. doi: 10.1186/s12864-015-2212-y
- Annicchiarico, P., Scotti, C., Carelli, M., and Pecetti, L. (2010). Questions and avenues for lucerne improvement. *Czech J. Genet. Plant Breed.* 46 (1), 1–13. doi: 10.17221/90/2009-CJGPB
- Anonymous (1999). *Annual Report (Rabi 1998-99), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2000). *Annual Report (Rabi 1999-2000), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2001). *Annual Report (Rabi 2000-01), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2002). *Annual Report (Rabi 2001-02), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2003). *Annual Report (Rabi 2002-03), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2004). *Annual Report (Rabi 2003-04), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2005). *Annual Report (Rabi 2004-05), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2006). *Annual Report (Rabi 2005-06), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2007). *Annual Report (Rabi 2006-07), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2008). *Annual Report (Rabi 2007-08), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2009). *Annual Report (Rabi 2008-09), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2010). *Annual Report (Rabi 2009-10), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2011). *Annual Report (Rabi 2010-11), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2012). *Annual Report (Rabi 2011-12), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2013). *Annual Report (Rabi 2012-13), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2014). *Annual Report (Rabi 2013-14), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2015). *Annual Report (Rabi 2014-15), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2016). *Annual Report (Rabi 2015-16), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2017). *Annual Report (Rabi 2016-17), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2018). *Annual Report (Rabi 2017-18), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2019). *Annual Report (Rabi 2018-19), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2020). *Annual Report (Rabi 2019-20), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2021). *Annual Report (Rabi 2020-21), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Anonymous (2022). *Annual Report (Rabi 2021-22), AICRP on Forage Crops and Utilization* (Jhansi: ICAR-Indian Grassland and Fodder Research Institute).
- Available at: <https://aicrponforagecrops.icar.gov.in/pdfs/Leucerne.pdf> (Accessed 03.04.2023).
- Available at: <https://seednet.gov.in/material/Indianseedsector.htm> (Accessed 03.04.2023).
- Babu, C., Iyanar, K., and Kalamani, K. V. A. (2014). A high yielding Lucerne variety CO-2. *Electronic J. Plant Breed.* 5 (3), 345–349.
- Bouton, J. H. (2012). Breeding lucerne for persistence. *Crop Pasture Sci.* 63 (2), 95–106. doi: 10.1071/CP12009
- Chand, S., Chandra, K., and Khatik, C. L. (2020). "Varietal release, notification and denotification system in India," in *Plant Breeding-Current and Future Views*. Ed. I. Y. Abdurakhmonov (London, United Kingdom: IntechOpen), 1–12. doi: 10.5772/intechopen.94212
- Chand, S., Singh, N., Prasad, L., Nanjundan, J., Meena, V. K., Chaudhary, R., et al. (2022a). Inheritance and Allelic Relationship among Gene (s) for White Rust Resistance in Indian Mustard [*Brassica juncea* (L.) Czern & Coss]. *Sustainability* 14 (18), 11620. doi: 10.3390/su141811620
- Chand, S., Singhal, R. K., and Govindasamy, P. (2022b). Agronomical and breeding approaches to improve the nutritional status of forage crops for better livestock productivity. *Grass Forage Sci.* 77 (1), 11–32. doi: 10.1111/gfs.12557
- Chauhan, J. S., Chand, S., Choudhury, P. R., Singh, K. H., Agarwal, R. K., Bhardwaj, N. R., et al. (2021). A scenario of breeding varieties and seed production of forage crops in India. *Indian J. Genet.* 81 (3), 343–357. doi: 10.31742/IJGPB.81.3.1
- Chauhan, J. S., Prasad, S. R., Pal, S., Choudhury, P. R., and Bhaskar, K. U. (2016). Seed production of field crops in India: Quality assurance, status, impact and way forward. *Indian J. Agric. Sci.* 86 (5), 563–579. doi: 10.56093/ijas.v86i5.58233
- Chauhan, J. S., Roy, A. K., Pal, S., Kumar, D., Choudhury, P. R., Mall, A. K., et al. (2017). Forage seed production scenario in India: Issues and way forward. *Indian J. Agric. Sci.* 87 (2), 147–158. doi: 10.56093/ijas.v87i2.67533
- Gawel, E. (2008). The effect of the way and frequency of utilization of lucerne-grass mixtures on their yield, botanical composition and quality. *Water-Environment-Rural Areas* 24, 5–18.
- HanChinal, R. R. (2012). "An overview of developments in Indian seed sector and future challenges," in *National Seed Congress on Welfare and economic prosperity of the Indian farmers through seeds* (Raipur, Chhattisgarh), 1–12.
- Hiremath, U., Gowda, B., Ganiger, B. S., and Lokesh, G. Y. (2020). Role of formal and informal seed sector in augmenting seed replacement rate in Raichur district of Karnataka, India. *Int. J. Curr. Microb. Appl. Sci.* 9 (6), 182–186. doi: 10.20546/ijemas.2020.906.230
- Irwin, J. A. G., Lloyd, D. L., and Lowe, K. F. (2001). Lucerne biology and genetic improvement—an analysis of past activities and future goals in Australia. *Aust. J. Agric. Res.* 52 (7), 699–712. doi: 10.1071/AR00181
- Joshi, A. K., Mishra, B., Chatrath, R., Ortiz Ferrara, G., and Singh, R. P. (2007). Wheat improvement in India: present status, emerging challenges and future prospects. *Euphytica* 157 (3), 431–446. doi: 10.1007/s10681-007-9385-7
- Kaithwas, M., Singh, S., Prusty, S., Mondal, G., and Kundu, S. S. (2020). Evaluation of legume and cereal fodders for carbohydrate and protein fractions, nutrient digestibility, energy and forage quality. *Range Manage. Agroforestry* 41 (1), 126–132.
- Kapadia, V. N. (2019). Breeding for high forage yield in lucerne. *Genet. Plant Breed.*, 1–33.
- Krishna, V. V., Spielman, D. J., and Veettil, P. C. (2016). Exploring the supply and demand factors of varietal turnover in Indian wheat. *J. Agric. Sci.* 154 (2), 258–272. doi: 10.1017/S0021859615000155
- Lamb, J. F., Sheaffer, C. C., Rhodes, L. H., Sulc, R. M., Undersander, D. J., and Brummer, E. C. (2006). Five decades of alfalfa cultivar improvement: Impact on forage yield, persistence, and nutritive value. *Crop Sci.* 46 (2), 902–909. doi: 10.2135/cropsci2005.08-0236
- Mishra, C. N., Sharma, A., Kamble, U., Singh, S. K., and Singh, G. P. (2022). "Accelerating varietal replacement in wheat through strengthening of seed systems," in *New Horizons in Wheat and Barley Research: Global Trends, Breeding and Quality Enhancement* (Singapore: Springer), 63–79.
- Nasrullah, M., Javed, K., Bhatti, J. A., Khosa, A. N., Marghazani, I. B., Sales, J., et al. (2015). Nutrient intake and digestibility of various winter fodders fed to beetal goats and lohi sheep. *J. Anim. Plant Sci.* 25 (4), 1206–1209.
- Parihar, A. K., and Dixit, G. P. (2016). Varietal spectrum of seed production of pulses in India: an updated approach. *Proc. Natl. Acad. Sci. India Section B: Biol. Sci.* 86, 247–252. doi: 10.1007/s40011-014-0456-y
- Prasad, G. S., Rao, C. S., Suneetha, K., Muralidharan, K., and Siddiq, E. A. (2022). Impact of breeder seed multiplication and certified quality seed distribution on rice production in India. *CABI Agric. Biosci.* 3 (1), 33. doi: 10.1186/s43170-022-00099-2
- Roy, A. K., Agrawal, R. K., Chand, S., Ahmad, S., Kumar, R. V., Mall, A. K., et al. (2020). "Database of forage crop varieties: 2020," in *AICRP on Forage Crops & Utilization*, vol. 362. (Jhansi, Uttar Pradesh, India: ICAR-IGPRI).
- Roy, A. K., Malaviya, D. R., and Kaushal, P. (2016). Genetic improvement of fodder legumes especially dual purpose pulses. *Indian J. Genet.* 76 (4), 608–625. doi: 10.5958/0975-6906.2016.00076.6

- Singh, R. P. (2015). Varietal replacement rates among field crops: current status, constraints, impact, challenges and opportunities for the Indian seed industry. *Seed Times (National Seed Assoc. India)* 7 (3), 71–89.
- Singh, R. P., Chintagunta, A. D., Agarwal, D. K., Kureel, R. S., and Kumar, S. J. (2020). Varietal replacement rate: Prospects and challenges for global food security. *Global Food Secur.* 25, 100324. doi: 10.1016/j.gfs.2019.100324
- Vishnu, K., Raj, K., Verma, R. P. S., Ajay, V., and Indu, S. (2013). Recent trends in breeder seed production of barley (*Hordeum vulgare*) in India. *Indian J. Agric. Sci.* 83 (5), 576–578.
- Wang, Z., and Şakiroğlu, M. (2021). “The origin, evolution, and genetic diversity of Alfalfa,” in *The Alfalfa Genome*. Eds. L. X. Yu and C. Kole (Cham: Springer). doi: 10.1007/978-3-030-74466-3_3
- Woodfield, D. R. (1999). “Genetic improvements in New Zealand forage cultivars,” in *Proceedings of the New Zealand Grassland Association*. (Dunedin, New Zealand). 61, 3–7. Available at: https://www.grassland.org.nz/publications/nzgrassland_publication_485.pdf
- Woodfield, D. R., and Brummer, E. C. (2001). “Integrating molecular techniques to maximise the genetic potential of forage legumes,” in *Molecular Breeding of Forage Crops: Proceedings of the 2nd International Symposium, Molecular Breeding of Forage Crops, Lorne and Hamilton, Victoria, Australia, November 19–24, 2000* (Netherlands: Springer), 51–65.
- Yadav, R. N., Kumar, P. R., Hussain, Z., Yadav, S., Lal, S. K., Kumar, A., et al. (2022). “Maintenance breeding,” in *Fundamentals of Field Crop Breeding*. Eds. D. K. Yadava, H. K. Dikshit, G. P. Mishra and S. Tripathi (Singapore: Springer Nature Singapore), 703–744.



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

REVIEWED BY

Zhiqiang Wu,
Kunpeng Institute of Modern Agriculture at
Foshan, China
Linchun Shi,
Chinese Academy of Medical Sciences and
Peking Union Medical College, China

*CORRESPONDENCE

Zinian Wu

✉ wuzinian@caas.cn

Zhiyong Li

✉ lizhiyong@caas.cn

RECEIVED 25 July 2023

ACCEPTED 20 November 2023

PUBLISHED 08 December 2023

CITATION

Liu Q, Wu Z, Tian C, Yang Y, Liu L, Feng Y
and Li Z (2023) Complete mitochondrial
genome of the endangered *Prunus*
pedunculata (Prunoideae, Rosaceae)
in China: characterization
and phylogenetic analysis.
Front. Plant Sci. 14:1266797.
doi: 10.3389/fpls.2023.1266797

COPYRIGHT

© 2023 Liu, Wu, Tian, Yang, Liu, Feng and Li.
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.

Complete mitochondrial genome of the endangered *Prunus pedunculata* (Prunoideae, Rosaceae) in China: characterization and phylogenetic analysis

Qian Liu¹, Zinian Wu^{1,2*}, Chunyu Tian¹, Yanting Yang¹,
Lemeng Liu¹, Yumei Feng¹ and Zhiyong Li^{1,2*}

¹Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot, China, ²Key Laboratory of Grassland Resources and Utilization of Ministry of Agriculture, Hohhot, China

Introduction: *Prunus pedunculata* (Prunoideae: Rosaceae), a relic shrub with strong resistance and multiple application values, is endangered in China. Extensive research had been devoted to gene expression, molecular markers, plastid genome analysis, and genetic background investigations of *P. pedunculata*. However, the mitochondrial genome of this species has not been systematically described, owing to the complexity of the plant mitogenome.

Methods: In the present research, the complete mitochondrial genome of *P. pedunculata* was assembled, annotated, and characterized. The genomic features, gene content and repetitive sequences were analyzed. The genomic variation and phylogenetic analysis have been extensively enumerated.

Results and discussion: The *P. pedunculata* mitogenome is a circular molecule with a total length of 405,855 bp and a GC content of 45.63%, which are the smallest size and highest GC content among the known *Prunus* mitochondrial genomes. The mitogenome of *P. pedunculata* encodes 62 genes, including 34 unique protein-coding genes (PCGs, excluding three possible pseudogenes), three ribosomal RNA genes, and 19 transfer RNA genes. The mitogenome is rich in repetitive sequences, counting 112 simple sequence repeats, 15 tandem repeats, and 50 interspersed repetitive sequences, with a total repeat length of 11,793 bp, accounting for 2.91% of the complete genome. Leucine (Leu) was a predominant amino acid in PCGs, with a frequency of 10.67%, whereas cysteine (Cys) and tryptophan (Trp) were the least adopted. The most frequently used codon was UUU (Phe), with a relative synonymous codon usage (RSCU) value of 1.12. Selective pressure was calculated based on 20 shared PCGs in the mitogenomes of the 32 species, most of which were subjected to purifying selection ($K_a/K_s < 1$), whereas *ccmC* and *ccmFn* underwent positive selection. A total of 262 potential RNA editing sites in 26 PCGs were identified. Furthermore, 56 chloroplast-derived fragments were ascertained in the mitogenome, ranging from 30 to 858 bp, and were mainly located across IGS (intergenic spacer) regions or rRNA genes. These findings verify the occurrence of intracellular gene transfer events from the chloroplast to the mitochondria. Furthermore, the phylogenetic relationship of

P. pedunculata was supported by the mitogenome data of 30 other taxa of the Rosaceae family. Understanding the mitochondrial genome characteristics of *P. pedunculata* is of great importance to promote comprehension of its genetic background and this study provides a basis for the genetic breeding of *Prunus*.

KEYWORDS

Prunus pedunculata, endangered plants, mitochondrial genome, gene transfer, RNA editing, phylogenetic analysis

1 Introduction

Prunus pedunculata Pall. (Prunoideae, Rosaceae), the longstalk almond, also known as *Amygdalus pedunculata* Pall., is a nationally endangered relic shrub (Ministry of Ecology and Environment of the People's Republic of China and Science, 2020; Yan et al., 2022) mainly distributed in the desert and mountain lands of arid and semi-arid regions in northwest China, Mongolia, and Russia (Wang et al., 2018c; He et al., 2021). Due to being constantly exposed to extreme climates in these regions, *P. pedunculata* has evolved great adaptability and resistance to water deficiency, low temperature, high wind, and barren soil, making it an optimal species for environmental restoration and sand fixation (Chu et al., 2013). *P. pedunculata* is an excellent oil-bearing plant and its seeds are rich in unsaturated fatty acids, which have high antihyperlipidemic and antioxidant activities (Gao et al., 2016). Longstalk almond nuts also contain health-promoting compounds, such as phytosterols, polyphenols, amygdalin (Chau and Wu, 2006), vitamins, minerals, and the essential amino acids (Gao et al., 2016; Wang et al., 2018b); thus, they possess great nutritional and medicinal value. *P. pedunculata* is also a valuable germplasm resource for wild fruit and feed plants (Wang et al., 2018c). Notwithstanding its multiple application values, *P. pedunculata* has become endangered due to extreme environmental conditions and anthropogenic activities, such as overexploitation, overgrazing, and environmental pollution (Chu et al., 2017). *P. pedunculata* has attracted significant attention ever since it was identified as a key protected wild plant (Class III) and endangered plant (Class II) of Inner Mongolia in the 1990s (Zhao, 1992; Chu et al., 2017). In addition, *P. pedunculata* was classified as a national near-threatened species by the China Biodiversity Red List – Higher Plants (Ministry of Ecology and Environment of the People's Republic of China and Science, 2020) and been assessed by the IUCN Red List of Threatened Species (Rhodes and Maxted, 2016). Extensive investigations have focused on chloroplast (cp) genome analysis (Duan et al., 2020; Du et al., 2021), chemical compounds (Ma, 2013; Lu et al., 2018; Yao et al., 2018; Li et al., 2020), genetic diversity (Zuo, 2016; Bao et al., 2021), resistance to various abiotic stress (Ma, 2006; Jiang, 2008; Luo, 2009; Guo, 2014), cultivation technology (Li, 2017; Wang et al., 2020b; Wang et al., 2021b), protection and utilization (Chu et al., 2015; Liu, 2017; Xiong et al., 2018; Bao et al., 2021) of *P. pedunculata*. Previous studies have indicated that *P. pedunculata* possesses many resistance genes. Understanding the genetic composition and phylogenetic status of

P. pedunculata is of great academic value for its conservation and application.

Mitochondria, known as “the powerhouse of the cell”, are involved not only in adenosine triphosphate (ATP) synthesis through oxidative phosphorylation (Skippington et al., 2015), but also in programmed cell death, cell signaling, male sterility, and other angiosperm bioprocesses (McBride et al., 2006), and are semi-autonomous organelles found in most eukaryotic cells (Gualberto et al., 2014). Originated from endosymbiotic events of alpha-proteobacterial 1.5 billion years ago (Mower et al., 2012), the mitochondrial genome has evolved rapidly via multiple structural variation and rearrangements and gene transfers (Wu et al., 2020b). Plant mitogenomes vary in size, gene content, and genomic configuration compared with compact animal and fungal mitogenomes (Smith and Keeling, 2015; Morley and Nielsen, 2017). Besides, some unique characteristics exist in plant mitogenomes, including uncompact gene distribution, RNA editing, gene loss, DNA sequence transfer, and exogenous sequences acquisition (Knoop et al., 2011; Mower et al., 2012; Sloan et al., 2012). Plant mitochondrial genomes are very large and vary tremendously in size, even between close relatives (Kubo and Newton, 2008). Most mitogenomes range from 200–800 kb in length. The mitogenomes of spermatophytes are the greatest in size among all organelle genomes, which can be as large as 11.7 Mb in *Siberian larch* (Putintseva et al., 2020) and 11.3 Mb in *Silene Conica* (Sloan et al., 2012), whereas the smallest mitogenome by far is only 66 Kb, found in *Viscum scurruloideum* (Skippington et al., 2015). This tremendous variation in mitogenome size is assumed to be a consequence of repetitive sequences, DNA transfer from other organisms and large intragenic segments acquisition or loss (Bergthorsson et al., 2003; Wynn and Christensen, 2019; Wu et al., 2020b). Generally, plant mitochondrial genomes are circular double-linked DNA molecules (or circularly mapping molecules); such as single circular structure in *Arabidopsis thaliana* (Sloan et al., 2018b). In addition to the typical circular structure, branched, linear and multichromosomal architectures have also been observed in *Cucumis sativus*, *Oryza sativa*, *Silene noctiflora* (Bellot et al., 2016; Kazama and Toriyama; Kozik et al., 2019; Wu et al., 2020b), as well as an extreme example *Silene Conica*, which contains numerous circular chromosomes (Sloan et al., 2012; Wu et al., 2020b). Moreover, massive occurrence of gene transfer and RNA editing may lead to the gene content variation and sequence diversity of functional protein-coding genes (Rice et al., 2013; Wu et al., 2017; Sloan et al., 2018a).

The complexity of mitogenome architectures due to genome recombination, duplication, and rearrangement makes the sequencing and assembly of mitogenomes much more complex and difficult than that of other organelle genomes (Fang et al., 2021). Hence, the full panoramic plant mitogenome description remains a bottleneck in evolutionary biology, and most plant phylogenetic studies have focused on nuclear and chloroplast genomes. Owing to the development of high-throughput sequencing technologies and the rising of next-generation phylogenomics, many software programs applicable to the mitogenome sequencing assembling were developed, such as GetOrganelle (Jin et al., 2020), Mitofiner (Allio et al., 2020), GSAT (He et al., 2023), and PMAT (<https://github.com/bichangwei/PMAT>), etc. The sequencing and assembly of mitogenome become much more accurately and efficiently.

To date (As of June 30, 2023), 895 complete plant mitogenomes have been deposited in the National Center for Biotechnology Information (NCBI) database (<https://www.ncbi.nlm.nih.gov/>), including 60 species of the Rosaceae family, among which 14 are *Prunus* species (including four cultivars). Most deposited mitogenomes maintain ‘master circle’ model (Wu et al., 2020b). Complete mitogenomes of the genus *Prunus* have been released in recent years (Pervaiz et al., 2015; Fang et al., 2021), however systematic studies have rarely been conducted. Phylogenetic relationship of *P. pedunculata* has been investigated based on morphology (Yazbek and Oh, 2013), microsatellite DNA (SSR) markers (Zhang et al., 2018), and molecular markers from the coding regions or non-coding regions of nuclear and chloroplast genes (Dong, 2015; Wang et al., 2020a), as well as chloroplast genomes (Wang et al., 2018a; Duan et al., 2020; Wang et al., 2020a). However, the phylogenetic affinities of *P. pedunculata* have not yet been determined from the perspective of the mitogenome. Elucidating the mitochondrial genome of *P. pedunculata* is a prerequisite for accurate molecular identification and genetic breeding of this endangered species. In this report, the mitogenome of this species was comprehensively assembled and analyzed. Genomic features, repetitive sequences, codon usage of PCGs, RNA editing sites, synonymous substitution rates, and DNA sequence transfer events in the *P. pedunculata* mitogenome have been extensively enumerated. Given the paucity of plant mitogenome information, the phylogenetic analysis was referring to available mitogenome data for only 30 previously annotated species of the Rosaceae family based on 20 conserved PCGs, which further clarify the evolutionary relationships and genetic background of *Prunus* species. Concurrently, the decipherment of the mitogenome enriches molecular markers and genetic resources for *Prunus* breeding and provides in-depth knowledge of organelle genome evolution.

2 Materials and methods

2.1 DNA extraction, genome sequencing, and assembly

Fresh leaves of *P. pedunculata* were collected from Hohhot, Inner Mongolia, China (40.57°N, 111.93°E) and deposited in the National Medium-Term Genebank Forage Germplasm (Hohhot,

China). Genomic DNA was extracted from fresh leaves using a Plant DNA Isolation Kit (Tiangen, Beijing, China) and sequenced using an Illumina MiSeq platform (Novogene Co., Ltd., Tianjing, China). Around 7.52 Gb clean data with 50.16 million reads were yielded and used for mitogenome *de novo* assembling. The assembly of mitochondria was performed using the software GetOrganelles V 1.7.5.3 (Jin et al., 2020) with default parameters (-R 50 -k 21,45,65,85,105,115,127 -P 1000000) (Jin et al., 2020). The accuracy of the assembly results was checked using the visualization software Bandage (Wick et al., 2015) and by mapping clean reads using Bowtie2 (Langmead and Salzberg, 2012). Afterwards, the average coverage depth was assessed to be 242.9x by SAMtools (Li et al., 2009) (Supplementary Figure S1). The coverage was visualized by Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2012). The cp genome of *P. pedunculata* was assembled in a similar manner. The complete mitochondrial genome sequence was deposited in GenBank (accession number: OQ 556854.1) and the complete chloroplast genome sequence was deposited with accession numbers of OR343251.

2.2 Genome annotation

P. pedunculata mitogenomes were annotated using GeSeq (Tillich et al., 2017) with reference to previously released mitogenome data of *Prunus* species and then manually adjusting the data into a circular mitogenome model. The cp genome was annotated using Plastid Genome Annotator (PGA) tools (Qu et al., 2019). Subsequently, Geneious V9.0.2 was used to amend mistaken codons (Kearse et al., 2012). The genome map was visualized using the Organellar Genome Draw (OGDRAW) software (Greiner et al., 2019).

2.3 Repeat sequence identification

Simple sequence repeats (SSRs) of the *P. pedunculata* mitochondrial genome were identified using the MISA software (Beier et al., 2017) with the parameters of minimum nucleotide numbers of mono-10, di-6, tri-4, tetra-3, penta-3 and hexa-3, respectively. Tandem Repeats Finder (TRF) (Benson, 1999) was used to identify tandem repeats with default parameters, whereas dispersed repeats larger than 70 bp were identified as forward, reverse, palindromic, and complementary repeats using the online tool REPuter (Kurtz et al., 2001) with a Hamming distance of 3 and a cutoff *e*-value of $1e^{-5}$.

2.4 Codon usage bias analysis

The RSCU value of PCGs and their amino acid composition were calculated by the Molecular Evolutionary Genetics Analysis software (MEGA v11.0.26) (Tamura et al., 2021), codon preferences were configured using Perl scripts.

2.5 Selective pressure calculation

Non-synonymous (K_a) and synonymous (K_s) substitution rates were calculated using DnaSP 6.12.0 (Rozaš et al., 2017), based on a total of 20 shared PCGs (*atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *ccmB*, *ccmC*, *ccmFc*, *ccmFn*, *cob*, *cox2*, *cox3*, *mttB*, *nad2*, *nad3*, *nad4*, *nad5*, *nad6*, *nad9*, and *rps13*) between the mitogenomes of *P. pedunculata* and 29 other Rosaceae species and *Oryza sativa*, *Triticum aestivum*.

2.6 Prediction of RNA editing sites

Based on three RNA-seq datasets of *P. pedunculata* deposited in the SRA database (<https://www.ncbi.nlm.nih.gov/sra/>; accession numbers: SRR13261917, SRR13261918 and SRR13261919) (Bao et al., 2021), we identified the putative RNA editing sites in mitochondrial PCGs. Mapping the RNA-seq data onto the sequences of mitochondrial PCGs by using BWA v0.7.15 (Li and Durbin, 2010) software. Then we called single nucleotide polymorphism sites (SNPs) by using SAMtools v1.17 (Li et al., 2009) and BCFtools v1.17 (Danecek et al., 2021). To identify and annotate RNA editing sites, the SNP-calling data were processed using REDO v 1.0 (Wu et al., 2018), a specialized tool for easily identifying RNA editing sites in plant organelles. To exclude the false positive RNA editing sites, BWA v0.7.15 was used to mapping the DNA-Seq data to *P. pedunculata* mitogenome. The SNP-calling method was conducted using BCFtools, then eliminating the RNA editing sites detected in genomic SNPs.

2.7 Chloroplast-derived mitochondrial sequence identification

The chloroplast genome data for *P. pedunculata* were obtained from our assemblies. Homologous sequences between the chloroplast genome and mitogenome were identified, and the transferred DNA fragments were screened using BLASTN with a cutoff value of $1e-5$. Gene transfer from the chloroplasts to mitochondria was visualized using TB tools (Chen et al., 2020).

2.8 Phylogenetic analysis

Phylogenetic analysis based on 20 shared PCGs (as mentioned in Section 2.5) derived from the complete mitogenomes of 30 selected Rosaceae species was performed, with *O. sativa* and *T. aestivum* as the outgroups. Mitogenome data of the reference accessions were downloaded from the NCBI (Supplementary Table S1). The corresponding nucleotide sequences of the PCGs in the chosen genomes were concatenated, and the MAFFT program (Katoh and Standley, 2013) was used to perform multiple sequence alignment. Both the ML algorithm and Bayesian methods were used to construct the phylogenetic tree, and the best models were selected using ModelFinder (Kalyaanamoorthy et al., 2017). The ML method was conducted

using RAxML (Stamatakis, 2006) with the GTRGAMMA model and bootstrap of 1000 replicates. Bayesian inferences (BI) using MrBayes v3.2.6. (Ronquist et al., 2012) were calculated to select the best-of-fit model GTR+I+G4.

3 Results

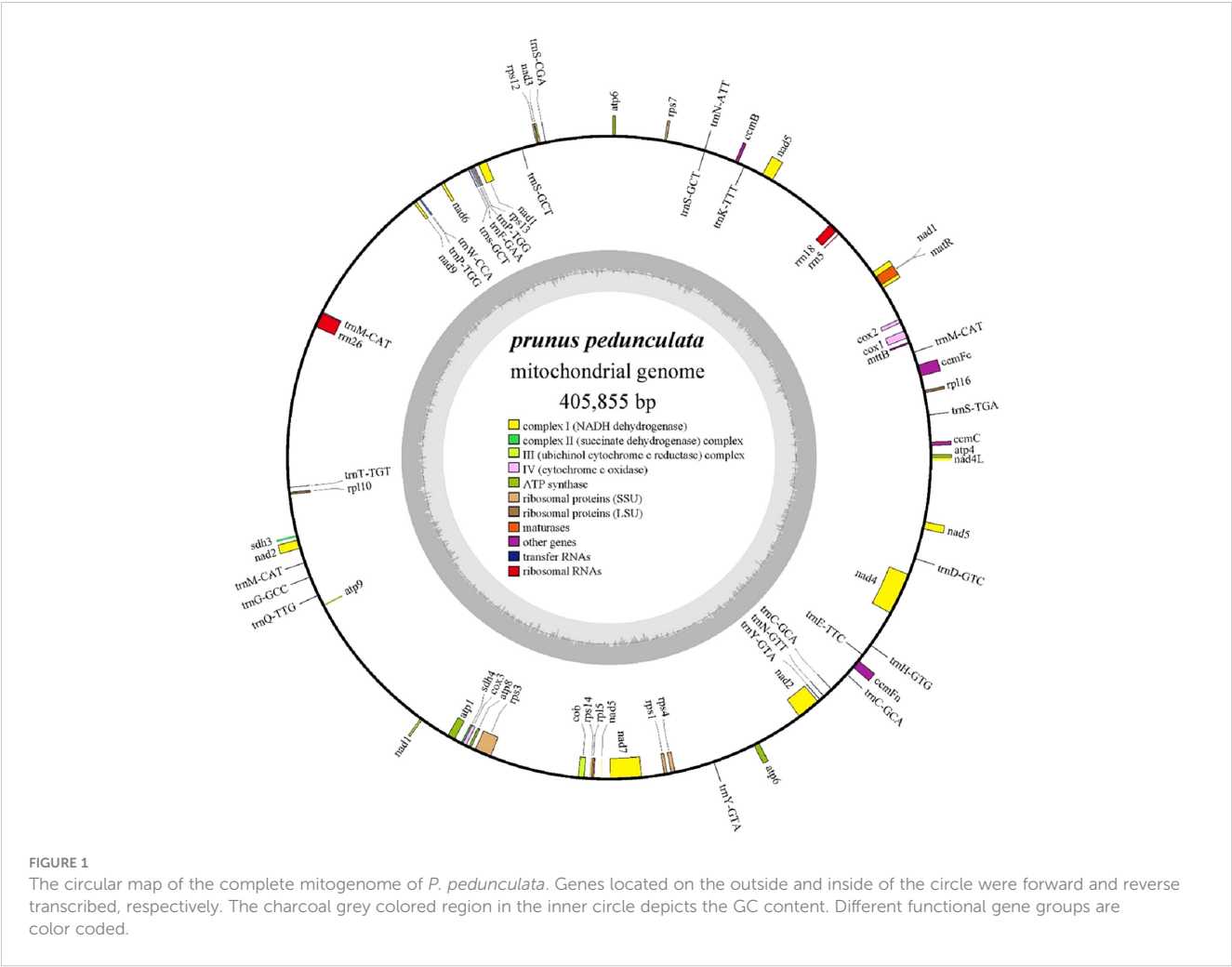
3.1 Genomic features of the *P. pedunculata* mitogenome

The complete mitochondrial genome of *P. pedunculata* was assembled into a typical single circular molecule with a size of 405,855 bp (Figure 1) and a GC content of 45.63% (Table 1). The mitogenome comprised 27.06% adenine, 27.31% thymine, 22.77% guanine, and 22.85% cytosine. Although 84.82% of the mitogenome was composed of non-coding regions, the proportion of PCGs and cis-spliced-introns was 7.39% and 6.58% in the mitogenome, respectively, and that of tRNA and rRNA accounted for 0.45% and 1.27%, respectively.

The mitochondrial genome of *P. pedunculata* contains 14 core genes, including five ATP synthase genes (*atp1*, *atp4*, *atp6*, *atp8*, *atp9*), four cytochrome C biogenesis genes (*ccmB*, *ccmC*, *ccmFc*, *ccmFn*), three cytochrome c oxidase genes (*cox1*, *cox2*, *cox3*), one ubiquinol cytochrome c reductase gene (*cob*), and one maturase gene (*matR*) (Table 2). In addition, an *atp6*-like pseudogene was detected. The genome also comprises 20 variable genes, three ribosomal RNAs (*rrn18*, *rrn26*, *rrn5*), and 19 transfer RNAs. These variable genes included nine NADH dehydrogenase genes (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*), two large ribosome protein subunits (*rpl10*, *rpl5*), six small ribosome protein subunits (*rps1*, *rps12*, *rps13*, *rps3*, *rps4*, *rps7*), one transport membrane protein (*mttB*), and two succinate dehydrogenase genes (*sdh3*, *sdh4*). In addition, the *rpl16*-like and a *rps14*-like sequences were identified as pseudogenes. As for transfer RNAs, *trnS-GCT* and *trnM-CAT* had three copies, while *trnC-GCA*, *trnP-TGG*, and *trnY-GTA*, had two copies. Other tRNA genes were represented by identical copies in the mitogenome. Intron sequences were identified in nine genes (Table 2), among which *nad1*, *nad2*, *nad5*, and *nad7* contained four introns; *nad4* contained three introns; and *ccmFc*, *cox2*, *rps3*, and one copy of *trnY-GTA* possessed one intron region.

3.2 Anatomization of repeat sequences

Repeat sequences included SSRs, tandem repeats, and dispersed repeats. A total of 50 interspersed repeats were identified in the *P. pedunculata* mitogenome, including 33 palindromic repeats (P, 66%) and 17 forward repeats (F, 34%); no reverse or complement repeats were detected. The lengths of the repeats were unevenly distributed. Most of the repeat sizes were between 70 bp–190 bp (78%), and 11 repeats (22%) exceeded 200bp, whereas only one repeat was longer than 1 kb, which was the largest repeat (1257 bp) (Figure 2; Supplementary Table S2).



With regards to the SSRs, the distribution and structure of 112 SSR repeats in the *P. pedunculata* mitogenome were analyzed, comprised 45 single-nucleotide motifs (40.18%), five dinucleotide repeats (4.46%), 12 trinucleotide repeats (10.71%), 40 tetranucleotide repeats (35.71%), four pentanucleotide repeats (3.57%), and six hexanucleotide repeats (5.36%) (Figure 3; Supplementary Table S3). Among all the SSRs, most repeats (82, 72.32%) were rich in A/T. Notably, 34 SSRs were entirely composed of A/T, including 26 monomer units (A/T), three dimer units (AT/

TA), two trimer units (AAT/TTA, TAT/ATA), and three tetramer units (AAAT/ATTT, AATT/TTAA). The A/T richness of the 48 SSRs ranged from 50 to 80%. The majority of SSRs (95) were located in IGS region, while nine SSRs were distributed on introns, two on exons, and one on the ORF region of *rps1*; only one SSR repeat was positioned across the intron region of *nad1* and part of the coding region of *matR* (Supplementary Table S4). These extensive SSRs provide abundant potential molecular markers for the identification and genetic study of *Prunus*.

TABLE 1 Genomic features of the *P. pedunculata* mitogenome.

Feature	A (%)	T (%)	G (%)	C (%)	GC (%)	Size (bp)	Proportion in Genome (%)
Genome	27.06	27.31	22.77	22.85	45.63	405855	100
Protein-coding genes	26.27	30.8	21.62	21.31	42.93	29973	7.39
Cis-spliced-intron	25.67	22.74	27.05	24.54	51.59	26711	6.58
tRNA	22.88	26.52	27.93	22.66	50.6	1840	0.45
rRNA	26.17	21.88	29.19	22.76	51.95	5136	1.27
Non-coding regions	27.34	27.34	22.75	22.57	45.32	344248	84.82

TABLE 2 Gene composition in the *P. pedunculata* mitogenome.

	Group of genes	Gene name
Core genes	ATP synthase	# <i>atp6</i> , <i>atp1</i> , <i>atp4</i> , <i>atp6</i> , <i>atp8</i> , <i>atp9</i>
	Cytochrome c biogenesis	<i>ccmB</i> , <i>ccmC</i> , <i>ccmFc*</i> , <i>ccmFn</i>
	Ubiquinol cytochrome c reductase	<i>cob</i>
	Cytochrome c oxidase	<i>cox1</i> , <i>cox2*</i> , <i>cox3</i>
	Maturases	<i>matR</i>
Variable genes	Transport membrane protein	<i>mttB</i>
	NADH dehydrogenase	<i>nad1****</i> , <i>nad2****</i> , <i>nad3</i> , <i>nad4***</i> , <i>nad4L</i> , <i>nad5****</i> , <i>nad6</i> , <i>nad7****</i> , <i>nad9</i>
	Ribosomal proteins (LSU)	# <i>rpl16</i> , <i>rpl10</i> , <i>rpl5</i>
	Ribosomal proteins (SSU)	# <i>rps14</i> , <i>rps1</i> , <i>rps12</i> , <i>rps13</i> , <i>rps3*</i> , <i>rps4</i> , <i>rps7</i>
	Succinate dehydrogenase	<i>sdh3</i> , <i>sdh4</i>
rRNA genes	Ribosomal RNAs	<i>rrn18</i> , <i>rrn26</i> , <i>rrn5</i>
tRNA genes	Transfer RNAs	<i>trnC</i> -GCA (2), <i>trnD</i> -GTC, <i>trnE</i> -TTC, <i>trnF</i> -GAA, <i>trnG</i> -GCC, <i>trnH</i> -GTG, <i>trnK</i> -TTT, <i>trnM</i> -CAT (3), <i>trnN</i> -ATT*, <i>trnN</i> -GTT, <i>trnP</i> -TGG (2), <i>trnQ</i> -TTG, <i>trnS</i> -CGA, <i>trnS</i> -GCT(3), <i>trnS</i> -TGA, <i>trnT</i> -TGT*, <i>trnW</i> -CCA, <i>trnY</i> -GTA, <i>trnY</i> -GTA*

Asterisks (*) beside genes represent intron numbers; Pound (#) before genes indicates Pseudogene; Numbers (2 or 3) after genes show the number of copies of multi-copy genes.

Additionally, 15 tandem repeat sequences ranging from 14 to 51 bp were evenly distributed in the mitogenome of *P. pedunculata* with a similarity match greater than 80%. These tandem repeats were predominantly located in the IGS, and only one repeat resided on *rrn26* (Supplementary Table S5). However, not all the repeat sequences were copied. Certain sequences had multiple non-integral copies. For instance, the TACATATTCGAGAA motif was repeated twice in the IGS between *atp9* and *nad1*, whereas the GACTATGAAACAGATCGC repeat unit was present in *rrn26* with a repetition number of 2.4.

3.3 Codon usage analysis of PCGs

The codon usage of 34 PCGs in the *P. pedunculata* mitogenome with a total length of 29,973 bp, encoding 9991 codons was analyzed. The results of RSCU analysis are shown in Figure 4. Leucine (Leu) was a predominant amino acid in PCGs with a frequency of 1066 (10.67%), followed by serine (8.98%) and isoleucine (7.97%), whereas cysteine (Cys) and tryptophan (Trp) were the least adopted amino acids, which only occurred 146 and

145 times (1.46% and 1.45%, respectively). The PCGs had the highest preference for UUU (Phe), which was used 364 times in PCGs, with an RSCU value of 1.12. The TAG termination codon was the least frequently used (Table 3; Supplementary Table S6). Alanine (Ala) had a preference for GCT, which occasionally occurred 248 times in PCGs, with a maximum RSCU value of 1.56.

Interestingly, in the PCGs, A/T bases preferentially appeared in the third codon position, rather than C/G. Codons ending in A/T had RSCU values greater than 1. In addition, almost all PCGs began with the typical start codon ATG, except for *nad1* which started with ACG. Five models of termination codons were observed in the PCGs. TNAs (TAA, TGA, and N for A, T, C, or G, respectively) were the most dominant codons in the 24 PCGs. Six PCGs (*atp4*, *mttB*, *matR*, *nad7*, *rps1*, and *ccmFn*) were terminated with TAG. Meanwhile, *ccmFc*, *atp9*, and *sdh4* were stopped by CGA and *atp6* ended with CAA, which may be incomplete codons (Supplementary Table S7).

3.4 The substitution rates of mitochondrial PCGs

Non-synonymous and synonymous substitution ratios (Ka/Ks) were calculated for the mitogenomes of 32 species based on 20 homologous PCGs. *P. pedunculata* was used as the reference. In most PCGs, the Ka/Ks values were notably less than 1 (Figure 5; Supplementary Table S8), implied that these genes were dominated by purifying selection during evolution. Conversely, the Ka/Ks ratios of *ccmC* in *P. anserina*, *ccmFn* in *R. chinensis* and *R. rugosa* versus *P. pedunculata* were greater than 1 (1.02956, 1.12327, and 1.05751, respectively), inferring positive selection. In the case of *nad2* from the *P. kanzakura* and *P. yedoensis* mitogenome, the Ka/Ks ratios were close to 1 (0.99014), suggesting a tendency of neutral selection. Additionally, 68 pairwise Ka/Ks values were 0 and 120 were pairwise with non-applicable (NA) Ka/Ks values. The highest average Ka/Ks ratios were observed for cytochrome c biogenesis genes (0.62132). Meanwhile, the lowest values of average Ka/Ks ratios were noted for *atp9* (0.0539), *nad3* (0.09952) and *nad9* (0.06257). In particular, the maximum substitution ratios of *atp9*, *atp1*, and *nad9* were as low as 0.235319, 0.325493, and 0.326979, respectively. The low Ka/Ks values suggested that these genes may have been highly conserved during the evolution of the *P. pedunculata* mitogenome.

3.5 RNA editing sites prediction

RNA editing events, especially the C-to-U editing sites, are enriched in plant mitogenomes. A total of 262 RNA editing sites in 26 analyzed PCGs of *P. pedunculata* mitogenome were identified among which 249 sites exhibiting C-to-U RNA editing (Figure 6; Supplementary Table S9). Most of the predicted RNA editing events occurred at the first (76, 29.01%) or second (173, 66.03%) positions of the codons, only 13 were found in the third positions (4.96%). The largest number of RNA editing sites was detected in the NADH dehydrogenase genes (129), among them *nad7* (34) maintained the most editing sites in all mitochondrial genes, then followed by *nad4* (28). There is only one editing site were predicted in *rps1* and *rps3*,

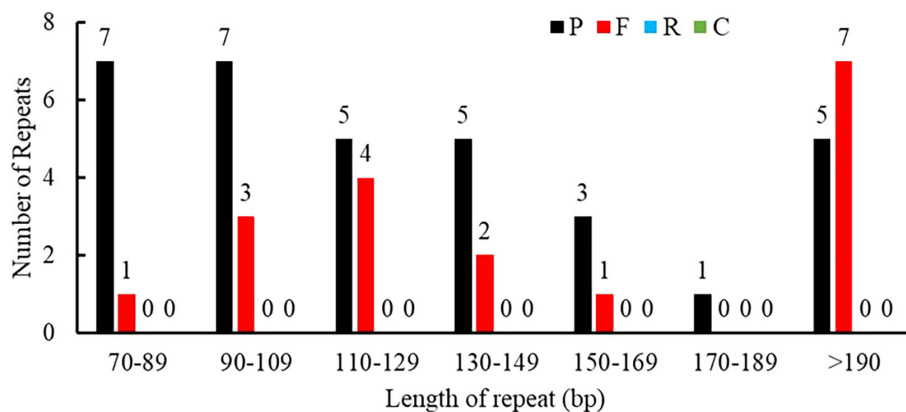


FIGURE 2

Allocation of the lengths of dispersed repeats in the *P. pedunculata* mitogenome. The X-axis indicates the types of dispersed repeats and the ordinate indicates the number of scattered repeats.

respectively. No potential RNA editing sites were identified in *atp6*, *ccmB*, *ccmC*, *ccmFc*, *mttB*, *nad4L*, *nad9*, *rps7* and *rps13*, no RNA editing sites were found in two assumed pseudogenes (*rpl16*, *rps14*), either. The majority of RNA editing sites (247) were non-synonymous variations, synonymous editing sites were merely (15) found. A total of 31 types of amino acid conversion were identified at these RNA editing sites (Supplementary Table S9), including two special sites in *atp9* and *nad7*, which convert to termination codons. The most frequently occurred amino acid changes among all of the identified mutations were histidine (H) to leucine (L) and serine (S) to leucine (L) change, with frequency of 61 times (23.28%) and 55 times (20.99%), respectively. Besides, *nad1* was initiated with ACG as its start codon (Supplementary Table S7), implying an alteration caused by RNA editing event.

3.6 Chloroplast-derived mitogenomic sequences

The mitogenome (405,855 bp) of *P. pedunculata* was approximately 2.57 times larger than the chloroplast genome (157,830 bp), so the distribution of mitochondrial genes in *P. pedunculata* was relatively sparse compared to that of chloroplast genes (Figure 7). In this study, 56 chloroplast-like fragments that might have undergone gene transfer were identified in the mitogenome, based on sequence similarity between the chloroplast and mitochondrial genomes of *P. pedunculata* (Figure 7; Supplementary Table S10). These inserted fragments, ranging from 30 to 863 bp, were distributed on the mitochondrial genome, with a total length of 11,582 bp, which comprised 2.85% of

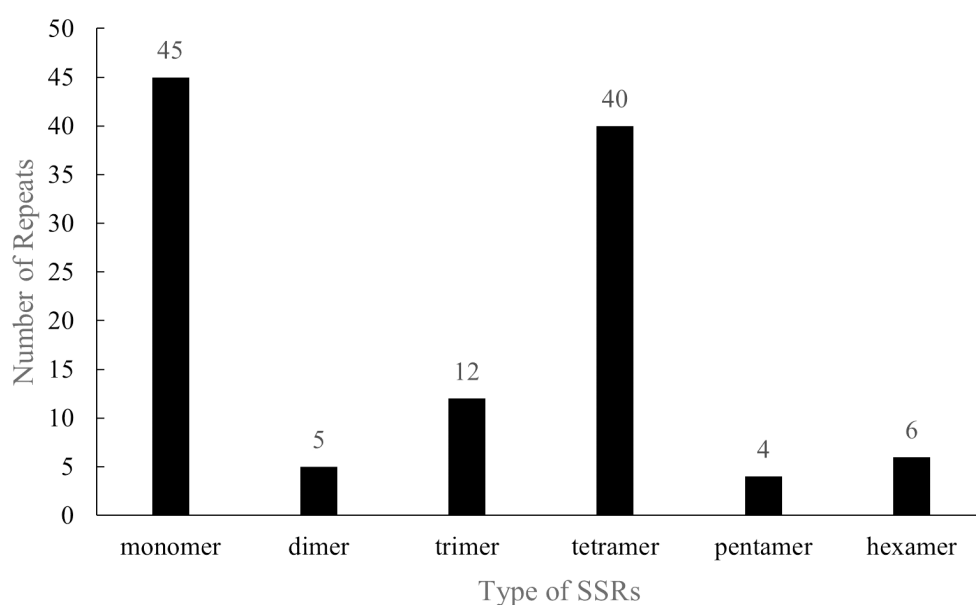
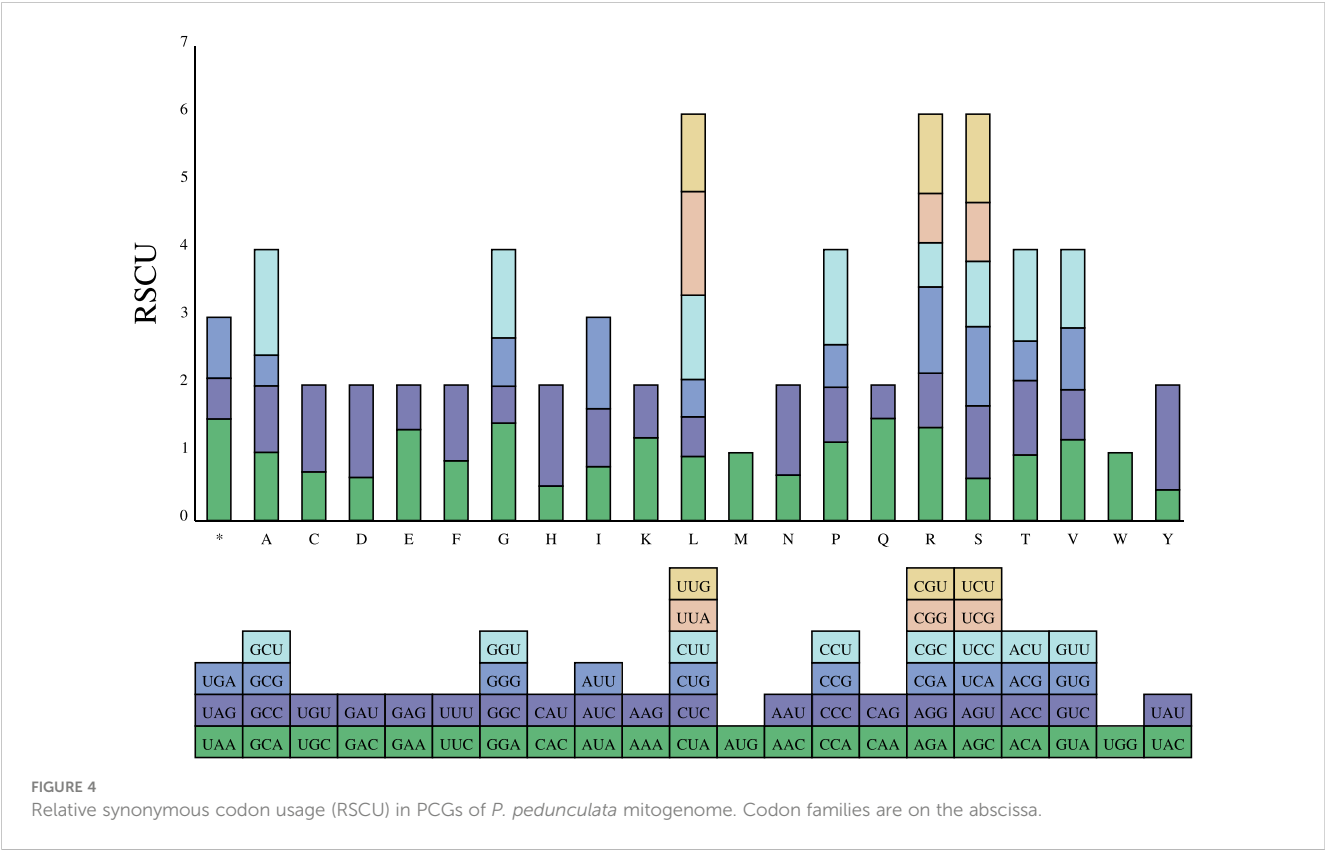


FIGURE 3

Distribution of SSRs in the *P. pedunculata* mitogenome. The X-axis indicates the type of SSRs, and the ordinate indicates the number of SSR repeats.



the complete mitogenome. The longest sequence (863 bp) was transferred from *rrn16S* and *rrn16S-2* in the cp genome to *rrn18* in the mitogenome. These migrated sequences were mainly located in the IGS (19) or rRNA genes (20) of the mitogenome. There are 22 *rrn16S* and *rrn23S* sequences of the *P. pedunculata* chloroplast genome that were inserted into the mitogenome and were mostly transferred into *rrn18* or *rrn26*, except for two fragments that were transferred to IGS regions. Among the remaining chloroplast-like

TABLE 3 Codon counts in the *P. pedunculata* mitochondrial PCGs.

Codon	Count	Codon	Count	Codon	Count	Codon	Count
UAA(*)	15	GGC(G)	93	AUG(M)	271	AGU(S)	160
UAG(*)	6	GGG(G)	122	AAC(N)	110	UCA(S)	175
UGA(*)	9	GGU(G)	223	AAU(N)	219	UCC(S)	144
GCA(A)	160	CAC(H)	64	CCA(P)	160	UCG(S)	130
GCC(A)	156	CAU(H)	188	CCC(P)	112	UCU(S)	195
GCG(A)	72	AUA(I)	211	CCG(P)	87	ACA(T)	123
GCU(A)	248	AUC(I)	227	CCU(P)	194	ACC(T)	140
UGC(C)	52	AUU(I)	358	CAA(Q)	217	ACG(T)	74
UGU(C)	93	AAA(K)	237	CAG(Q)	71	ACU(T)	172
GAC(D)	104	AAG(K)	151	AGA(R)	149	GUA(V)	191
GAU(D)	223	CUA(L)	168	AGG(R)	87	GUC(V)	118
GAA(E)	284	CUC(L)	104	CGA(R)	138	GUG(V)	146
GAG(E)	139	CUG(L)	98	CGC(R)	71	GUU(V)	185
UUC(F)	287	CUU(L)	221	CGG(R)	79	UGG(W)	146
UUU(F)	364	UUA(L)	272	CGU(R)	127	UAC(Y)	70
GGA(G)	246	UUG(L)	203	AGC(S)	93	UAU(Y)	239

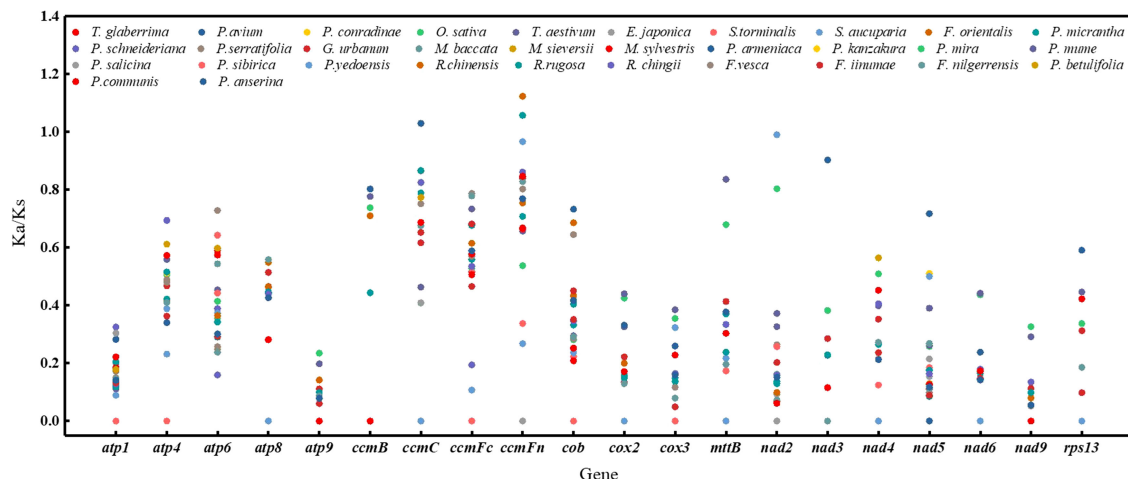


FIGURE 5

A dot plot of the Ka/Ks values of 20 protein-coding genes in mitogenomes of *P. pedunculata* versus 32 species.

sequences, nine fragments were located on tRNA genes, three on core genes (*atp1*, *ccmC*), and two on the ribosomal protein *rps12*. Some transferred fragments were intact genes (*atp1*, *rrn18*, *rrn26*, *rps12*, *ccmC*, *trnD-GTC*, *trnH-GTG*, *trnN-GTT*, *trnF-GAA*, *trnP-TGG*, *trnQ-TTG*), whereas others were partial sequences (*trnW-CCA*, *trnP-TGG-2*, *trnN-GTT*, *trnM-CAT-3*).

Migration may occur from gene to gene, from gene to IGS, from IGS to IGS, or from gene to introns/exons. Our findings showed that 22 rRNA sequences were transferred from chloroplasts to the mitochondria, mostly maintaining the function of ribosomal RNAs; whilst only two were inserted into the IGS region and became nonfunctional. Most sequences from the encoded genes of the chloroplasts lost their function and relocated to the IGS of the

mitochondria. In contrast, *atpA* from the chloroplasts was transformed into *atp1* in the mitochondria. Five segments that immigrated from the IGS remained in the IGS regions. The transferred sequences of 14 tRNA genes were shared between the chloroplast and mitochondrial genomes, accounting for 25% of the total transferred sequences. Relocated sequences were also observed in the intron and exon regions of some genes.

3.7 Phylogenetic analysis

Phylogenetic trees between *P. pedunculata* and 30 other Rosaceae species were constructed using the ML and BI methods,

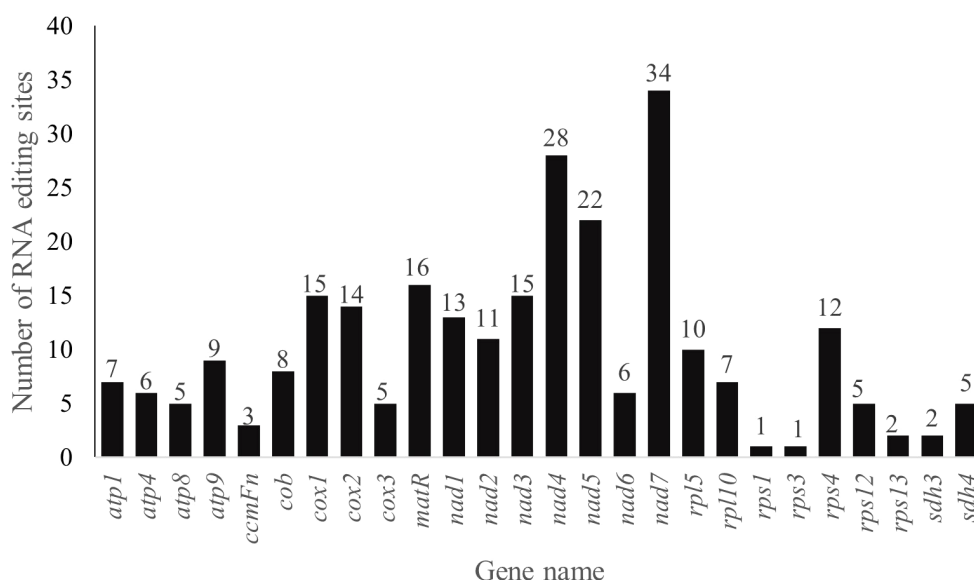


FIGURE 6

The distribution of RNA editing sites in the mt PCGs of *P. pedunculata*.

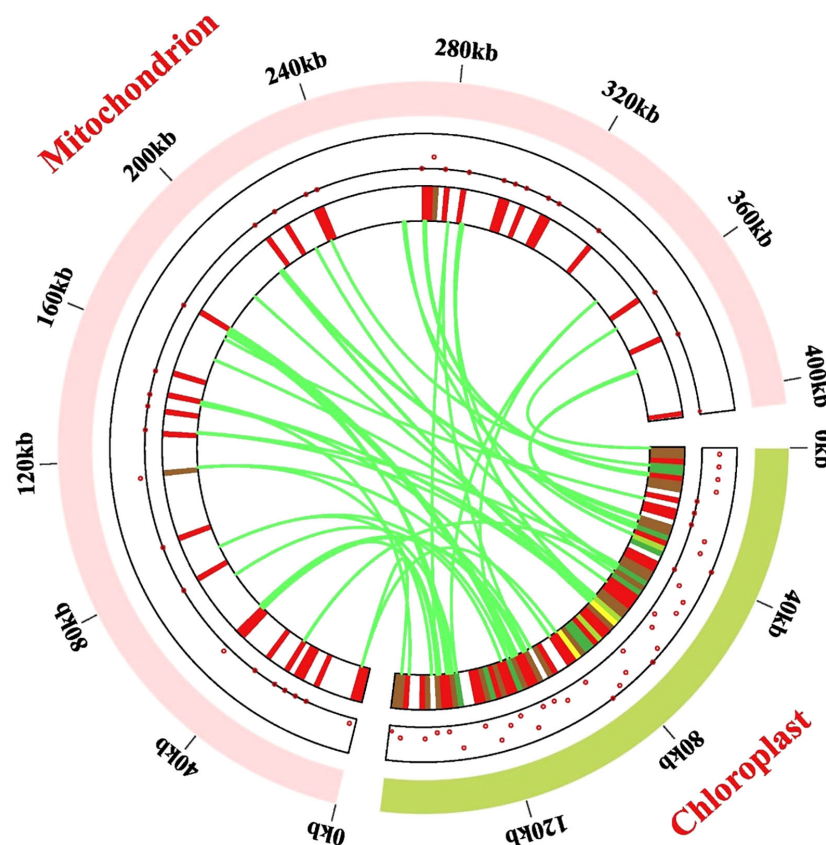


FIGURE 7

Gene transfer events between the chloroplast and mitochondrial genome. Dots and heat maps inside the two chromosomes demonstrate where the migrated genes are located. The light-green circular segment represents the chloroplast genome, and the pink circular segments depict the mitogenome. The lawn-green lines in the circle portray the routes of chloroplast-like sequences inserted from the cp genome into the mitogenome.

based on 20 shared mitochondrial PCGs and 72 shared chloroplast PCGs, respectively. *O. sativa* and *T. aestivum* functioned as outgroups. (Figure 8; Supplementary Table S1). Both ML and BI analyses indicated that most branches of the phylogenetic tree had high support values and the topology presented high consistency. The evolution relationships of both chloroplast and mitochondrial genomes among all taxa were separated into three clades, as deduced from the phylogenetic trees. The first large clade consisted of 11 *Prunus* species, which were further clustered into three secondary clades. *P. armeniaca*, *P. sibirica*, *P. mume*, and *P. salicina* form the *Prunus* subclade. As is shown in Figure 8, *P. pedunculata* was settled as a single monophyletic branch and group into the *Prunus* subgenus *Amygdalus* clade together with *P. mira* in both organelle genome trees. The subgenus *Cerasus*, including *P. avium*, *P. conradinae*, *P. schneideriana*, *P. yedoensis*, and *P. kanzakura* constitutes the third subclade of the *Prunus* genus. These results are consistent with the classification taxonomy.

Ten other ligneous species of Rosaceae, comprising species from the genera *Eriobotrya*, *Sorbus*, *Torminalis*, *Pyrus*, *Photinia*, and *Malus*, formed Clade 2, among which *E. japonica* and *S. aucuparia* formed distinct subclades, the rest of the species in the group clustered into another subclade in the mitogenome tree, whereas *S.*

aucuparia was more related to *Pyrus* in the cp genome tree. Species from the genera *Fragaria*, *Rosa* and *Potentilla* comprised Clade 3. Most species in Clade 3 were herbaceous plants, in addition to three shrub species: *R. chingii*, *R. rugosa*, and *R. chinensis*. *R. chingii* and *G. urbanum* accessions diverged into two independent subclades with different genetic distance in the mitogenome tree and cp genome tree, though they are separated from the other relatives in this group.

4 Discussion

4.1 Genomic features of the *P. pedunculata* mitogenome

Mitochondria have more complex genomes in plants than those of animals, owing to variation and repeated sequences. Due to the complexity of plant mitogenomes, extensive research has been focused on plastids, leaving multifarious mitogenomes to be investigated (Zardoya, 2020). Each of the 60 Rosaceae mitogenomes deposited in NCBI database was assembled into cyclic structure with remarkable variation in size, ranging from

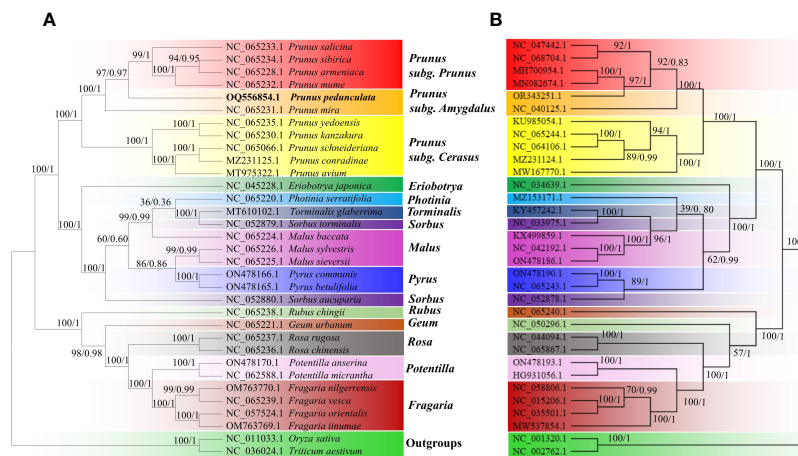


FIGURE 8

The phylogenetic relationships of *P. pedunculata* compared with that of 30 Rosaceae species. (A) Phylogenetic analysis of mitogenomes based on 20 shared protein-coding genes from the mitogenome; (B) phylogenetic analysis of cp genomes based on 72 shared protein-coding genes. *Oryza sativa* and *Triticum aestivum* were selected as outgroups. The ML bootstrap support values/BI posterior probabilities are shown for each node. Colored blocks indicate the subtype that the specific species belongs to.

270, 143bp (*Rosa* hybrid cultivar: OQ628291.1) to 535,727bp (*P. mume*: NC_065232.1). The genome size within the *Prunus* species varies slightly (Supplementary Table S1), among which *P. pedunculata* mitogenome (405, 855bp) was the smallest. In this survey, the complete mitogenome of *P. pedunculata* was thoroughly characterized and assembled into a typical circular structure. The GC contents of Rosaceae mitogenomes are relatively static, ranging from 43.31% in *R. chingii* to 45.62% in *P. avium* (Sun et al., 2022). Interestingly, *P. pedunculata* exhibited the highest percentage of GC base (45.63%). The corresponding chloroplast genome sequence of *P. pedunculata* was found to be 157,830 bp long, with a GC content of 36.8%.

Plant mitogenomes contain highly conserved genes, relatively low gene densities, abundant non-coding sequences, and RNA editing occurrences (Niu et al., 2023). The high account of non-coding regions (84.82%) of the *P. pedunculata* mitogenome might be a consequence of sequence duplication during evolution (Ma et al., 2022). Mitochondrial DNA (mtDNA) encodes tRNAs, rRNAs, and a dynamic number of ribosomal proteins (Kubo and Newton, 2008). A total of 34 PCGs were identified in the *P. pedunculata* mitogenome, which is higher than that of most mitogenomes of Rosaceae species, aside from *P. salicina*, which has 39 PCGs in its mitogenome. The high content of PCGs indicated that fewer original mitochondrial genes had transfer into nuclear region during the evolutionary history of *P. pedunculata*. No gene loss events were observed in PCGs of the *P. pedunculata* mitogenome. Thirty-five genes were detected in most *Prunus* mitogenomes, except for *rpl16* which was found only in the *P. avium* cultivar Glory, the *P. avium* cultivar Staccato, *P. tenella*, *P. armeniaca*, and *P. salicina* (Sun et al., 2022). Additionally, similar with *O. sativa* (Kazama and Toriyama), the *P. pedunculata* mitogenome contains two duplicated loci of the *atp6* gene, one of them was assumed to be a pseudo-copy. Most Rosaceae species contain three ribosomal RNAs in their mitogenomes, except some

species that possess four rRNAs (*G. urbanum*, *M. domestica* cultivar Gala, *M. domestica* cultivar Yantai fuji 8, *M. sylvestris* and *P. serratifolia*), and *S. aucuparia* has two rRNAs (Sun et al., 2022). In accordance with the known mitogenomes of *Prunus*, the mitogenome of *P. pedunculata* contains three rRNAs (*rrn18*, *rrn26*, and *rrn5*). In general, transfer RNAs in the mitogenome of angiosperm do not form a complete set (Michaud et al., 2011). Eight tRNAs (*trnP-AGG*, *trnP-GGG*, *trnR-TCT*, *trnS-ACT*, *trnS-GGA*, *trnT-GGT*, *trnV-GAC*, and *trnI-GAT*) were absent in the mitogenomes of *Prunus*, including *P. pedunculata*. Specifically, there are two copies of *trnC-GCA* in *Prunus* species and *trnN-ATT* appears solely in *Prunus* mitogenomes. Three copies of *trnM-CAT*, two copies of *trnP-TGG*, and *trnY-GTA* have been identified in most *Prunus* species. Moreover, *trnI-TAT* and *trnL-TAA* were absent from *P. pedunculata*, *P. schneideriana*, *P. conradinae*, and *P. tenella*, whereas *trnK-TTT* and *trnG-GCC* were present in these four species, differs greatly from other *Prunus* species (Sun et al., 2022). As for the intron-containing tRNAs (*trnN-ATT*, *trnT-TGT* and *trnY-GTA*) in *P. pedunculata* mitogenome, splicing their introns is a critical step of tRNA maturation (Hayne et al., 2022).

Repetitive sequences, especially abundant SSRs (Freitas et al., 2022), increase the mitogenome size, generation of multiple copies of genes, genome diversification, and structural variations (Fang et al., 2021). SSRs are characterized by cross-species transfer and high polymorphism, which allow them to be applied as molecular markers in phylogenetic analyses (Rossetto et al., 2002). In our study, 112 SSRs of *P. pedunculata* mitogenome distributed across different genomic regions: ORFs, introns, exons, or intergenic regions (IGS). The most abundant type of SSR in the mtDNA was mononucleotides (40.18%), consistent with a previous study (Kuntal and Sharma, 2011). However, the second-most common SSR in *P. pedunculata* mitogenome was tetranucleotide repeats (35.71%), rather than dinucleotide repeats. A near-universal A/T bias (Smith and Keeling, 2015) was also observed in *P. pedunculata*

mitogenome, these SSRs were composed of motifs rich in A and T, which agrees with previous observations (Kuntal and Sharma, 2011), and corroborated the correlation between AT content of the complete mitogenome and SSRs (Freitas et al., 2022).

4.2 The genomic variation of *P. pedunculata* mitogenome

The non-synonymous and synonymous substitution ratios (Ka/Ks) have a complex relationship with biological functions of PCGs, and reflect gene selective pressures, thus contributing to our understanding of the evolutionary dynamics of PCGs among related species (Wang et al., 2010). In the current study, for the cytochrome c biogenesis genes *ccmC* and *ccmFn* that underwent positive selection (Ka/Ks>1), non-synonymous variations were favored because of functional adaptation or useful mutations. The average Ka/Ks value was the largest (0.62132) for cytochrome c biogenesis genes. Stabilizing selection (Ka/Ks<1) have been reported for other plants mitochondrial genes (Bi et al., 2020; Cheng et al., 2021; Yu et al., 2022). In contrast, genes that with Ka/Ks ratios below 1 may be highly conserved and have undergone negative selection during evolution (Greilhuber et al., 2012). Nonsynonymous variations are likely to be disadvantageous, therefore, genes with crucial functions subjected to stabilizing selection are partial to having lower Ka/Ks values. The lowest average values were observed for genes related to NADH dehydrogenase (0.18653). Because of constraint on protein function, synonymous substitutions are more common in most mitochondrial PCGs of *P. pedunculata* that dominated by purifying selection, which concordant with the fact that PCGs in the mitogenome are conserved across green plants (Mower et al., 2007; Cheng et al., 2021). With the Ka/Ks value approximately equal to 1, the protein function of *nad2* might not constrain evolutionary changes. These results implied that the majority of genes were subjected to purifying selection. Thus, further research of the gene selection and evolution of *Prunus* species is still required.

Selection events for biased codon usage and recognition motifs for RNA editing sites occur in angiosperm mitogenomes (Liu and B., 2005). Deaminated cytosines in RNA transcripts become uridines at RNA editing sites in numerous mitochondrial DNA-encoded genes (Morley and Nielsen, 2017; Hao et al., 2021). According to the RSCU analysis, the most unique PCGs encoded by the *P. pedunculata* mitogenome contained the classical start codon ATG, in accordance with the allocation of amino acid compositions in other plant mitogenomes (Sloan et al., 2018b). Nevertheless, *nad1* starts with ACG, presumably as a consequence of C-to-U editing at the second site. The A/T bias in the third codon position of PCGs with higher RSCU values (>1) indicates that A/T (U) stews across the mitochondrial genome of *P. pedunculata*, which may be a consequence of the A/T mutation bias common in plant mitogenomes (Smith and Keeling, 2015; Bi et al., 2020). CGA acts as a stop codon for *ccmFc*, *atp9*, and *sdh4*; however, analogous observations have rarely been reported. The *cox1* gene of many lepidopteran species uses CGA as a start codon (Wu et al.,

2020a; Jiang et al., 2021) rather than a stop codon. CAA acts as a stop codon for *atp6* in the *P. pedunculata* mitogenome. Similar results have been reported for *atp9* in wild carrots (Mandel et al., 2012).

It is believed that RNA editing events plays a pivotal role in the plant development regulation and stress resistance (Tang and Luo, 2018). In general, plant mitochondria contain more RNA editing sites than chloroplast (Small et al., 2020). We have predicted 262 RNA editing sites in *P. pedunculata* mitogenome, which is lower in comparison to other angiosperms (Giegé and Brennicke, 1999; Sloan et al., 2010; Richardson et al., 2013; Shan et al., 2023). Most RNA editing events in plant organelles arise from the site-specific C-to-U conversion (Small et al., 2020). Similar results were found in the current research. Particularly, 3 reverse changes (U-to-C conversion) were also been identified, which is more likely enriched in basal plants (Chen et al., 2023). In addition, G-to-A, G-to-U, A-to-G, A-to-U, U-to-A and C-to-A conversion types were also observed once and G-to-C changes were found twice among all RNA editing sites. RNA editing events may trigger the variation of amino acids and change the start or stop codons of PCGs. There are editing sites being transformed to stop codons in *atp9* and *nad7*. ACG was used as a start codon in *nad1* gene of *P. pedunculata* mitochondria directly, possibly due to a C-to-U conversion at the second site without editing in mitogenome (Zanduetta-Criado and Bock, 2004). RNA editing events were reported existing in the initiation codon of *nad1* in other plants (Li et al., 2018; Fang et al., 2021; Wee et al., 2022). Consistent to most plants, there is few RNA editing events occur in rRNA, tRNA and introns (Giegé and Brennicke, 1999). In view of previous studies, RNA editing sites barely exist at the third codon position (Bi et al., 2020; Ma et al., 2022), largely because of the limitations of the methodology. Therefore, further research on the precisely predictive methods is still required.

Intergenicomic gene transfer between organellar genomes has been an important phenomenon throughout the long-term evolution of higher plants (Sloan and Wu, 2014). As an important type of intracellular gene transfer (IGT) (Wang et al., 2018d) in the mitochondria, the mechanism of transfer events remains challenging because most transferred DNA sequences are located in non-coding regions. Plant mitogenomes maintain massive mitochondrial plastid fragments (MTPTs) (Lai et al., 2022). The proportion of plastid-derived fragments in mitogenomes varies from 0.44% (*P. serratifolia*) to 16.34% (*R. chingii*) in the family Rosaceae (Sun et al., 2022). The integration rate of chloroplast-derived mitogenomic sequences in *P. pedunculata* was 2.85%, which was much higher than that in other *Prunus* species (less than 1.16%). Similarly, the total length of the transferred sequences in the *P. pedunculata* mitogenome was 11,582 bp, which was the longest among the *Prunus* species. The species with the highest total length of plastid-derived sequences in the Rosaceae family is *R. chingii* (77,163 bp), which also comprises the highest proportion (16.34%) (Sun et al., 2022). Mostly, chloroplast-derived sequences are non-functional (Zhang et al., 2023). Remarkably, fragments from the exons of two genes (*rps12*, *rps12-2*) in the chloroplast genome of *P. pedunculata* changed to functional *rps12* after transfer. Conversely, as a

consequence of gene transfer events, the original functional gene sequence of *ndhF* was relocated to the exon region of *nad5*. The *rrn16-rrn23* regions were used to construct transformation vectors, indicating that the *rrn16-rrn23* chloroplast region might affect transformation efficiency (Lopez-Ochoa et al., 2015). The lost tRNAs in mitogenomes can be replaced by tRNAs inserted into other organelles (Sloan et al., 2010). The transfer of tRNA gene sequences from chloroplasts to mitochondrial genomes are common in plants (Bergthorsson et al., 2003). Inserted tRNAs accounted for 25% of the total transferred sequences in the *P. pedunculata* mitogenome, which contradicts the hypothesis that most transferred genes are tRNA genes (Chang et al., 2013).

Polyploidy of *P. pedunculata* leads to difficulties in the evolutionary analysis of genome sequences and phenotypes (Wang et al., 2020a). Phylogenetic inference of *P. pedunculata* has been controversial. According to previous publications, *P. pedunculata* was placed outside the monophyletic subgenus *Amygdalus* based on molecular data. In terms of morphological classification, it is clustered as a sister clade to peach (Yazbek and Oh, 2013). Previous studies have suggested that *P. pedunculata* is closely associated with *P. tomentosa* and *P. triloba* (Wang et al., 2018a; Duan et al., 2020; Wang et al., 2020a; Wang et al., 2021a). Nevertheless, there were no complete mitochondrial genome data for these two species available for phylogenetic analysis. In this study, within the *Prunus* clade, *P. pedunculata* was divided into an independent branch, suggesting that the mitogenome divergence process produced distinctive maternal lines of *P. pedunculata*. The analysis provided strong evidence of the phylogenetic affinities of *P. pedunculata* from the perspective of 20 conserved PCGs in the mitogenome and provided a better understanding of the phylogenetic relationships among Rosaceae species. Nevertheless, additional investigations of more accessions of *Prunus* species and more molecular data are required to comprehensively understand the phylogeny of *Prunus* (Pervaiz et al., 2015).

5 Conclusions

In this study, the complete mitochondrial genome of *P. pedunculata* was assembled and characterized. Extensive analyses have been conducted on the mitogenome features. The *P. pedunculata* mitogenome is a circular molecule with a total length of 405,855 bp, which encodes 62 genes, including 34 PCGs, three ribosomal RNAs, and 19 transfer RNAs. The mitogenome contained 112 SSRs, 15 tandem repeats, and 50 interspersed repetitive sequences with a total repeat length of 11,793 bp. The codon usage of PCGs, Ka/Ks ratios, RNA editing and gene transfer events in the *P. pedunculata* mitogenome were thoroughly evaluated. This phylogenetic relationship analysis was supported by mitogenome information of 30 other taxa of the Rosaceae family. In summary, understanding the mitochondrial genome characteristics of *P. pedunculata* is of great importance to promote the understanding of the evolution of the genetic background and provide a basis for genetic breeding of *Prunus*.

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. The raw data were deposited at NCBI SRA database <https://www.ncbi.nlm.nih.gov/sra>; with the accession number: SRR25636420.

Author contributions

ZW: Formal analysis, Methodology, Writing – review & editing. QL: Formal analysis, Investigation, Validation, Writing – original draft. CT: Resources, Writing – review & editing. YY: Investigation, Writing – review & editing. LL: Validation, Writing – review & editing. YF: Formal analysis, Writing – review & editing. ZL: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research was funded by the Central Public-interest Scientific Institution Basal Research Fund (No. Y2023PT02), Inner Mongolia Science and Technology Plan (2021CG0019, 2020GG0127 and 2022YFHH0140), Hohhot Science and Technology Plan (2022-she-zhong-1-2), High-level talents for scientific research funds for Inner Mongolia Autonomous Region (2202000010131440006).

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.2023.1266797/full#supplementary-material>

References

- Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdociimi, F., Nabholz, B., and Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* 20, 892–905. doi: 10.1111/1755-0998.13160
- Bao, W., Ao, D., Wang, L., Ling, Z., Chen, M., Bai, Y., et al. (2021). Dynamic transcriptome analysis identifies genes related to fatty acid biosynthesis in the seeds of *Prunus pedunculata* Pall. *BMC Plant Biol.* 21, 152–168. doi: 10.1186/s12870-021-02921-x
- Beier, S., Thiel, T., Munch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198
- Bellet, S., Cusimano, N., Luo, S., Sun, G., Zarre, S., Gröger, A., et al. (2016). Assembled plastid and mitochondrial genomes, as well as nuclear genes, place the parasite family cynomoriaceae in the Saxifragales. *Genome Biol. Evol.* 8, 2214–2230. doi: 10.1093/gbe/evw147
- Benson, G. (1999). Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27 (2), 573–580. doi: 10.1093/nar/27.2.573
- Bergthorsson, U., Adams, K. L., Thomason, B., and Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* 424, 197–201. doi: 10.1038/nature01743
- Bi, C., Lu, N., Xu, Y., He, C., and Lu, Z. (2020). Characterization and analysis of the mitochondrial genome of common bean (*Phaseolus vulgaris*) by comparative genomic approaches. *Int. J. Mol. Sci.* 21, 3378. doi: 10.3390/ijms21113778
- Chang, S., Wang, Y., Lu, J., Gai, J., Li, J., Chu, P., et al. (2013). The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. *PLoS One* 8, e56502. doi: 10.1371/journal.pone.0056502
- Chau, C. F., and Wu, S. H. (2006). The development of regulations of Chinese herbal medicines for both medicinal and food uses. *Trends Food Sci. Technol.* 17, 313–323. doi: 10.1016/j.tifs.2005.12.005
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009
- Chen, H., Huang, L., Yu, J., Miao, Y., and Liu, C. (2023). Mitochondrial genome of *Artemisia argyi* L. suggested conserved mitochondrial protein-coding genes among genera *Artemisia*, *Tanacetum* and *Chrysanthemum*. *Gene* 871, 147427. doi: 10.1016/j.gene.2023.147427
- Cheng, Y., He, X., Priyadarshani, S. V. G. N., Wang, Y., Ye, L., Shi, C., et al. (2021). Assembly and comparative analysis of the complete mitochondrial genome of *Suaeda glauca*. *BMC Genomics* 22 (1), 167. doi: 10.1186/s12864-021-07490-9
- Chu, J., Li, Y., Zhang, L., Li, B., Gao, M., Tang, X., et al. (2017). Potential distribution range and conservation strategies for the endangered species *Amygdalus pedunculata*. *Biodiversity Sci.* 25, 799–806. doi: 10.17520/biods.2015218
- Chu, J., Xu, X., and Zhang, Y. (2013). Production and properties of biodiesel produced from *Amygdalus pedunculata* Pall. *Bioresour. Technol.* 134, 374–376. doi: 10.1016/j.biortech.2012.12.089
- Chu, J., Yang, H., Lu, Q., and Zhang, X. (2015). Endemic shrubs in temperate arid and semiarid regions of northern China and their potentials for rangeland restoration. *Acta Bot. Sin.* 41, 61–63. doi: 10.1093/aobpla/plv063
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, 1–4. doi: 10.1093/gigascience/giab008
- Dong, S. P. (2015). *Phylogeny Research of Prunus Triloba and Related Species Based on Chromosome Kayotype and Single-copy Nuclear Gene DNA Sequences*. Master's thesis. (Beijing: Beijing Forestry University).
- Du, Z., Lu, K., Zhang, K., He, Y., Wang, H., Chai, G., et al. (2021). The chloroplast genome of *Amygdalus* L. (Rosaceae) reveals the phylogenetic relationship and divergence time. *BMC Genomics* 22, 645. doi: 10.1186/s12864-021-07968-6
- Duan, C., Zhang, K., and Duan, Y. (2020). Comparison of complete chloroplast genome sequences of *Amygdalus pedunculata* Pall. *Sheng Wu Gong Cheng Xue Bao* 36, 2850–2859. doi: 10.13345/j.cjb.200188
- Fang, B., Li, J., Zhao, Q., Liang, Y., and Yu, J. (2021). Assembly of the complete mitochondrial genome of Chinese plum (*Prunus salicina*): characterization of genome recombination and RNA editing sites. *Genes (Basel)* 12 (12), 1970. doi: 10.3390/genes12121970
- Freitas, K., Busanello, C., Viana, V. E., Pegoraro, C., Victoria, F., and Maia, L. (2022). An empirical analysis of mtSSRs: could microsatellite distribution patterns explain the evolution of mitogenomes in plants? *Funct. Integr. Genomics* 22, 35–53. doi: 10.1007/s10142-021-00815-7
- Gao, Y., Li, C., Chen, B., Shen, Y. H., Han, J., and Zhao, M. G. (2016). Anti-hyperlipidemia and antioxidant activities of *Amygdalus pedunculata* seed oil. *Food Funct.* 7, 5018–5024. doi: 10.1039/c6fo01283c
- Giegé, P., and Brennicke, A. (1999). RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *PNAS* 96, 15324–15329. doi: 10.1073/pnas.96.26.15324
- Greilhuber, J., Doležal, J., and Leitch, I. J. (2012). *Plant genomes, their residents, and their evolutionary dynamics*. In *Plant genome diversity*. Ed. J. F. Wendel (Springer-Verlag Wien). doi: 10.1007/978-3-7091-1130-7
- Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238
- Gualberto, J. M., Milesina, D., Wallet, C., Niaz, A. K., Weber-Lotfi, F., and Dietrich, A. (2014). The plant mitochondrial genome: dynamics and maintenance. *Biochimie* 100, 107–120. doi: 10.1016/j.biochi.2013.09.016
- Guo, G. G. (2014). *Studies on Drought Resistance of Different Regional Amygdalus pedunculata* Pall. Master's thesis. (Yangling: Northwest Agriculture & Forestry University).
- Hao, W., Liu, G., Wang, W., Shen, W., Zhao, Y., Sun, J., et al. (2021). RNA editing and its roles in plant organelles. *Front. Genet.* 12. doi: 10.3389/fgene.2021.757109
- Hayne, C. K., Lewis, T. A., and Stanley, R. E. (2022). Recent insights into the structure, function, and regulation of the eukaryotic transfer RNA splicing endonuclease complex. *WIREs RNA* 13, e1717. doi: 10.1002/wrna.1717
- He, Y., Xiang, K., Chen, C., Wang, J., and Wu, Z. (2023). Master graph: an essential integrated assembly model for the plant mitogenome based on a graph-based framework. *Briefings Bioinf.* 24 (1), 1–13. doi: 10.1093/bib/bbac522
- He, Y., Pan, L., Yang, T., Wang, W., Li, C., Chen, B., et al. (2021). Metabolomic and confocal laser scanning microscopy (CLSM) analyses reveal the important function of flavonoids in *Amygdalus pedunculata* pall leaves with temporal changes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.648277
- Jiang, B. (2008). *Study on Cold Resistance of Amygdalus pedunculata* Pall. Master's Thesis. (Yangling: Northwest A&F University).
- Jiang, K., Su, T., He, B., Zhao, F., Lin, G., and Huang, Z. (2021). Complete mitochondrial genome of *Casmara patrona* (Lepidoptera: Oecophoridae). *Mitochondrial DNA B Resour* 6, 325–326. doi: 10.1080/23802359.2020.1863872
- Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kazama, T., and Toriyama, K. (2016). Whole mitochondrial genome sequencing and re-examination of a cytoplasmic male sterility-associated gene in boro-taichung-type cytoplasmic male sterile rice. *PLoS One* 11 (7), e0159379. doi: 10.1371/journal.pone.0159379
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199
- Knoop, V., Volkmar, U., Hecht, J., and Grewe, F. (2011). “Mitochondrial genome evolution in the plant lineage,” in *Plant Mitochondria*, 3–29. Ed. F. Kempken (Advances in Plant Biology, vol. 1. New York, NY: Springer). doi: 10.1007/978-0-387-89781-3_1
- Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michels, R. W., et al. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet.* 15, e1008373. doi: 10.1371/journal.pgen.1008373
- Kubo, T., and Newton, K. J. (2008). Angiosperm mitochondrial genomes and mutations. *Mitochondrion* 8, 5–14. doi: 10.1016/j.mito.2007.10.006
- Kuntal, H., and Sharma, V. (2011). In silico analysis of SSRs in mitochondrial genomes of plants. *OMICS* 15, 783–789. doi: 10.1089/omi.2011.0074
- Kurtz, S., Choudhuri, J., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29 (22), 4633–4642. doi: 10.1093/nar/29.22.4633
- Lai, C., Wang, J., Kan, S., Zhang, S., Li, P., Reeve, W. G., et al. (2022). Comparative analysis of mitochondrial genomes of *Broussonetia* spp. (Moraceae) reveals heterogeneity in structure, syntenic, intercellular gene transfer, and RNA editing. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1052151
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Li, B. (2017). *Soft-wood Cutting Propagation Technique and Rooting Physiology Research of Amygdalus pedunculata*. Master's Thesis. (Beijing: Chinese academy of forestry sciences).
- Li, J., Bi, C., Tu, J., and Lu, Z. (2018). The complete mitochondrial genome sequence of *Boechera stricta*. *Mitochondrial DNA Part B* 3, 896–897. doi: 10.1080/23802359.2018.1501323
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, C., Yang, J., Yao, L., Qin, F., Hou, G., Chen, B., et al. (2020). Characterisation, physicochemical and functional properties of protein isolates from *Amygdalus pedunculata* Pall seeds. *Food Chem.* 311, 125888. doi: 10.1016/j.foodchem.2019.125888
- Liu, Z. R. (2017). *Study on Resources and Priority Conservation of Rare and Endangered Plants in Inner Mongolia*. Doctoral Dissertation. (Hohhot: Inner Mongolia Agricultural University).
- Liu, T. L., and Bundschuh, R. (2005). Model for codon position bias in RNA editing. *Phys. Rev. Lett.* 95 (8), 88101. doi: 10.1103/PhysRevLett.95.088101
- Lopez-Ochoa, L. A., Apolinar-Hernández, M. M., and Peña-Ramírez, Y. J. (2015). Characterization of chloroplast region rrn16-rrn23S from the tropical timber tree *Cedrela odorata* L. and *de novo* construction of a transplasmic expression vector suitable for Meliaceae trees and other economically important crops. *Genet. Mol. Res.* 14, 1469–1478. doi: 10.4238/2015.February.20.2
- Lu, C., Li, H., Li, C., Chen, B., and Shen, Y. (2018). Chemical composition and radical scavenging activity of *Amygdalus pedunculata* Pall leaves' essential oil. *Food Chem. Toxicol.* 119, 368–374. doi: 10.1016/j.fct.2018.02.012
- Luo, S. W. (2009). *Study on Photosynthetic Characteristics of Amygdalus pedunculata Pall*. Master's Thesis. (Yangling: Northwest A&F University).
- Ma, X. W. (2006). *Research of drought resistance mechanism of Amygdalus pedunculatus Pall*. Master's Thesis. (Yangling: Northwest A&F University).
- Ma, H. (2013). *Research on Nutrition Ingredient of Amygdalus Pedunculata Pall. and Related Species of Six Kind*. Master's Thesis (Northwest A&F University).
- Ma, Q., Wang, Y., Li, S., Wen, J., Zhu, L., Yan, K., et al. (2022). Assembly and comparative analysis of the first complete mitochondrial genome of *Acer truncatum* Bunge: a woody oil-tree species producing nervonic acid. *BMC Plant Biol.* 22, 29. doi: 10.1186/s12870-021-03416-5
- Mandel, J. R., McAssey, E. V., Roland, K. M., and McCauley, D. E. (2012). Mitochondrial gene diversity associated with the atp9 stop codon in natural populations of wild carrot (*Daucus carota* ssp. *carota*). *J. Hered.* 103, 418–425. doi: 10.1093/jhered/esr142
- McBride, H. M., Neuspiel, M., and Wasiak, S. (2006). Mitochondria: more than just a powerhouse. *Curr. Biol.* 16, R551–R560. doi: 10.1016/j.cub.2006.06.054
- Michaud, M., Cognat, V., Duchene, A. M., and Marechal-Drouard, L. (2011). A global picture of tRNA genes in plant genomes. *Plant J.* 66, 80–93. doi: 10.1111/j.1365-3113X.2011.04490.x
- The China Red List of Biodiversity - Higher Plants (2020). *Kunming: Ministry of Ecology and Environment of the People's Republic of China and Chinese Academy of Sciences*, 2023.
- Morley, S. A., and Nielsen, B. L. (2017). Plant mitochondrial DNA. *Front. Biosci. (Landmark Ed)* 22 (6), 1023–1032. doi: 10.2741/4531
- Mower, J. P., Sloan, D. B., and Alverson, A. J. (2012). Plant mitochondrial genome diversity: the genomics revolution. *Plant Genome Diversity* 1, 123–144. doi: 10.1007/978-3-7091-1130-7_9
- Mower, J. P., Touzet, P., Gummow, J. S., Delph, L. F., and Palmer, J. D. (2007). Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC Evolutionary Biol.* 7, 135. doi: 10.1186/1471-2148-7-135
- Niu, Y., Zhang, T., Chen, M., Chen, G., Liu, Z., Yu, R., et al. (2023). Analysis of the complete mitochondrial genome of the bitter melon (*Momordica charantia*). *Plants (Basel)* 12 (8), 1686. doi: 10.3390/plants12081686
- Pervaiz, T., Sun, X., Zhang, Y., Tao, R., Zhang, J., and Fang, J. (2015). Association between Chloroplast and Mitochondrial DNA sequences in Chinese *Prunus* genotypes (*Prunus persica*, *Prunus domestica*, and *Prunus avium*). *BMC Plant Biol.* 15, 4. doi: 10.1186/s12870-014-0402-4
- Putintseva, Y. A., Bondar, E. I., Simonov, E. P., Sharov, V. V., Oreshkova, N. V., Kuzmin, D. A., et al. (2020). Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. *BMC Genomics* 21 (1), 654. doi: 10.1186/s12864-020-07061-4
- Qu, X. J., Moore, M. J., Li, D. Z., and Yi, T. S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15, 50. doi: 10.1186/s13007-019-0435-7
- Rhodes, L., Pollard, R. P., and Maxted, N. (2016). "Amygdalus pedunculata," in *The IUCN Red List of Threatened Species* IUCN Standards and Petitions Subcommittee, vol. 2016, e.T50025884A50025905.
- Rice, D. W., Alverson, A. J., Richardson, A. O., Young, G. J., Sanchez-Puerta, M. V., Munzinger, J., et al. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* 342, 1468–1473. doi: 10.1126/science.1246275
- Richardson, A. O., Rice, D. W., Young, G. J., Young, G. J., Alverson, A. J., Alverson, A. J., Palmer, J. D., and Palmer, J. D. (2013). The "fossilized" mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol.* 11, 29. doi: 10.1186/1741-7007-11-29
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029
- Rossetto, M., McNally, J., and Henry, R. J. (2002). Evaluating the potential of SSR flanking regions for examining taxonomic relationships in the Vitaceae. *Theor. Appl. Genet.* 104, 61–66. doi: 10.1007/s001220200007
- Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248
- Shan, Y., Li, J., Zhang, X., and Yu, J. (2023). The complete mitochondrial genome of *Amorphophallus albus* and development of molecular markers for five *Amorphophallus* species based on mitochondrial DNA. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1180417
- Skippington, E., Barkman, T. J., Rice, D. W., and Palmer, J. D. (2015). Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all *nad* genes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E3515–E3524. doi: 10.1073/pnas.1504491112
- Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., et al. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 10, e1001241. doi: 10.1371/journal.pbio.1001241
- Sloan, D. B., Alverson, A. J., Storchova, H., Palmer, J. D., and Taylor, D. R. (2010). Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. *BMC Evol. Biol.* 10, 274. doi: 10.1186/1471-2148-10-274
- Sloan, D. B., Warren, J. M., Williams, A. M., Wu, Z., Abdel-Ghany, S. E., Chicco, A. J., et al. (2018a). Cytonuclear integration and co-evolution. *Nat. Rev. Genet.* 19, 635–648. doi: 10.1038/s41576-018-0035-9
- Sloan, D. B., and Wu, Z. (2014). History of plastid DNA insertions reveals weak deletion and at mutation biases in angiosperm mitochondrial genomes. *Genome Biol. Evol.* 6, 3210–3221. doi: 10.1093/gbe/evu253
- Sloan, D. B., Wu, Z., and Sharbrough, J. (2018b). Correction of persistent errors in arabidopsis reference mitochondrial genomes. *Plant Cell* 30, 525–527. doi: 10.1105/tpc.18.00024
- Small, I. A.-O., Schallenberg-Rüdinger, M. A.-O., Takenaka, M., Mireau, H., and Ostersetzter-Biran, O. A.-O. (2020). Plant organellar RNA editing: what 30 years of research has revealed. *Plant J.* 101, 1040–1056. doi: 10.1111/tpj.1457
- Smith, D. R., and Keeling, P. J. (2015). Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci.* 112, 10177–10184. doi: 10.1073/pnas.1422049112
- Stamatakis, A. (2006). RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446
- Sun, M. Y., Zhang, M. Y., Chen, X. N., Liu, Y. Y., Liu, B. B., Li, J. M., et al. (2022). Rearrangement and domestication as drivers of Rosaceae mitogenome plasticity. *BMC Biol.* 20, 181. doi: 10.1186/s12915-022-01383-3
- 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, W., and Luo, C. (2018). Molecular and functional diversity of RNA editing in plant mitochondria. *Mol. Biotechnol.* 60, 935–945. doi: 10.1007/s12033-018-0126-z
- Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017
- Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391
- Wang, X. C., Chen, H., Yang, D., and Liu, C. (2018d). Diversity of mitochondrial plastid DNAs (MTPTs) in seed plants. *Mitochondrial DNA Part A* 29, 635–642. doi: 10.1080/24701394.2017.1334772
- Wang, W., Li, Z. J., Zhang, Y. L., and Xu, X. Q. (2021a). Current situation, global potential distribution and evolution of six almond species in China. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.619883
- Wang, X. Q., Wang, J. X., Ma, X., Zhang, Y. Y., Wang, H. B., Wang, Y. B., et al. (2021b). Effects of Leaf Extracts of *Amorpha fruticosa* on Seed Germination and Seedling Growth of *Amygdalus pedunculata*. *Ying Yong Sheng Tai Xue Bao* 32, 57–65. doi: 10.13287/j.1001-9332.202101.008
- Wang, W., Wang, H. L., Xiao, X. Z., and Xu, X. Q. (2018a). Characterization of the complete chloroplast genome of longstalk almond (*Prunus pedunculata* (Pall.) Maxim.), an important sand-fixation shrub plant endemic to northern China. *Conserv. Genet. Resour.* 11, 419–421. doi: 10.1007/s12686-018-1039-7
- Wang, W., Wang, H. L., Xiao, X. Z., and Xu, X. Q. (2018b). Wild almond (*Amygdalus pedunculata* Pall.) as potential nutritional resource for the future: studies on its chemical composition and nutritional value. *J. Food Measurement Characterization* 13, 250–258. doi: 10.1007/s11694-018-9939-5
- Wang, W., Xu, X. Q., Zhang, Y. L., Shi, F. D., Liu, X. D., Liu, J. Z., et al. (2018c). *Amygdalus pedunculata* Pall (Beijing: China Forestry Publishing).

- Wang, W., Yang, T., Wang, H. L., Li, Z. J., Ni, J. W., Su, S., et al. (2020a). Comparative and phylogenetic analyses of the complete chloroplast genomes of six almond species (*Prunus* spp. L.). *Sci. Rep.* 10, 10137. doi: 10.1038/s41598-020-67264-3
- Wang, X., Zhang, R., Wang, J., Di, L., Wang, H., and Sikdar, A. (2020b). The effects of leaf extracts of four tree species on *Amygdalus pedunculata* seedlings growth. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.587579
- Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3
- Wee, C.-C., Nor Muhammad, N. A., Subbiah, V. K., Arita, M., Nakamura, Y., and Goh, H.-H. (2022). Mitochondrial genome of *Garcinia mangostana* L. variety Mesta. *Sci. Rep.* 12 (1), 9840. doi: 10.1038/s41598-022-13706-z
- Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383
- Wu, Z. Q., Liao, X. Z., Zhang, X. N., Tembrock, L. R., and Broz, A. (2020b). Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *J. Systematics Evol.* 60, 160–168. doi: 10.1111/jse.12655
- Wu, S., Liu, W., Aljohi, H. A., Alromaih, S. A., Alanazi, I. O., Lin, Q., et al. (2018). REDO: RNA editing detection in plant organelles based on variant calling results. *J. Comput. Biol.* 25, 509–516. doi: 10.1089/cmb.2017.0214
- Wu, Z., Sloan, D. B., Brown, C. W., Rosenblueth, M., Palmer, J. D., and Ong, H. C. (2017). Mitochondrial retroprocessing promoted functional transfers of rpl5 to the nucleus in grasses. *Mol. Biol. Evol.* 34, 2340–2354. doi: 10.1093/molbev/msx170
- Wu, Y. P., Su, T. J., and He, B. (2020a). Complete mitochondrial genome of *Plodia interpunctella* (Lepidoptera: Pyralidae). *Mitochondrial DNA B Resour* 5, 583–585. doi: 10.1080/23802359.2019.1710590
- Wynn, E. L., and Christensen, A. C. (2019). Repeats of unusual size in plant mitochondrial genomes: identification, incidence and evolution. *G3 (Bethesda)* 9, 549–559. doi: 10.1534/g3.118.200948
- Xiong, Q., Bai, Q., Li, C., He, Y., Shen, Y., and Uyama, H. (2018). A cellulose acetate/*Amygdalus pedunculata* shell-derived activated carbon composite monolith for phenol adsorption. *RSC Adv.* 8, 7599–7605. doi: 10.1039/c7ra13017a
- Yan, H., Ma, S. M., Wei, B., Zhang, H. X., and Zhang, D. (2022). Historical distribution patterns and environmental drivers of relict shrub: *Amygdalus pedunculata*. *Chin. J. Plant Ecol.* 46, 766–774. doi: 10.17521/cjpe.2021.0406
- Yao, L., Li, H., Yang, J., Li, C., and Shen, Y. (2018). Purification and characterization of a hydroxynitrile lyase from *Amygdalus pedunculata* Pall. *Int. J. Biol. Macromol* 118, 189–194. doi: 10.1016/j.ijbiomac.2018.06.037
- Yazbek, M., and Oh, S. H. (2013). Peaches and almonds: phylogeny of *Prunus* subg. *Amygdalus* (Rosaceae) based on DNA sequences and morphology. *Plant Systematics Evol.* 299, 1403–1418. doi: 10.1007/s00606-013-0802-1
- Yu, X., Duan, Z., Wang, Y., Zhang, Q., and Li, W. (2022). Sequence Analysis of the Complete Mitochondrial Genome of a Medicinal Plant, *Vitex rotundifolia* Linnaeus f. (Lamiales: Lamiaceae). *Genes* 13 (5), 839. doi: 10.3390/genes13050839
- Zanduetta-Criado, A., and Bock, R. (2004). Surprising features of plastid ndhD transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. *Nucleic Acids Res.* 32 (2), 542–550. doi: 10.1093/nar/gkh217
- Zardoya, R. (2020). Recent advances in understanding mitochondrial genome diversity. *F1000Res* 9 (F1000 Faculty Rev), 270. doi: 10.12688/f1000research.21490.1
- Zhang, S., Wang, J., He, W., Kan, S., Liao, X., Jordan, D. R., et al. (2023). Variation in mitogenome structural conformation in wild and cultivated lineages of sorghum corresponds with domestication history and plastome evolution. *BMC Plant Biol.* 23 (1), 91. doi: 10.1186/s12870-023-04104-2
- Zhang, L., Yang, X., Qi, X., Guo, C., and Jing, Z. (2018). Characterizing the transcriptome and microsatellite markers for almond (*Amygdalus communis* L.) using the Illumina sequencing platform. *Hereditas* 155, 14. doi: 10.1186/s41065-017-0049-x
- Zhao, Y. (1992). *Rare and Endangered Plants in Inner Mongolia*. (Beijing: China Agriculture Science and Technique Press).
- Zuo, S. Y. (2016). *Analysis on Genetics Characteristics and important traits evaluation of Prunus pedunculata Maxim population*. Master's Thesis. (Changsha: Central South University of Forestry and Technology).



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

REVIEWED BY

Yun-peng Du,
Beijing Academy of Agricultural and
Forestry Sciences, China
Adil Hussain,
Abdul Wali Khan University, Pakistan

*CORRESPONDENCE

Chunsong Cheng
✉ chengcs@lsbg.cn
Puxin Gao
✉ gaopx@lsbg.cn

†These authors have contributed equally to
this work

RECEIVED 08 September 2023

ACCEPTED 20 November 2023

PUBLISHED 18 December 2023

CITATION

Yang J, Fan S, Guo M, Xie Z, Cheng Q,
Gao P and Cheng C (2023) DNA barcoding
and comparative RNA-Seq analysis provide
new insights into leaf formation using
a novel resource of high-yielding
Epimedium koreanum.
Front. Plant Sci. 14:1290836.
doi: 10.3389/fpls.2023.1290836

COPYRIGHT

© 2023 Yang, Fan, Guo, Xie, Cheng, Gao and
Cheng. 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.

DNA barcoding and comparative RNA-Seq analysis provide new insights into leaf formation using a novel resource of high-yielding *Epimedium koreanum*

Jiaxin Yang^{1†}, Siqing Fan^{1†}, Min Guo¹, Zhaoqi Xie¹,
Qiqing Cheng^{1,2}, Puxin Gao^{1*} and Chunsong Cheng^{1,3*}

¹Lushan Botanical Garden, Chinese Academic of Sciences, Jiujiang, China, ²School of Pharmacy, Hubei University of Science and Technology, Xianning, China, ³National Resource Center for Chinese Materia Medica, Chinese Academy of Chinese Medical Sciences, Beijing, China

Epimedium koreanum Nakai, a well-known traditional Chinese medicinal herb, has been widely used to treat osteoporosis and sexual dysfunction for thousands of years. However, due to the decreasing population of East Asian natural resources, yearly output of *Epimedium* crude herb has been in low supply year by year. In this study, an unusual variety of *E. koreanum* was discovered in Dunhua, Jilin Province, the northernmost area where this variety was found containing 6 individuals, with three branches that had 27 leaflets, which is much more than the typical leaflet number of 9. Firstly, the novel *E. koreanum* variety was identified using DNA barcodes. Then, 1171 differentially expressed genes (DEGs) were discovered through parallel RNA-seq analysis between the newly discovered variety and wild type (WT) *E. koreanum* plant. Furthermore, the results of bioinformatics investigation revealed that 914 positively and 619 negatively correlated genes associated with the number of leaflets. Additionally, based on RNA-Seq and qRT-PCR analysis, two homologous hub *TCP* genes, which were commonly implicated in plant leaf development, and shown to be up regulated and down regulated in the discovered newly variety, respectively. Thus, our study discovered a novel wild resource for leaf yield rewarding medicinal *Epimedium* plant breeding, provided insights into the relationship between plant compound leaf formation and gene expression of *TCPs* transcription factors and other gene candidates, providing bases for creating high yield cultivated *Epimedium* variety by using further molecular selection and breeding techniques in the future.

KEYWORDS

Epimedium, DEGs, compound leaf, *EkTCP*, plant selection, leaf yield

1 Introduction

Epimedium is the largest herbaceous genus in berberidaceae family (Zhang et al., 2007; Zhang et al., 2022). Herbal *Epimedium*, first described in the *Shen Nong Herbal Classics*, is a prominent traditional Chinese medicinal plant, that has been widely used to treat osteoporosis and sexual dysfunction for thousands of years. There are currently 62 *Epimedium* species, 52 species are indigenous to China (Zhang et al., 2022). The published papers showed that *Epimedium* leaves contain a large number of flavonoid chemical components, which provides beneficial properties including anti-cancer effects and the treatment of cardiovascular diseases, rheumatoid arthritis, osteoporosis, and immune enhancement (Li et al., 2001; Ma et al., 2011; Jiang et al., 2015; Indran et al., 2016; Jiao et al., 2021; Liu et al., 2021; Zhu et al., 2021). In the *Pharmacopoeia of the People's Republic of China* (Ch. P), total flavonoids and flavonoid glycoside quantitation serve as indicators of the herbal *Epimedium* quality. Additionally, polysaccharides, another medicinal component of *Epimedium*, have antiviral, anti-aging, and immune-regulating activities.

The majority of *Epimedium* research focused on the therapeutic benefit and chemical potential of its metabolites. However, there have been few studies that dived into botany and plant physiology, specifically the molecular processes that regulate the growth and development of herbal *Epimedium*. Among the statutory 5 medicinal *Epimedium* species recorded in the Ch. P, *Epimedium koreanum* is one of the species with a large amount of wild resources across in the two northeastern provinces of China, Jilin and Liaoning provinces. It primarily spreads in Eastern Asia, with notable distribution in China, Korea, and Japan (Lee et al., 2016; Qian et al., 2023). While the published papers on *E. koreanum* have mainly focused on recourse protection and sustainable utilization, including resource collection, cultivation, geographical distribution characteristics, and medical and pharmaceutical applications of its metabolites (Zhong et al., 2017; Zhang et al., 2020), and there have been few investigations on its plant physiology and molecular characteristics. (Lee et al., 2016). *E. koreanum* has been in low supply in the Chinese herbal medicine market in recent years due to the depletion of natural resources across East Asia, and its price has been rising year by year. With the foreseeable rising market demands, the crude herb of wild *E. koreanum* supplies are depleting, and cultivated *E. koreanum* will become the most important raw materials in the primary market. So, it is critical to use molecular techniques to create and select new *E. koreanum* varieties.

The principal photosynthetic organs of flowering plants, with great diversity in number, shape, and structure. Leaves are scientificlly classified into two categories based on the number of leaflets and the structure of the leaf: simple leaves and complex leaves with many leaflets. A compound leaf, distinguished by its potential to take on many forms such as pinnate and palmate compound leaves, is made up of numerous discontinuous leaf units attached to the rachis and petiole, as opposed to a simple leaf, which is a single unit (Kim et al., 2003). Each leaflet in a complex leaf provides the same photosynthetic function as a simple leaf. From a functional aspect, each leaflet fulfills the same job as a simple leaf; hence, the development of complex leaves

may boost plants' capacity for photosynthetic energy generation and plant survival rate (Laura et al., 2010).

At present, the mechanism of compound leaf development and formation is not well understood. Leaf shape develops from the apical meristem (SAM) of the plant, while the leaf primordium cells develop from the flanks of SAM. Leaf development must arise in three distinct and overlapping stages: The first stage is leaf initiation, in which the leaf primordium differentiates from the flank of SAM; this is followed by primary morphogenesis (PM), in which leaf margin structures such as blade, serrate, and lobes begin to form; and finally, secondary morphogenesis (SM), which determines the final size and shape of the leaf (Dengler and Tsukaya, 2001; Bar and Ori, 2015). Leaf development is regulated by transcription factors and phytohormone networks. One of the essential genes involved in regulating leaf growth is the plant specific transcription factors *TEOSINTE BRANCHED1 CYCLOIDEA PROLIFERATING CELL FACTOR* (TCP). Various TCP transcription factors have been identified as critical modulators of leaf architecture. For instance, in tomato (*Solanum lycopersicum*), the TCP gene family member *LANCEOLATE* (LA) was showed exhibiting the premature leaf differentiation in the gain-of-function mutant *La-2*, resulting in a single-leaf pattern. Conversely, the loss of function mutant *la-6* displayed highly fragmented leaf margin shape, indicating the involvement of LA in leaf development (Ori et al., 2007). In *Arabidopsis*, the microRNA *miR319* was found to down-regulate the expression of TCP gene family members, thereby influencing leaf morphogenesis (Koyama et al., 2017). Similarly, *TCP13* was showed to regulate leaf and root growth in response to drought conditions in *Arabidopsis* (Urano et al., 2022). Furthermore, in lettuce (*Lactuca sativa* L.), the *LsAP2* gene promoted the leaf division by inhibiting TCP transcription factor activity, emphasizing the significance of TCPs in leaf development across different plant species (Luo et al., 2021). In addition, aside from the compound leaf plant tomato, there have few number of studies investigating compound leaf development in other plant species such as medicago (*Medicago truncatula*) and pea (*Pisum sativum*). In medicago, the *PINNATE LIKE PENTAFOLIATA1* (PINNA1) gene was identified as a key regulator of terminal leaflet morphogenesis. It works by inhibiting the expression of the *FLORICAULA/LFY* homologous gene, *SINGLE LEAFLET1* (SGL1), thereby suppressing the formation of lateral leaflets (He et al., 2020). Similarly, in peas, research focused on the *afila* (*af*) mutant, which exhibits increased leaf complexity. This increase in leaf complexity is accompanied by elevated expression of the *UNIFOLIATA* (UNI) gene. Further investigations involving the double mutant *af tendril* (*tl*) also revealed a synergistic effect, with heightened UNI expression suggesting that AF and TL jointly inhibit leaflet formation (Mishra et al., 2009; Demason et al., 2013). While there has been considerable research on compound leaf development in tomato, pea, and medicago, studies investigating the mechanism of compound leaf development in medicinal plants are currently lacking for scientific community. This knowledge gap is especially concerning in light of the scientific and practical consequences for medicinal or commercial plants that relays on complex leaves as harvesting and therapeutic components. Understanding the development and regulation of leaflets in medicinal plant species holds immense

potential for enhancing their cultivation, yield, and medicinal properties. Therefore, imperious demands are expected to investigate compound leaf development in medicinal plants and uncover the underlying molecular mechanisms.

During extensive field investigations and resource collection encompassing various species of *Epimedium*, our research team made an intriguing discovery that the presence of a distinct *Epimedium* plant displayed remarkable characteristics in the primary growth region of *E. koreanum* in northeast China. This unique wild resource exhibited an unprecedented number of leaflets, ranging from 11 to 27, surpassing the typical 9 leaflets observed in *E. koreanum*. To unravel the taxonomic implications associated with this finding, we rigorously employed DNA barcode analysis and constructed an evolutionary tree, aiming to ascertain whether this variant represented a new species of *E. koreanum*. Subsequently, we performed comprehensive investigations on this “super *Epimedium*” plant. Fresh leaf tissues were carefully sampled and subjected to global RNA sequencing (RNA-Seq) and qRT-PCR analysis, allowing us to delve into the intricate molecular mechanisms underlying its exceptional leaf development. Through rigorous analysis of differential gene expression (DEGs) and comparison of gene expression patterns, we gained valuable insights into the critical signaling pathways involved in leaf development. By integrating these findings with important molecular pathways, we aimed to gain a holistic understanding of compound leaf morphogenesis in *E. koreanum*. Ultimately, our study endeavored to improve our understanding of compound leaf morphogenesis in *E. koreanum* while also opening up novel prospects for future selection and breeding.

2 Materials and methods

2.1 Sample collection

The material used in this study were *E. koreanum* and the discovered newly variation. The newly variation named *E. koreanum* var. *polyphylla* CS Cheng (EKP) in this study collected from in Dunhua City Jilin Province (E: 128.0369, N: 43.1156, A: 670) and ex-situ cultivated in the E115°59', N29°51' at Lushan Botanical Garden.

2.2 Sequence mining and primer design

The nucleic acid sequences of each species within the genus *Epimedium* were downloaded from the NCBI website (<https://www.ncbi.nlm.nih.gov>) by searching with “‘*Epimedium*’ [organism] and ‘*matK*’ [All files]”. The *ITS*, *rbcl*, and *trnL-trnF* sequences were also searched and downloaded in a similar manner. Save the downloaded sequences with Geneious primer 2021 software (Kearse et al., 2012) for the subsequent steps of nucleic acid sequence SNP analysis. The *matK* sequence (GenBank: AB069837.1) of *E. koreanum* species was used as the reference

sequence and Primer premier 5.0 was used for primer design (Li et al., 2009). The primer sequences were designed as:

matK forward primer (FmatK): 5'-TATGACAATAAATCC AGTTC-3'

matK reverse primer (RmatK): 5'-ATGCCCCGATACGTTA CAAA-3'

2.3 DNA extraction and DNA sequencing

Genomic DNA from leaves was isolated using the standard Cetyltrimethyl ammonium bromid (CTAB) extraction protocol (Helliwell et al., 2016). The targeted sequences were amplified with specific primers. The standard 50 µL PCR reaction mixture contained 25 µL of 2 × PrimeSTAR[®] Max DNA Polymerase (Takara, Code NO. R045A) and 10 ng of template DNA, 0.2 µM of each primer. The samples were amplified using a Verit 96- Well Fast Thermal Cycler (Applied Biosystems, Foster City, CA, USA) under the following conditions: initial denaturation at 94 °C for 5 min, followed by 30 cycles of denaturation at 98 °C for 10 s, annealing at 54 °C for 10 s, extension at 72 °C for 30 s, and a final elongation step at 72 °C for 7 min. The PCR products were confirmed by 1.0% agarose gel electrophoresis in 1 × TAE buffer to detect whether the target sequences were cloned successfully. The amplicons were purified with an TaKaRaMinBEST Agarose Gel DNA Extraction Kit Ver.4.0 (Takara, Code No.9762) and quantified with a NaoDrop 2000 spectrophotometer (Thermo Fisher Scientific) (Lin et al., 2021). And then the target fragments were sent to BGI (Wuhan, China) for bidirectional sequencing.

2.4 Sequence analysis

To discover the SNP of sequences, sequenced sequences and *matK* sequences downloaded from NCBI were subjected to multiple sequence alignments through Geneious primer 2021 software with default settings. Polymorphic locis were counted after multiple alignment for different species. In the *Geneious* software, the head and tail of the aligned *matK* sequences were cut off while removing the gaps in different species, respectively. Then, intraspecific genetic distances were calculated as SNP%. Six species in the *Epimedium* genus were randomly selected, and all *matK* sequences of each species were randomly sampled with put-back six times for sequence alignments, and then the calculated SNP% was used as the interspecific genetic distance. The *ITS*, *rbcl*, and *trnL-trnF* sequences were also handled as described above.

2.5 Phylogenetic analysis

Test sequences from the collected samples and *matK* sequences downloaded from NCBI for all species of the *Epimedium* genus were used for multiple sequence alignments through MEGA-X (Kumar et al., 2018) with the default setting. Then, they were

performed cutting the head and tail, as well as aligning the gaps in the sequence. To compare the evolutionary relationships, the results of the above alignments and processing were used to construct the phylogenetic tree using MEGA-X with Maximum Likelihood (ML) method. The phylogenetic tree was then visualized by EVOLVIEW. (<https://www.evolgenius.info/evolview/#login>).

To build phylogenetic tree of *TCP* proteins, the *TCP* protein sequences were first aligned with MAFFT (Version 7.037b) (model: “BLOSUM62”, strategy: “L-INS-i”) (Kato and Standley, 2013), then refined conserved sequences from the alignments by Gblocks (Version 0.91b) (Maximum number of contiguous nonconserved positions: 32000, Minimum length of a block: 2, Allowed gap positions: all) (Castresana, 2000). The neighbor-joining phylogenetic tree was finally generated with Mega 6 (Tamura et al., 2013).

2.6 Geographic analysis

According to the origin information of different species of *Epimedium* genus (<http://www.iplant.cn/>, <http://www.plantsoftheworldonline.org/>), the latitude and longitude of origins were also found and recorded by Google Earth. The locations were projected to the provincial boundary map of China depended on the latitude and longitude data by ArcGIS10 software to observe and analysis the distribution of different groups.

2.7 mRNA sequencing

The newly developed leaves of *E. koreanum* in the rainy season of August in Jilin Province were collected for molecular sequencing analyses. The two groups of *E. koreanum* with significant differences in leaf shape and number of leaflets of compound leaves were named variety and normal. And, each sample was blended with several leaflets from the same biological source, and at least three biological duplicate samples were chosen for each group in this investigation. All samples were powdered by liquid nitrogen quick-freezing and then RNA was extracted. Nanodrop 2000 (ThermoFisher) was used for purity and concentration detection of the extracted RNA, RNA integrity was detected by agarose gel electrophoresis, and Agilent 2100 was used to determine the RIN value. Four samples were sent to BGI Genomics Co., Ltd. (East Lake Development Area, Wuhan, China) for library preparation and RNA sequencing. Using the BGISEQ-500 sequencing platform, sequencing data quality control included sequencing data statistics, original data statistics and quality control data statistics.

2.8 Transcriptome data analysis

Data filtering: The raw data obtained from sequencing was filtered using the filtering software fastp (Chen et al., 2018) to remove reads containing adapters (adapter contamination), reads with unknown base N content greater than 5%, and low-quality reads (reads with a quality value below 15 that account for more than 20% of the total bases in the read). *De novo* assembly and

quality assessment: Clean reads were assembled *de novo* using Trinity (Grabherr et al., 2011), and their assembly quality was evaluated using BUSCO. Reference gene alignment: Clean data was aligned to reference gene sequences using Bowtie 2 (v2.2.5) (Langmead and Salzberg, 2012) software, and gene and transcript expression levels were calculated using RSEM software (Li and Dewey, 2011). CDS prediction: Candidate coding regions within transcripts were identified using Transdecoder software (Kim et al., 2015), and BLASTed against SwissProt and searched for Pfam protein homologous sequences were using the Hmmscan to predict coding regions. Gene annotation: Transcripts were annotated with seven major functional databases (KEGG, GO, NR, NT, SwissProt, Pfam, and KOG). WGCNA analysis: Gene co-expression networks were analyzed using WGCNA (v1.48). Differentially expressed genes: Group difference gene analysis was performed using DESeq 2 (Love et al., 2014), with the condition that Fold Change ≥ 1 and padj value (after multiple correction) was less than 0.05. Based on GO annotation results and official classification, differentially expressed genes were classified functionally, and GO enrichment analysis was performed using the clusterProfile package (Wu et al., 2021). A threshold of $q\text{-value} \leq 0.05$ was used in where a definition of significant enrichment in candidate genes was met.

2.9 Validation of differential gene expression by qRT-PCR analysis

RNA extraction and cDNA synthesis: Total RNA was extracted from the frozen leaf samples with the MiniBEST Plant RNA Extraction Kit (TaKaRa, China). The NanoDrop ND1000 spectrophotometer (Thermo, USA) was used to calculate the RNA concentration and assess purity. The RNA samples with a 260/280 nm absorbance ratio of 1.8–2.0 were retained for further analyses. The RNA integrity was evaluated by 1% agarose gel electrophoresis. The HiScript@III RT SuperMix for qPCR (Vazyme) was used to synthesize cDNA. The qRT-PCR assay was completed with SYBR qPCR Master Mix (Vazyme) and The LightCycle 480 Instrument II (Roche). The reaction solution consisted of 5 μL SYBR qPCR Master Mix (Vazyme), 4 μL cDNA (100 ng), 0.5 μL 10 μM forward primer, 0.5 μL 10 μM reverse primer for a final volume of 10 μL (Table 1). The amplification conditions were as follows: 95°C for 3 min; 40 cycles of 95°C for 10 s, 58°C for 10 s, and 72°C for 25 s, followed by a melting curve analysis from 60 to 95°C. The gene expression levels for each sample were determined based on three replicates.

2.10 Statistical analysis

Graphpad Prism 7 software was used to analyze the data. All numerical values were presented as mean \pm SEM and the sequence type was indicated in the legends. Statistically significant differences between inter- and intraspecific distance were determined by pair using t-test, with p values < 0.05 . The three-dimensional structure of TCP proteins were built in SWISS-MODEL (<https://swissmodel.expasy.org/>) (Waterhouse et al., 2018).

TABLE 1 Primer sequences used in this study.

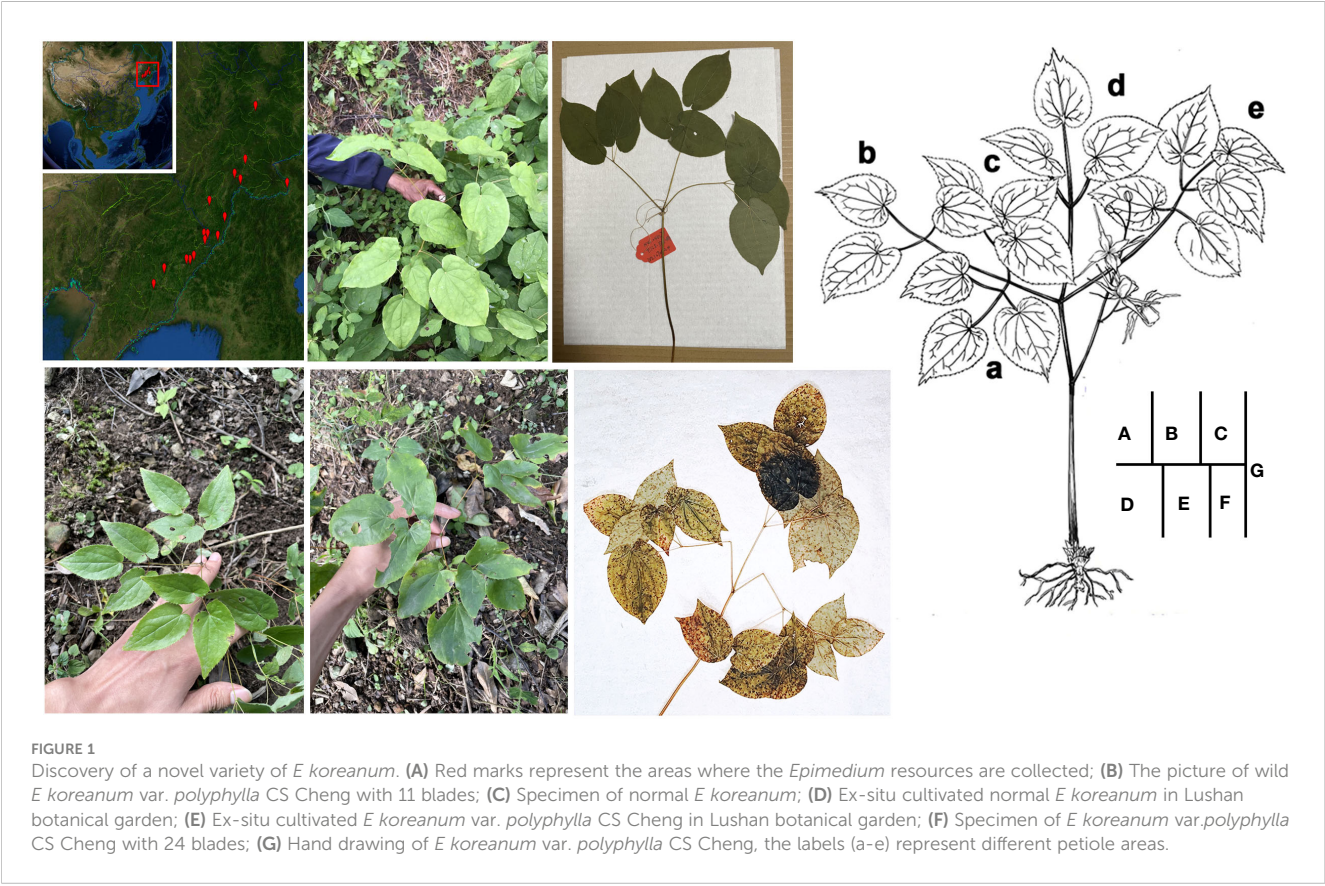
Primer name	Primer sequence
EkTCP14-QF	5'-ATGGGAGATACCAAACCAAGTGAAA-3'
EkTCP14-QR	5'-TCCACCTTTGTGTCTGTCCT-3'
EkTCP9-QF	5'-CCGATGTGGGCGATGGT-3'
EkTCP9-QR	5'-GAACTTGAATAGCGTTTGCCAT-3'
EkACT2-QF	5'-GCCATTCAAGGCTGTTCTTC-3'
EkACT2-QR	5'-GGTAAGATCGCGACCTGCTA-3'
EkSERK1-QF	5'-CACTTTTCTATTGCTGTCGC-3'
EkSERK1-QR	5'-TTTACCAAAACCTCTCTACCC-3'
EkACC2-QF	5'-GAGACGAATCATACCGCTCACC-3'
EkACC2-QR	5'-GAGCGAAGAGCCGAACCTAC-3'
EkSUS4-QF	5'-GCTTTTCATTCCTTTCCCGT-3'
EkSUS4-QR	5'-AGAAGAGAGACTAATTCGTTGCG-3'
EkSPL1-QF	5'-TGGATTCTGGGGCATAG-3'
EkSPL1-QR	5'-CCAATAAACCAACCAAGCCTT-3'
EkARP8-QF	5'-TTGGGATACTATTGTTTCGCC-3'
EkARP8-QR	5'-CGAGGCATATTTACTCCAACCA-3'
EkCDKE-1-QF	5'-ATCTTCTACCGCCCTACTTTT-3'
EkCDKE-1-QR	5'-GCCTGTTGGGATTGTTGTAGT-3'

3 Results

3.1 Distinctive resources discovered in the wild population of *Epimedium koreanum*

Referring to the records of plant specimens collected in China mainland. The sample collection of resources of *E. koreanum* was implemented in August 2021, covering the whole areas along the Yalu River in Jilin and Liaoning provinces (Figure 1A). A distinctive

wild population of *Epimedium* was discovered in Dunhua City Jilin Province (E: 128.0369, N: 43.1156, A: 670). The discovered newly resource contains 6 individuals which grow mixed with ordinary *E. koreanum* under the same masson pine forest. It was hard to tell whether a newly discovered exceptional resource belongs to a budding mutation or a new species on the spot. Its reproductive organs were not appreciably different from those of the typical *E. koreanum* (Figures 1C, D). The number of their leaves generally exceeds 9, and reaching a maximum of 27, its flower stem often has more than 1 biternate leaf, which is a significant feature of this population (Figures 1B, E–G). In details, this wild *Epimedium* is also a perennial herb, with 20–45 cm tall, rhizome creeping, triternate leaves, 11–27 foliolate, leaflets ovate, abaxially pallid but adaxially dark green, 6–16 × 4–12 cm, papery, glabrous, base deeply cordate with usually round lobes, margin minutely serrate, apex acute or acuminate. Flowering stem with 2–3 biternate leaf. Simple raceme inflorescence 12–17 cm with 6–16 flowered, glabrous. Pedicel 1–2 cm, flowers yellowish-white, 2–5 cm in diam. Outer sepals reddish or pale yellow, 5–8 mm, inner sepals narrowly ovate to lanceolate and apically acute. Petals nearly twice longer than the inner sepals, spurs slender, elongate and tapering subulate, 1–2 cm. The stamens ca 6.5 mm, anther ca 5 mm, filaments ca 1.5 mm, pistil ca 8 mm, ovary ca 5 mm, capsules 6–12 mm long, and 2–4 mm broad. Seeds usually 5–7. Fl. May, fr. May. Based on the phenotype of the newly resource with strong biomass advantage and the identification of medicinal plant taxonomy by expert Dr. Cheng Chung, we suspected that this is a variety of *E. koreanum*. So, we tentatively named the distinctive resource with *E. koreanum* var. *polyphylla* CS



Cheng (EKP) in this study. This exceptional germplasm resource was grown ex-situ at the Lushan Botanical Garden. Moreover, after one year of phenological records performed, most of the phenotypes of complex leaf leaflets may still be maintained (Figure 1E), however the number of leaflets may fluctuate due to soil nutrition or climatic environment.

3.2 Molecular identification and phylogenetic tree analysis

The morphological identification of *Epimedium* species and the molecular classification based on DNA barcoding are both challenging (Zhang et al., 2016; Ren et al., 2018; Zhang et al., 2022). However, the application of DNA sequencing and barcoding is undoubtedly crucial for evaluating a new plant resource. In this study, the applied DNA sequences (DNA barcodes) (Vijayan and Tsou, 2010; Kim et al., 2016), including *ITS*, *matK*, *rabL* and *trnL-trnF* were evaluated for usability by using the degree of differences in DNA sequences within and between species, also known as intraspecific distance and interspecific distance. According to the DNA sequences statistics of *Epimedium* plants published by NCBI, only *matK* sequence was considered to be suitable for next step of genetic analysis and species identification of the new plant resource, because its interspecific difference was significantly higher than intraspecific difference ($P < 0.01$, $N > 6$) (Figure 2A). The other investigated sequences included *ITS*, *rabL*, and *trnL-trnF*, although all showed an average interspecific distance greater than the average intraspecific distance, but the *t*-test result showed no significant differences ($P > 0.05$, $N > 6$). Herein, specific primers were designed

at the beginning and end of the referenced sequence to maximize the amplification of the *matK* sequence. In order to balance the credibility of public sequences with the longest possible sequence length, the referenced *matK* sequences were artificially divided into three parts (a, b, and c), differential DNA base count results showed that part b covers the most SNPs in *Epimedium* plants ($N > 50$, $P < 0.001$) (Figures 2B, C).

The PCR products were sequenced by Sanger sequencing and verified by positive and negative sequencing (Figure 3A). A specific homozygous mutation (T) was generated at site 333, compared to the typical *E. koreanum* population (G) (Figures 3B, C). Mutations specific to this site appeared to be rare, since all public *matK* sequences showed that only *E. perralderianum* and *E. pinnatum* from the highest latitudes or altitudes ($N > 34^\circ$) have genotype T/T at this site (Figures 3B, 4A, S1). This homozygous SNP site was obviously not created by sequencing errors (Figure 3A), and the function of the *matK* gene corresponding to the SNP at this site may be relevant to the evolution of high-latitude plant species. Overall, the confirmation of this discovered SNP verified our idea that this novel resource discovered in wild *E. koreanum* population can be identified as a variety. Although the study of the novel SNP and its kinase function corresponding to the *matK* gene belongs to an interest scientific issue, as one of the rapidly evolving genes in plant chloroplast genome, *matK* sequence is often used to assist in the identification of species below plant genera. Therefore, in this study, only the phylogenetic tree of *matK* sequence was discussed.

As shown in Figure 4A, the maximum likelihood method based phylogenetic tree was conducted and revealed that the monophyletic origin in *Epimedium* genus, which was consistent with other reports

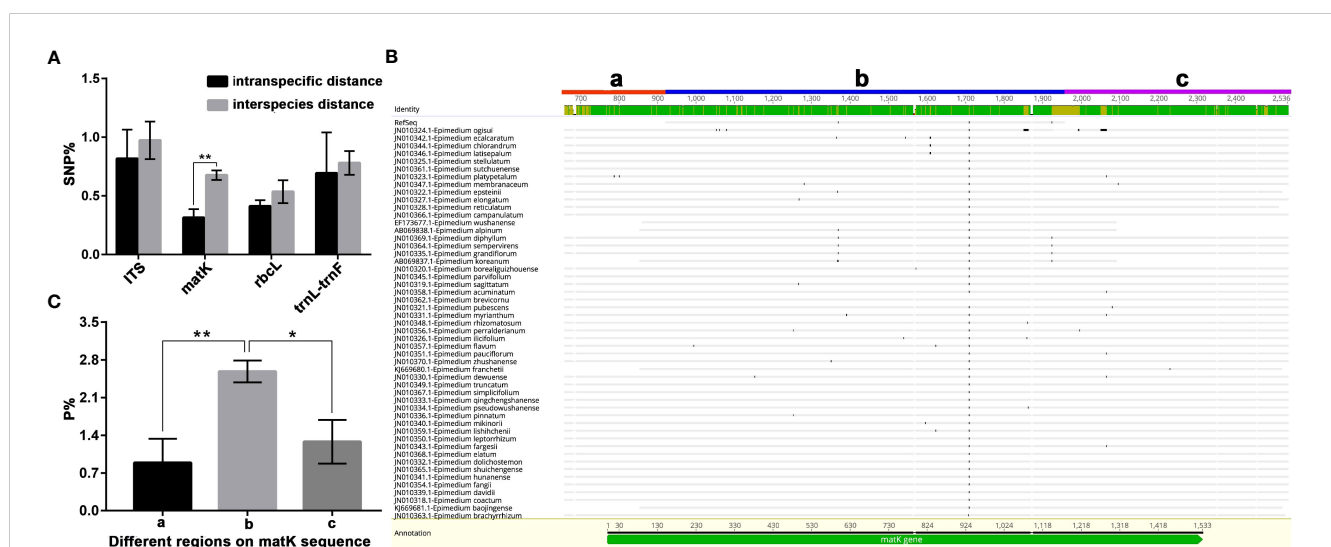


FIGURE 2

Scientific screening of commonly used DNA barcodes for assisted identification of *Epimedium* species. (A) The SNP statistics of *Epimedium* genus DNA barcodes within and between species; (B) Homologous alignment analysis of *matK* sequence in the chloroplast genome fingerprint region of *Epimedium* species, a and c represent the 3' and 5' ends of the *matK* sequence, respectively (prone to sequencing errors or deletions due to primers), and b represents the region with high degree of sequence homology in *Epimedium* genus; (C) The distribution and statistics of SNPs in the *matK* sequence of the genus *Epimedium*. Data presented as means \pm SEM ($n > 6$). T-test, $**p < 0.01$ $*p < 0.05$; the dots represent single nucleotide mutation sites, with the first sequence as a reference.

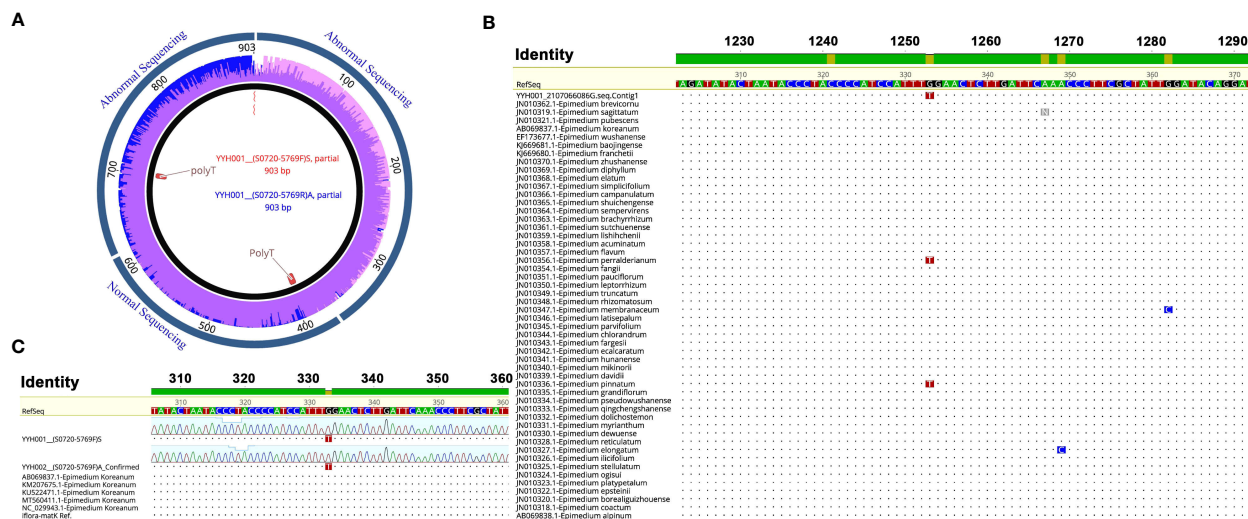


FIGURE 3

Discovery of mutation sites on *thematK* sequence of *E. koreanum* var. *polyphylla* CS Cheng. (A) The sequencing electropherogram of the *matK* sequence of *E. koreanum* var. *polyphylla* CS Cheng. (B) Alignment of *matK* sequences among species of the *Epimedium* genus. (C) Alignment of *matK* sequences from *E. koreanum*.

(Zhang et al., 2007; Guo et al., 2022). An important group including the *E. koreanum*, which is entirely located at high latitudes and is of particular concern to this study. There were 8 *Epimedium* species in this group, including *E. pinnatum*, *EKP*, *E. grandiflorum*, *E. sempervirens*, *E. diphyllosum*, *E. alpinum*, *E. pinnatum* and *E. perralderianum*. Furthermore, haplotype parsimony network analysis of *matK* sequences diversity obtained from sampling 52 sequences (Figure 4B). The network analysis revealed a more distinct genealogical connect among the eight *Epimedium* species indicated above, which are found in high latitudes all over the world.

3.3 DEGs analysis for exploring molecular mechanism of leaflets increasing

We investigated the differences in transcription expression levels between *EKP* and *E. koreanum* by second-generation transcriptome sequencing. We used Trimmed Mean of M-values (TMM) to analyze gene transcription levels. The volcano map showed the number of differed genes between *EKP* and normal *E. koreanum*, and showed significantly up-regulated, down-regulated, and non-significant genes in different colors

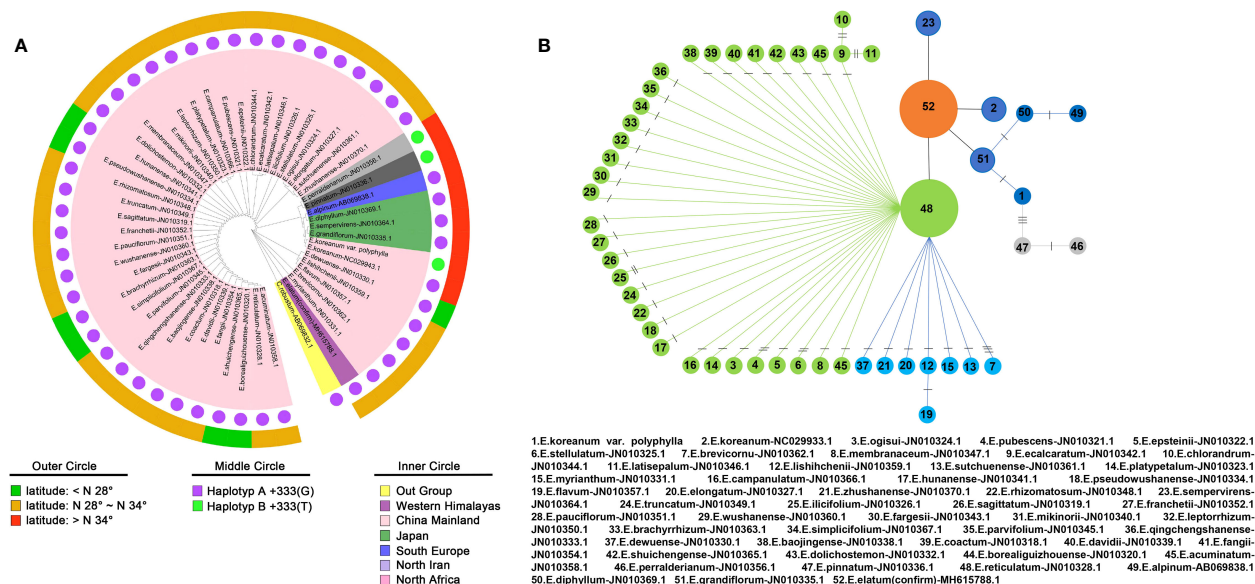
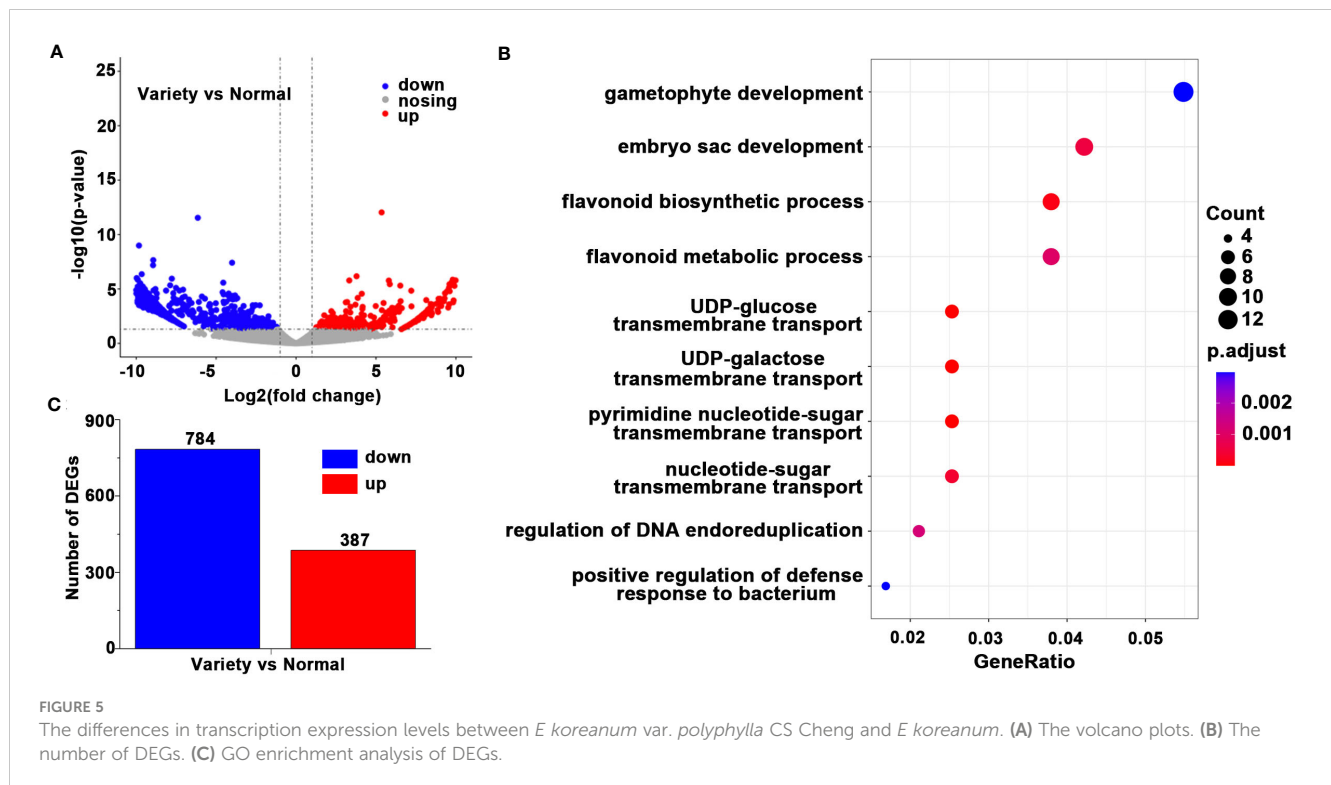


FIGURE 4

Research on the monophyletic origin of the *Epimedium* genus. (A) Phylogenetic tree based on the maximum likelihood method. (B) Haplotype parsimony network analysis of the diversity of 52 *matK* sequences in the *Epimedium* genus.



(Figure 5A). The genes that were considerably up-regulated and down-regulated in *EKP* were statistically analyzed in a bar chart, and it was found that 387 genes were significantly up-regulated and 784 genes were significantly down-regulated when compared to the typical *E. koreanum* (Figure 5C). Table was used to show DEGs in this study (Table S1). With these transcripts, we then performed GO enrichment analysis, and we chose the top 10 modules to present (Figure 5B). It was discovered that DEGs were primarily enriched in the development of gametophytes, embryo sacs, flavonoid biosynthetic pathways, and flavonoid metabolic pathways. The majority of the genes involved in the differentiation and proliferation of leaf primordium cells were associated with gametophyte and embryo sac development. The involvement of plant hormones such as auxin and ethylene is widely regarded as important roles in the synergistic regulation of gametophyte and embryo sac development. So these findings provided new insights into *E. koreanum* leaf growth and flavonoid metabolism.

3.4 Weighted gene co-expression network analysis for discovery of leaf development related genes

In order to identify the DEGs related to leaf development in *E. koreanum*, we further analyzed leaf development as a trait by WGCNA (Langfelder and Horvath, 2008). Twenty-one modules were discovered and colored differently. The twenty-one modules' gene counted ranged from 120 to 3154 (Figures 6A, C). In this research, we focused on the two main modules. The pink module involved 914 transcripts, was positively correlated with leaf

development. The green yellow module involved 619 transcripts were negatively correlated with leaf development (Figure 6B). Thus, these results provided a further understanding of leaf development of *E. koreanum*.

3.5 Identification of hub genes regulate to the leaf development in *E. koreanum*

We performed a visual network analysis of these two modules, identified the top 30 most reliable nodes for visualization, and discovered that these hub genes were involved in leaf development (Figures 7A, B). Because no whole genome data for *E. koreanum* is now publicly available, we used homologous alignment to annotate these genes and discovered *TRINITY_DN4142_c0_g1_i1*, a *TCP14* transcription factor homologous gene in green yellow module, which has previously been reported to inhibit cell proliferation in leaf tissues (Kieffer et al., 2011). Besides, we also found other genes may involve in regulating leaf development, such as *SERK1* in pink module, *SPL1*, *ARP8* and *SUS4* in green yellow module. Although the homologous genes of these genes in other species have rarely been reported to regulate leaf development in other species, they may be involved in leaf development in *Epimedium*. The expression levels of these genes were shown in Figure 7C.

Based on the visual network's cues, we used the *TCP* conserved domain homologous comparison method to identify potential *TCP* transcription factors from our transcript data and demonstrated their expression levels (Figure 7C). The transcription of *TRINITY_DN4142_c0_g1_i1* was found to be lower in *E. koreanum* var. *polyphylla* CS Cheng than in the typical *E. koreanum*, the decrease of *EktCP14* expression may promote the proliferation of

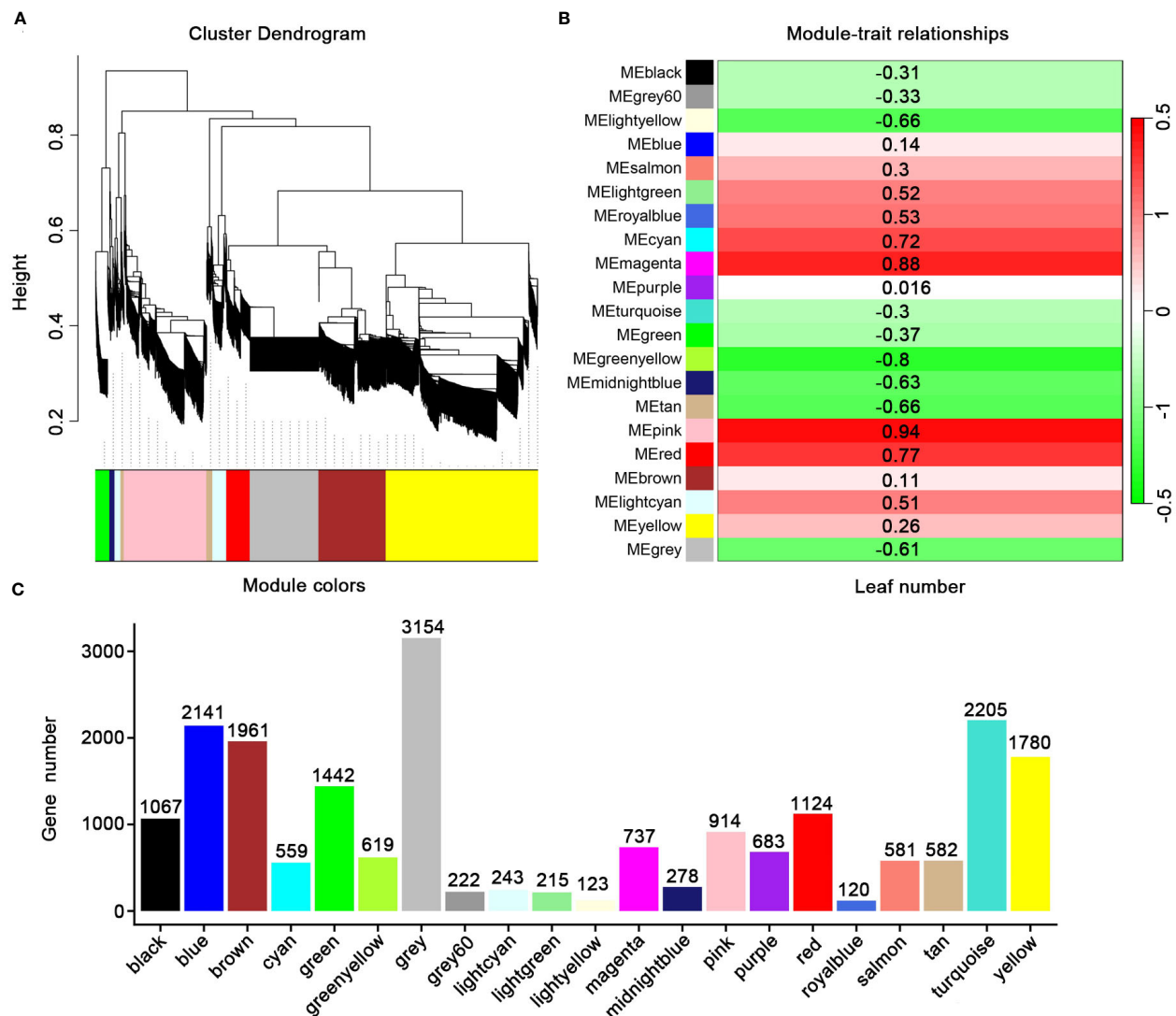


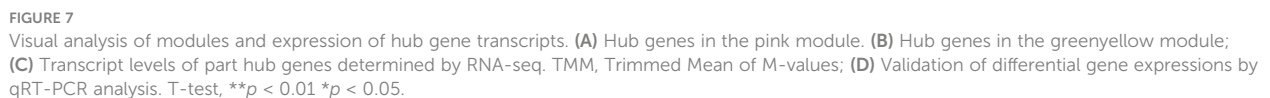
FIGURE 6

WGCNA analysis of genes related to leaf biomass in *E. koreanum*. (A) Clustering dendrogram of genes. Similarity is based on topological overlap, with module colors assigned. The 21 co-expression modules are displayed in different colors. (B) Correlation between the 21 modules. (C) Member count in the 21 modules.

cells and thus increase the number of leaves (Kieffer et al., 2011; Challa et al., 2021) (Figure 7C). Additionally, we discovered another *TCP9* similar gene, *TRINITY_DN3273_c2_g1_i2*, which is a member of the class I *TCP* family and whose transcripts were up-regulated in *E. koreanum* var. *polyphylla* CS Cheng. Class I *TCP* usually promotes cell proliferation, so its increased expression in *EKP* is also consistent with the phenotype (Li, 2015). So, these results first suggested that different *TCP* transcription factors have antagonistic effects on compound leaf development.

In this study, the *EktCP* genes and other hub genes mentioned in Figure 7D were then validated by qRT-PCR analysis. The results showed that the gene expression of *EktCP14*, *EktCP9*, *EkSERK1* and *EkACC2* was consistent with that in RNA-Seq analysis (t-test: $N=3$, $**P < 0.01$; $*P < 0.05$), but the expression of other 4 genes showed no significant difference compared to the

discovered newly variety and the normal *E. koreanum*. We also investigated the sequence differences of *EktCP* between the variety and normal *E. koreanum*, as well as the probable changes in spatial structure and evolutionary relationships. The results showed that *EktCP9* and *EktCP14* had multiple SNP sites, however, SNP sites were showed not in the *EktCP* domain, and only the spatial structure of *EktCP9* was slightly changed, so we speculated that there was no significant difference in the function of *EktCP9* and *EktCP14* (Figures S2A, B). In addition, we also carried out evolutionary analysis on the *EktCP* of berberaceae, and found that the variation of *EktCP* did not change much in evolution, were closest relatives to *Papaver somniferum* and *Coptis chinensis* (Figure S2C). Overall, the findings in this work implied that *EktCP* transcription factors may play an essential role in controlling *E. koreanum* leaf development.



Chinese medicine, with leaves gradually replacing other therapeutic portions since 2010 (Qian et al., 2023). The increased number of leaflets in this newly discovered variety significantly enhances its application value. Furthermore, the utilization of this resource on breeding and cultivation may contribute to alleviating the existing shortage of *E. koreanum* in the herb market. It is conceivable that the yield of this variety will be much higher than that of other *Epimedium* species. So, we used comparative RNA-Seq analysis to investigate the process of leaf development after discovering the novel *E. koreanum* varieties.

Currently, the discovered newly variety of *E. koreanum* has been cultivated in the Lushan Botanical Garden for one year, completing a phenological observation of the reproductive cycle. Although the newly identified germplasm still has an advantage in terms of the number of leaflets, the number of leaflets in ex situ grown plants has reduced in comparison to the original plants. The instability in the number of small leaf blades was observed, pointing to an innovative field of investigation into the molecular mechanism of complex leaf development.

4.2 Identification of hub genes regulating leaf biomass in *E. koreanum* var. *polyphylla* CS Cheng

The regulation of leaf development has been extensively studied in various plant species, including *Arabidopsis* (Koyama et al., 2017), tomato (Ori et al., 2007), and medicago (He et al., 2020). However, the complete genome sequence of *E. koreanum* is not yet available, making the functional study of its genes particularly challenging. *Epimedium* species, including *E. koreanum*, possess relatively large genomes and undergo numerous interspecific hybridizations, further complicating gene function analysis. To overcome these challenges, we employed RNA-seq, a powerful technique that allows for the analysis of gene expression and phenotypic characteristics. In this study, we performed comparative RNA-seq analysis on different varieties of *E. koreanum* and a control sample of normal *E. koreanum*. Our analysis identified 387 up-regulated genes and 784 down-regulated genes (Figure 1A). Actually, we intended to uncover essential genes with significant alterations in the hot reported genes associated with plant compound leaf production, and then confirm their function using RT-PCR with a larger sample size (Table S2). However, we found that none of the 10 hot genes mentioned in the study had significant differences between the newly resources and the typical *E. koreanum*. To gain insights into the functional implications of these differentially expressed genes, we conducted Gene Ontology (GO) enrichment analysis, which revealed the top ten most enriched gene categories (Figure 1B). These findings significantly contributed to our understanding of *E. koreanum*'s flavonoid metabolism and leaf growth processes.

Furthermore, we employed weighted gene co-expression network analysis (WGCNA) to uncover the modular structure and regulatory relationships among the identified genes. By performing network analysis, we successfully identified several hub genes that play a pivotal role in the regulation of leaf development in *Epimedium* (Figure 7A). Notably, our investigation highlighted the relevance of

TCP family genes in leaf development, a finding supported by studies in other plant species. TCP genes have been demonstrated to affect various aspects of leaf development, such as leaf shape (Palatnik et al., 2003; Qin et al., 2005; Schommer et al., 2008), size (Efroni et al., 2008; Tao et al., 2013), senescence (Danisman et al., 2013) and complexity (Koyama et al., 2010; Rubio-Somoza et al., 2014; Viola et al., 2023). Based on the integration of RNA-seq data and bioinformatics analysis, we made a significant discovery by identifying hub genes that are crucial for the regulation of leaf growth and development in *Epimedium*. These findings enhance our understanding of the molecular mechanisms underlying leaf growth and hold promise for potential applications in *Epimedium* breeding and leaf biomass improvement.

Currently, Transcriptome sequencing is the most effective approach for identifying differentially expressed genes in wild samples. It is required to investigate the spatiotemporal expression of genes involved to compound leaf development using real-time fluorescence quantification under an environmental control conditions. Specifically, absolute quantification of gene expression levels in samples with a gradient in leaflet number can assist discover the critical genes and regulatory mechanisms determining compound leaf biomass. In our validation study of differential gene expressions by using qRT-PCR analysis under an equal environmental control conditions, the results showed that the gene expression of EkTCP14, EkTCP9, EkSERK1 and EkACC2 was obvious consistent with that in RNA-Seq analysis, but the other 4 gene expressions showed no significant difference compared to the discovered newly variety and the normal *E. koreanum*. We speculate that the reason for this predicament is that direct transcriptome sequencing utilizing wild plant samples might provide some false positive findings. Therefore, the results of the qRT-PCR investigation revealed the regulatory involvement of the *EkTCP* transcription factors in the development of compound leaves.

4.3 Phytohormone regulates leaf development

Leaf development in plants is regulated by various phytohormones, including auxin and cytokinin (Navarro-Cartagena and Micol, 2023). The morphology of leaves varies across different plant species, primarily due to the distinct arrangement of leaf lobes or teeth in single leaves and leaflets in compound leaves (Hay and Tsiantis, 2006; Runions et al., 2017; Kierzkowski et al., 2019). Cytokinin plays a significant role in regulating leaf complexity in compound leaf species such as tomato and cardamine (*Cardamine hirsuta*). Additionally, in simple leaf species like *Arabidopsis thaliana*, both cytokinin and auxin co-regulate the morphogenesis of leaf margins. Transcription factors belonging to the TCP family are essential for maintaining the balance between cell proliferation and differentiation during leaf development. Class I TCP family members promote cell proliferation, while class II family members repress cell proliferation (Li, 2015). In *Arabidopsis*, Class I TCPs have been found to promote cytokinin responses (Steiner et al., 2012; Steiner et al., 2016). Notably, the up-regulation of Class II TCPs and *KNOX2* genes has been shown to repress the expression levels of *KNOX1* and *CUP-SHAPED COTYLEDON (CUC2)*, thereby enhancing

cytokinin response in *Nicotiana tabacum* BY-2 protoplasts and increasing the levels of active cytokinin (Cucinotta et al., 2018; Cucinotta et al., 2020; Challa et al., 2021).

Our findings indicate that the expression of *TCPs* in *EKP* significantly differed from that of normal *E. koreanum*. This suggested that abnormal leaf development in *EKP* might be regulated through phytohormone mediated abnormal expression of *TCPs*. However, it is important to note that we were unable to identify any genes directly involved in cytokinin regulation in this study. Although these genes may potentially exist, their specific functions remain unknown due to the lack of comprehensive genomic information. In conclusion, our study highlights the crucial roles of auxin and cytokinin in regulating leaf development. The complex interplay between phytohormones and key transcription factors, such as *TCPs*, is fundamental for determining leaf morphology and complexity across plant species. Further research is needed to unravel the precise mechanisms underlying cytokine in regulation and its direct genetic targets, which will contribute to a comprehensive understanding of leaf development in plants.

5 Conclusion

In this study, we have discovered a new class of *Epimedium* resources. Through DNA barcoding and phylogenetic tree analysis, it is believed that it belongs to the *E. koreanum* variety. This discovery not only identified a high leaf biomass of *Epimedium* and expanded the number of varieties of *Epimedium*, but also alleviated the current situation of resource shortage of *Epimedium*. Then, we analyzed the hub genes regulating the leaf biomass of *E. koreanum* through RNA-seq data. Based on the literature reports of related homologous genes and validation of differential gene expressions by qRT-PCR analysis, we suggested that *EkTCP9* and *EkTCP14* transcription factors play an important role in the leaf development of *E. koreanum*. The level of their expression was in line with expectations. We also investigated the SNP sites and possible spatial structure of these two *EkTCP*, and found that the *EkTCP* of these two *E. koreanum* has multiple SNP sites, but there was no significant effect on their spatial structure. In conclusion, our study not only found a new variety of *E. koreanum*, but also preliminarily analyzed the hub genes that regulate leaf development, providing new insights for future breeding of *Epimedium* with high leaf biomass.

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: Bioproject accession number PRJNA1008345.

Author contributions

JY: Investigation, Writing – original draft, Formal analysis, Methodology, Software. SF: Investigation, Methodology, Software, Data curation. ZX & MG: Methodology, Validation. QC:

Methodology, Validation, Investigation, Software. PG: Writing – review & editing. CC: Investigation, Writing – review & editing, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from National Natural Science Foundation of China (No.82260746). The authors gratefully acknowledge financial supports from Jiangxi Province Double Thousand Talent-Leader of Natural Science Project (jxsq2023101038), Jiangxi Province Urgently Overseas Talent Project (2022BCJ25027 & 2023BBG70014), Lushan Botanical Garden Project (2022ZWZX07 & 2023ZWZX08) and the Science and Technology Innovation Team Project in Key Areas of Jiujiang City Base and Talent Plan (S2022TJDS029).

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.2023.1290836/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Distribution of *E. perralderianum*, *E. pinnatum* and *E. koreanum*. The distribution of three species of *Epimedium* on the map. The red triangles represent different sample areas.

SUPPLEMENTARY FIGURE 2

Evolutionary relationships of TCP-containing proteins of *E. koreanum*. (A) SNP sites of *EkTCP9* and *EkTCP14*. *EkTCP14-1* and *EkTCP9-1* extracted from *E. koreanum*, *EkTCP14-2* and *EkTCP9-2* extracted from *E. koreanum* var. *polypphylla* CS Cheng. (B) The spatial structure of *TCP9* and *TCP14*. The red arrows represent the distinct areas. (C) The optimal tree with the sum of branch length = 14.79 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are in the units of the number of amino acid substitutions per site. The analysis involved 298 amino acid sequences with *TCP* domains, of which 294 were from the order Ranunculales. All positions containing gaps and missing data were eliminated.

SUPPLEMENTARY TABLE 1

Differentially expressed transcripts in variety Compared with normal *E. koreanum*.

References

- Bar, M., and Ori, N. (2015). Compound leaf development in model plant species. *Curr. Opin. Plant Biol.* 23, 61–69. doi: 10.1016/j.pbi.2014.10.007
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334
- Challa, K. R., Rath, M., Sharma, A. N., Bajpai, A. K., Davuluri, S., Acharya, K. K., et al. (2021). Active suppression of leaflet emergence as a mechanism of simple leaf development. *Nat. Plants* 7, 1264–1275. doi: 10.1038/s41477-021-00965-3
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinf. (Oxford England)* 34, i884–i890. doi: 10.1093/bioinformatics/bty560
- Cucinotta, M., Di Marzo, M., Guazzotti, A., de Folter, S., Kater, M. M., and Colombo, L. (2020). Gynoecium size and ovule number are interconnected traits that impact seed yield. *J. Exp. Bot.* 71, 2479–2489. doi: 10.1093/jxb/eraa050
- Cucinotta, M., Manrique, S., Cuesta, C., Benkova, E., Novak, O., and Colombo, L. (2018). CUP-SHAPED COTYLEDON1 (CUC1) and CUC2 regulate cytokinin homeostasis to determine ovule number in Arabidopsis. *J. Exp. Bot.* 69, 5169–5176. doi: 10.1093/jxb/ery281
- Danisman, S., van Dijk, A. D., Binbo, A., van der Wal, F., Hennig, L., de Folter, S., et al. (2013). Analysis of functional redundancies within the Arabidopsis TCP transcription factor family. *J. Exp. Bot.* 64, 5673–5685. doi: 10.1093/jxb/ert337
- Demason, D. A., Chetty, V., Barkawi, L. S., Liu, X., and Cohen, J. D. (2013). Unifoliata-Afila interactions in pea leaf morphogenesis. *Am. J. Bot.* 100, 478–495. doi: 10.3732/ajb.1200611
- Dengler, N. G., and Tsukaya, H. (2001). Leaf morphogenesis in dicotyledons: current issues. *Int. J. Plant Sci.* 162, 459–464. doi: 10.1086/320145
- Efroni, I., Blum, E., Goldshmidt, A., and Eshed, Y. (2008). A protracted and dynamic maturation schedule underlies Arabidopsis leaf development. *Plant Cell* 20, 2293–2306. doi: 10.1105/tpc.107.057521
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Guo, M., Pang, X., Xu, Y., Jiang, W., Liao, B., Yu, J., et al. (2022). Plastid genome data provide new insights into the phylogeny and evolution of the genus *Epimedium*. *J. Adv. Res.* 36, 175–185. doi: 10.1016/j.jare.2021.06.020
- Hay, A., and Tsiantis, M. (2006). The genetic basis for differences in leaf form between Arabidopsis thaliana and its wild relative Cardamine hirsuta. *Nat. Genet.* 38, 942–947. doi: 10.1038/ng1835
- He, L., Liu, Y., He, H., Liu, Y., Qi, J., Zhang, X., et al. (2020). A molecular framework underlying the compound leaf pattern of *Medicago truncatula*. *Nat. Plants* 6, 511–521. doi: 10.1038/s41477-020-0642-2
- Helliwell, E. E., Vega-Arreguin, J., Shi, Z., Bailey, B., Xiao, S., Maximova, S. N., et al. (2016). Enhanced resistance in Theobroma cacao against oomycete and fungal pathogens by secretion of phosphatidylinositol-3-phosphate-binding proteins. *Plant Biotechnol. J.* 14, 875–886. doi: 10.1111/pbi.12436
- Indran, I. R., Liang, R. L., Min, T. E., and Yong, E. L. (2016). Preclinical studies and clinical evaluation of compounds from the genus *Epimedium* for osteoporosis and bone health. *Pharmacol. Ther.* 162, 188–205. doi: 10.1016/j.pharmthera.2016.01.015
- Jiang, J., Song, J., and Jia, X. B. (2015). Phytochemistry and ethnopharmacology of *epimedium* L. Species. *Chin. Herb. Med.* 7, 204–222. doi: 10.1016/S1674-6384(15)60043-0
- Jiao, W., Sun, J., Zhang, X., An, Q., Fu, L., Xu, W., et al. (2021). Improvement of Qilin pills on male reproductive function in tripterygium glycoside-induced oligoasthenospermia in rats. *Andrologia* 53, e13923. doi: 10.1111/and.13923
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinf. (Oxford England)* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199
- Kieffer, M., Master, V., Waites, R., and Davies, B. (2011). TCP14 and TCP15 affect internode length and leaf shape in Arabidopsis. *Plant J.* 68, 147–158. doi: 10.1111/j.1365-3113.2011.04674.x
- Kierzkowski, D., Runions, A., Vuolo, F., Strauss, S., Lymbouridou, R., Routier-Kierzkowska, A. L., et al. (2019). A growth-based framework for leaf shape development and diversity. *Cell* 177, 1405–1418. doi: 10.1016/j.cell.2019.05.011
- Kim, W. J., Ji, Y., Choi, G., Kang, Y. M., Yang, S., and Moon, B. C. (2016). Molecular identification and phylogenetic analysis of important medicinal plant species in genus *Paeonia* based on rDNA-ITS, matK, and rbcL DNA barcode sequences. *Genet. Mol. Res.* 15 (3), gmr.15038472. doi: 10.4238/gmr.15038472
- Kim, H. S., Lee, B. Y., Won, E. J., Han, J., Hwang, D. S., Park, H. G., et al. (2015). Identification of xenobiotic biodegradation and metabolism-related genes in the copepod *Tigriopus japonicus* whole transcriptome analysis. *Mar. Genomics* 24 Pt 3, 207–208. doi: 10.1016/j.margen.2015.05.011
- Kim, M., McCormick, S., Timmermans, M., and Sinha, N. (2003). The expression domain of PHANTASTICA determines leaflet placement in compound leaves. *Nature* 424, 438–443. doi: 10.1038/nature01820
- Koyama, T., Mitsuda, N., Seki, M., Shinozaki, K., and Ohme-Takagi, M. (2010). TCP transcription factors regulate the activities of ASYMMETRIC LEAVES1 and miR164, as well as the auxin response, during differentiation of leaves in Arabidopsis. *Plant Cell* 22, 3574–3588. doi: 10.1105/tpc.110.075598
- Koyama, T., Sato, F., and Ohme-Takagi, M. (2017). Roles of miR319 and TCP transcription factors in leaf development. *Plant Physiol.* 175, 874–885. doi: 10.1104/pp.17.00732
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9, 559. doi: 10.1186/1471-2105-9-559
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Laura, W., Angela, T. M., and Will, E. (2010). Not so simple after all: searching for ecological advantages of compound leaves. *Oikos* 120, 813–821. doi: 10.1111/j.1600-0706.2010.19344.x
- Lee, J. H., Kim, K., Kim, N. R., Lee, S. C., Yang, T. J., and Kim, Y. D. (2016). The complete chloroplast genome of a medicinal plant *Epimedium koreanum* Nakai (Berberidaceae). *Mitochondrial DNA A DNA Mapp Seq Anal.* 27, 4342–4343. doi: 10.3109/19401736.2015.1089492
- Li, S. (2015). The Arabidopsis thaliana TCP transcription factors: A broadening horizon beyond development. *Plant Signal. Behav.* 10, e1044192. doi: 10.1080/15592324.2015.1044192
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12, 323. doi: 10.1186/1471-2105-12-323
- Li, Z., Liu, M., Zhang, L., Zhang, W., Gao, G., Zhu, Z., et al. (2009). Detection of intergenic non-coding RNAs expressed in the main developmental stages in *Drosophila melanogaster*. *Nucleic Acids Res.* 37, 4308–4314. doi: 10.1093/nar/gkp334
- Li, S. Q., Pei, Z. G., and Liu, Y. M. (2001). Clinical study on effect of gushukang granule in preventing and treating primary osteoporosis. *Zhongguo Zhong xi yi jie he zhi Zhongguo Zhongxiyi jiehe zazhi = Chin. J. integrated traditional Western Med.* 21, 265–268.
- Lin, Q., Jin, S., Zong, Y., Yu, H., Zhu, Z., Liu, G., et al. (2021). High-efficiency prime editing with optimized, paired pegRNAs in plants. *Nat. Biotechnol.* 39, 923–927. doi: 10.1038/s41587-021-00868-w
- Liu, T., Zhao, M., Zhang, Y., Qiu, Z., Zhang, Y., Zhao, C., et al. (2021). Pharmacokinetic-pharmacodynamic modeling analysis and anti-inflammatory effect of Wangbi capsule in the treatment of adjuvant-induced arthritis. *Biomed. Chromatogr.* 35, e5101. doi: 10.1002/bmc.5101
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8
- Luo, C., Wang, S., Ning, K., Chen, Z., Wang, Y., Yang, J., et al. (2021). LsAP2 regulates leaf morphology by inhibiting CIN-like TCP transcription factors and repressing LsKAN2 in lettuce. *Hortic. Res.* 8, 184. doi: 10.1038/s41438-021-00622-y
- Ma, H., He, X., Yang, Y., Li, M., Hao, D., and Jia, Z. (2011). The genus *Epimedium*: an ethnopharmacological and phytochemical review. *J. Ethnopharmacol.* 134, 519–541. doi: 10.1016/j.jep.2011.01.001
- Mishra, R. K., Chaudhary, S., Kumar, A., and Kumar, S. (2009). Effects of MULTIFOLIATE-PINNA, AFILA, TENDRIL-LESS and UNIFOLIATA genes on leafblade architecture in *Pisum sativum*. *Planta* 230, 177–190. doi: 10.1007/s00425-009-0931-5
- Navarro-Cartagena, S., and Micol, J. L. (2023). Is auxin enough? Cytokinins and margin patterning in simple leaves. *Trends Plant Sci.* 28, 54–73. doi: 10.1016/j.tplants.2022.08.019
- Ori, N., Cohen, A. R., Etzioni, A., Brand, A., Yanai, O., Shleizer, S., et al. (2007). Regulation of LANCEOLATE by miR319 is required for compound-leaf development in tomato. *Nat. Genet.* 39, 787–791. doi: 10.1038/ng2036
- Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., et al. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 425, 257–263. doi: 10.1038/nature01958

- Qian, H. Q., Wu, D. C., Li, C. Y., Liu, X. R., Han, X. K., Peng, Y., et al. (2023). A systematic review of traditional uses, phytochemistry, pharmacology and toxicity of *Epimedium koreanum* Nakai. *J. Ethnopharmacol.* 318, 116957. doi: 10.1016/j.jep.2023.116957
- Qin, G., Gu, H., Zhao, Y., Ma, Z., Shi, G., Yang, Y., et al. (2005). An indole-3-acetic acid carboxyl methyltransferase regulates *Arabidopsis* leaf development. *Plant Cell* 17, 2693–2704. doi: 10.1105/tpc.105.034959
- Ren, L., Guo, M.-y., and Pang, X.-h. (2018). Identification and classification of medicinal plants in *Epimedium*. *Chin. Herb. Med.* 10, 249–254. doi: 10.1016/j.chmed.2018.05.004
- Rubio-Somoza, I., Zhou, C. M., Confraria, A., Martinho, C., von Born, P., Baena-Gonzalez, E., et al. (2014). Temporal control of leaf complexity by miRNA-regulated licensing of protein complexes. *Curr. Biol.* 24, 2714–2719. doi: 10.1016/j.cub.2014.09.058
- Runions, A., Tsiantis, M., and Prusinkiewicz, P. (2017). A common developmental program can produce diverse leaf shapes. *New Phytol.* 216, 401–418. doi: 10.1111/nph.14449
- Schommer, C., Palatnik, J. F., Aggarwal, P., Chételat, A., Cubas, P., Farmer, E. E., et al. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. *PloS Biol.* 6, e230. doi: 10.1371/journal.pbio.0060230
- Steiner, E., Efroni, I., Gopalraj, M., Saathoff, K., Tseng, T. S., Kieffer, M., et al. (2012). The *Arabidopsis* O-linked N-acetylglucosamine transferase SPINDLY interacts with class I TCPs to facilitate cytokinin responses in leaves and flowers. *Plant Cell* 24, 96–108. doi: 10.1105/tpc.111.093518
- Steiner, E., Livne, S., Kobinson-Katz, T., Tal, L., Pri-Tal, O., Mosquna, A., et al. (2016). The putative O-linked N-acetylglucosamine transferase SPINDLY inhibits class I TCP proteolysis to promote sensitivity to cytokinin. *Plant Physiol.* 171, 1485–1494. doi: 10.1104/pp.16.00343
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197
- Tao, Q., Guo, D., Wei, B., Zhang, F., Pang, C., Jiang, H., et al. (2013). The TIE1 transcriptional repressor links TCP transcription factors with TOPLESS/TOPLESS-RELATED corepressors and modulates leaf development in *Arabidopsis*. *Plant Cell* 25, 421–437. doi: 10.1105/tpc.113.109223
- Urano, K., Maruyama, K., Koyama, T., Gonzalez, N., Inzé, D., Yamaguchi-Shinozaki, K., et al. (2022). CIN-like TCP13 is essential for plant growth regulation under dehydration stress. *Plant Mol. Biol.* 108, 257–275. doi: 10.1007/s11103-021-01238-5
- Vijayan, K., and Tsou, C.H.J.C.S. (2010). DNA barcoding in plants: taxonomy in a new perspective. *Curr. Sci.* 99, 1530–1541.
- Viola, I. L., Alem, A. L., Jure, R. M., and Gonzalez, D. H. (2023). Physiological roles and mechanisms of action of class I TCP transcription factors. *Int. J. Mol. Sci.* 24 (6), 5437. doi: 10.3390/ijms24065437
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi: 10.1016/j.xinn.2021.100141
- Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., et al. (2016). The complete chloroplast genome sequences of five *epimedium* species: lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7, 306. doi: 10.3389/fpls.2016.00306
- Zhang, Y., Li, J., Wang, Y., and Liang, Q. (2022). Taxonomy of *Epimedium* (Berberidaceae) with special reference to Chinese species. *Chin. Herb. Med.* 14, 20–35. doi: 10.1016/j.chmed.2021.12.001
- Zhang, M.-L., Uhink, C., and Kadereit, J. (2007). Phylogeny and biogeography of *epimedium/vancouveria* (Berberidaceae): western north american - east asian disjunctions, the origin of european mountain plant taxa, and east asian species diversity. *Syst. Bot.* 32, 81–92. doi: 10.1600/036364407780360265
- Zhang, H., Wu, X., Wang, J., Wang, M., Wang, X., Shen, T., et al. (2020). Flavonoids from the leaves of *Epimedium Koreanum* Nakai and their potential cytotoxic activities. *Nat. Prod. Res.* 34, 1256–1263. doi: 10.1080/14786419.2018.1560283
- Zhong, Q., Shi, Z., Zhang, L., Zhong, R., Xia, Z., Wang, J., et al. (2017). The potential of *Epimedium koreanum* Nakai for herb-drug interaction. *J. Pharm. Pharmacol.* 69, 1398–1408. doi: 10.1111/jphp.12773
- Zhu, H., Shi, Y., Jiang, S., Jiao, X., Zhu, H., Wang, R., et al. (2021). Investigation of the mechanisms of chuankezhi injection in the treatment of asthma based on the network pharmacology approach. *Evid. Based Complement. Altern. Med.* 2021, 5517041. doi: 10.1155/2021/5517041



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Soraya Mousavi,
National Research Council (CNR), Italy
Axel Diederichsen,
Agriculture and Agri-Food Canada (AAFC),
Canada

*CORRESPONDENCE

Francisco Jesús Gómez-Gálvez
✉ franciscoj.gomez.galvez@
juntadeandalucia.es

RECEIVED 26 July 2023

ACCEPTED 13 December 2023

PUBLISHED 05 January 2024

CITATION

Gómez-Gálvez FJ, Ninot A, Rodríguez JC,
Compañ SP, Andreva JU, Rubio JAG,
Aragón IP, Viñuales-Andreu J,
Casanova-Gascón J, Šatović Z, Lorite IJ,
De la Rosa-Navarro R and Belaj A (2024)
New insights in the Spanish gene pool of
olive (*Olea europaea* L.) preserved *ex situ*
and *in situ* based on high-throughput
molecular markers.
Front. Plant Sci. 14:1267601.
doi: 10.3389/fpls.2023.1267601

COPYRIGHT

© 2024 Gómez-Gálvez, Ninot, Rodríguez,
Compañ, Andreva, Rubio, Aragón,
Viñuales-Andreu, Casanova-Gascón, Šatović,
Lorite, De la Rosa-Navarro and Belaj. 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.

New insights in the Spanish gene pool of olive (*Olea europaea* L.) preserved *ex situ* and *in situ* based on high-throughput molecular markers

Francisco Jesús Gómez-Gálvez^{1*}, Antònia Ninot²,
Juan Cano Rodríguez³, Sergio Paz Compañ⁴,
Javier Ugarte Andreva⁵, Javier Alfonso García Rubio⁵,
Isis Pinilla Aragón⁵, Javier Viñuales-Andreu⁶,
José Casanova-Gascón⁶, Zlatko Šatović^{7,8},
Ignacio Jesús Lorite¹, Raúl De la Rosa-Navarro^{1,9}
and Angelina Belaj¹

¹Mejora Vegetal y Biotecnología, Instituto Andaluz de Investigación y Formación Agraria,
Pesquera, Alimentaria y de la Producción Ecológica (IFAPA), Centro Alameda del Obispo,
Córdoba, Spain, ²Fruticultura, Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Mas Bové,
Constantí, Tarragona, Spain, ³Ingeniería y Tecnología Agroalimentaria, Instituto Andaluz de
Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica (IFAPA),
Centro Venta del Llano, Mengibar, Jaén, Spain, ⁴Olivicultura, Instituto Valenciano de
Investigaciones Agrarias (IVIA), Moncada, Valencia, Spain, ⁵Servicio de Investigación Agraria y
Sanidad Vegetal, Gobierno de La Rioja, Logroño, Spain, ⁶Universidad de Zaragoza, Escuela
Politécnica Superior de Huesca, Aragón, Spain, ⁷Department of Plant Biodiversity, Faculty of
Agriculture, University of Zagreb, Zagreb, Croatia, ⁸Centre of Excellence for Biodiversity and
Molecular Plant Breeding (CroP-BioDiv), Zagreb, Croatia, ⁹Department of Plant Breeding,
Institute for Sustainable Agriculture, Spanish National Research Council (IAS-CSIC),
Cordoba, Spain

In Spain, several local studies have highlighted the likely presence of unknown olive cultivars distinct from the approximately 260 ones previously described in the literature. Furthermore, recent advancements in identification techniques have significantly enhanced in terms of efficacy and precision. This scenario motivated a new nationwide prospecting effort aimed at recovering and characterizing new cultivated germplasm using high-throughput molecular markers. In the present study, the use of 96 EST-SNP markers allowed the identification of a considerable amount of new material (173 new genotypes) coming from areas with low intensification of production in different regions of Spain. As a result, the number of distinct national genotypes documented in the World Olive Germplasm Bank of IFAPA, Córdoba (WOGBC-ESP046) increased to 427. Likewise, 65 and 24 new synonymy and homonymy cases were identified, respectively. This rise in the number of different national cultivars allowed to deepen the knowledge about the underlying genetic structure. The great genetic variability of Spanish germplasm was confirmed, and a new hot spot of diversity was identified in the northern regions of La Rioja and Aragon. Analysis of the genetic structure showed a clear separation between the

germplasm of southern and northern-northeastern Spain and indicated a significantly higher level of admixture in the latter. Given the expansion of modern olive cultivation with only a few cultivars, this cryptic germplasm is in great danger of disappearing. This underlines the fact that maintaining as many cultivars as possible will increase the genetic variability of the olive gene pool to meet the future challenges of olive cultivation.

KEYWORDS

genetic characterization, *Olea europaea*, EST-SNPs, cultivars collection, conservation

1 Introduction

The olive tree (*Olea europaea* L.) is one of the quintessential emblems of Spain, constituting a crucial element in the economic, social and environmental framework of the country. Its cultivation in Spain could date back to the Bronze Age (Terral and Arnold-Simard, 1996; Terral et al., 2004), although Phoenicians, Greeks and especially Romans and Muslims were the main responsible for its expansion and cultivation by importing know-how and plant material from eastern Mediterranean (Kaniewski et al., 2012; Besnard et al., 2018). Thus, most of the cultivated genotypes of olive tree in Spain have their origin in the empirical selection made by farmers over the centuries, being most of them very old and confined to its presumed area of selection (Barranco, 2010). Besides, the number of cultivars obtained by olive breeding is very small compared to other fruit trees.

Nowadays, Spain is by far the country with the largest number of olive trees planted in the world and represents the leading olive-producing country with more than 20% of the world area and around 34% of the worldwide production (FAOSTAT, 2020). One of the reasons that have led to the achievement of this leading position is the modernization and technification of olive farming carried out during the last decades (Fernandez-Escobar et al., 2013). These changes have been linked to a certain reconversion in the cultivar landscape. A broad range of old, local, and traditional cultivars are being displaced by few cultivars that are well known for their desirable traits in terms of earliness of bearing, oil content and quality, and suitability for new harvesting and pruning techniques (de la Rosa et al., 2007). As an example, Spanish olive-growing areas are dominated by only three cultivars: ‘Arbequina’, ‘Picual’ and ‘Hojiblanca’ (Belaj et al., 2016). Moreover, these few cultivars represent the main source of supply for most of the national breeding programs, thus favouring even more a possible genetic erosion of the crop (Rallo et al., 2013; León et al., 2021; Yilmaz-Duzyaman et al., 2022). This genetic erosion poses a risk to the legacy of diversity built over generations of olive growers and compromises the added value of exclusivity provided by local cultivars in olive products. Moreover, this loss of diversity weakens the availability of a potentially valuable strategic reserve for breeders that could help for dealing with future challenges such

as temperature increase (potentially leading to a lack of chilling requirements for flowering and/or heat stress), water stress, salinity, emerging pests and diseases, farming in new edaphoclimatic areas, and new market trends such as the search for specific quality characters in EVOO (Pérez et al., 2019; Medina-Alonso et al., 2020; Serrano et al., 2020; Lorite et al., 2022). For these reasons, the recovery, conservation and study of minor cultivars is of increasing interest in Spain and elsewhere (Díez et al., 2011; Hmam et al., 2018; Ninot et al., 2018; Debbabi et al., 2020; Valeri et al., 2022; Marchese et al., 2023).

To address this need, the World Olive Germplasm Bank of Córdoba (WOGBC) plays a crucial role by conserving as much olive genetic patrimony as possible. WOGBC is established at the experimental field “Alameda del Obispo” of the Andalusian Institute for Research and Training in Agriculture, Fishery, Food and Organic Production (IFAPA), and represents a reference olive germplasm bank both at national (INIA-ESP046) and international level (IOC) (Belaj et al., 2016; Díez et al., 2018; Belaj et al., 2022). The accurate identification of olive material is a crucial task in germplasm banks (Atienza et al., 2013; Trujillo et al., 2014; El Bakkali et al., 2019). Historically, different morphological and molecular markers, especially simple sequence repeats (SSRs), have been used at WOGBC (Barranco et al., 2000; Belaj et al., 2012; Atienza et al., 2013; Trujillo et al., 2014). Currently, cultivar identification in WOGBC is performed by means of EST-SNP markers (Single-Nucleotide Polymorphism from Expressed Sequence Tags). EST-SNP markers have shown clear advantages over previously used markers: fully automation in high-throughput assays, cost-effective, lower genotyping error rates, and higher reproducibility across different laboratories, germplasm collections, and genotyping platforms (Belaj et al., 2018). In a recent research aimed at improving the management and use of the genetic resources maintained at WOGBC, a core set of 96 EST-SNP markers was evaluated for the fingerprinting of 1273 accessions from 29 countries. It allowed the accurate identification of the highest number of olive genotypes (668) up to date, that are currently maintained at the WOGBC collection. Among them, 38% belonged to Spanish cultivars (Belaj et al., 2022). Most of these Spanish cultivars were incorporated to the collection

thanks to the prospecting surveys conducted at the end of the past century (Barranco, 2010). Since then, several identification studies that explored different areas of the country at local level have shown evidences of uncatalogued cultivars that remained to be recover and preserve (Íñiguez et al., 2001; Viñuales-Andreu, 2007; Díez et al., 2011; Fernández i Martí et al., 2015; Ninot et al., 2018). These findings indicate that the real number of Spanish olive cultivars is still underestimated and point out the importance of continuous and systematic prospecting surveys.

Starting from this scenario, in which a precisely identified set of national material is available and a management protocol has been fairly refined, the enrichment of the WOGBC with national unknown cultivars has been seen as a must in the last years. Thus, a new wave of prospecting and collecting surveys on Spanish territory was deployed to recover this local untapped diversity. The present research is part of an ongoing project aimed at enriching WOGBC by introducing new local Spanish cultivated germplasm followed by its genetic and agronomic characterization. In particular, this work addresses: i) the search for and recovery of non-catalogued cultivars, ii) their fingerprinting and identification by means of 96 EST-SNP markers, and iii) the assessment of their genetic diversity and structure.

2 Material and methods

2.1 Plant material

The collection of plant material under study was made in two principal ways: 1) through collaborations and incorporation of plant accessions from regional collections such as the Institut de Recerca i Tecnologia Agroalimentaria, IRTA, (Catalonia,

Northeastern Spain), and the Servicio de Investigación Agraria y Sanidad Vegetal (La Rioja, Northern Spain); and 2) through ongoing local prospecting surveys conducted mainly in Aragon, La Rioja and Catalonia (Northern, Northeastern Spain) and Andalusia (Southern Spain) regions and at a minor scale in other regions of the country (Table 1, Figure 1, Supplementary Table 1). In the case of Andalusia, special focus was placed in prospecting uncatalogued cultivars previously identified by Díez et al. (2011) from monumental and centennial trees (trunk diameter ranging 1–2.72 m) as well as other minor local cultivars that escaped from previous surveys; as for Aragon and Catalonia, most prospecting material came from the mountainous system of Pre-Pyrenees region. In order to reduce redundancies, sampling collection was implemented following the strategy defined in Belaj et al. (2022). In this sense, as much characterization data as possible was collected to be included in the passport data (fruit and stone size and shape, and any other relevant agronomical information), and an *a priori* identification by means molecular markers was conducted before their introduction into WOGBC collection. For this identification, the genetic profiles of all the samples included in this study were compared among them and with the WOGBC database (see Data analysis section below). The *in situ* morphological and agronomical information of sampled trees served to ensure their cultivated status. The *a priori* identification was made from shoot samples collected from olive trees conserved *in situ*, which were georeferenced. However, when sampling involved vulnerable trees at risk of disappearance and/or growing in very remote areas with difficult access, they were vegetatively propagated, and further incorporated as new accessions at different propagation facilities of WOGBC. Thus, *a priori* identification was conducted in a total of 538 DNA samples obtained from plant shoots of olive trees surveyed in different sites *in situ* (Table 1, Figure 1, Supplementary

TABLE 1 Number of samples and accessions genotyped per region and number of different cultivars identified.

Type of plant material	Sampling sites/origin	Number of samples/accessions	Genotypes identified	New Genotypes*
DNA from plant shoots (collected <i>in situ</i>)	Andalusia	65	43	25
	Aragon	31	20	7
	Catalonia	79	68	34
	La Rioja	331	76	35
	Other Regions	32	27	14
	TOTAL	538	234 (215**)	115
WOGBC accessions in propagation facilities (collected <i>ex situ</i>)	Andalusia	81	54	32
	Aragon	1	1	1
	Catalonia	15	15	15
	La Rioja	9	9	9
	Galicia	1	1	1
	TOTAL	107	80	58

*New genotypes whose profile did not match that of any other sample of this study or any of the profiles included in the WOGBC database.

**Total genotypes omitting redundancies between regions: 15 genotypes were identified in two different regions, and two of them ('Picual' and 'Manzanilla de Sevilla') were identified in three different regions (i.e., 19 redundancies).

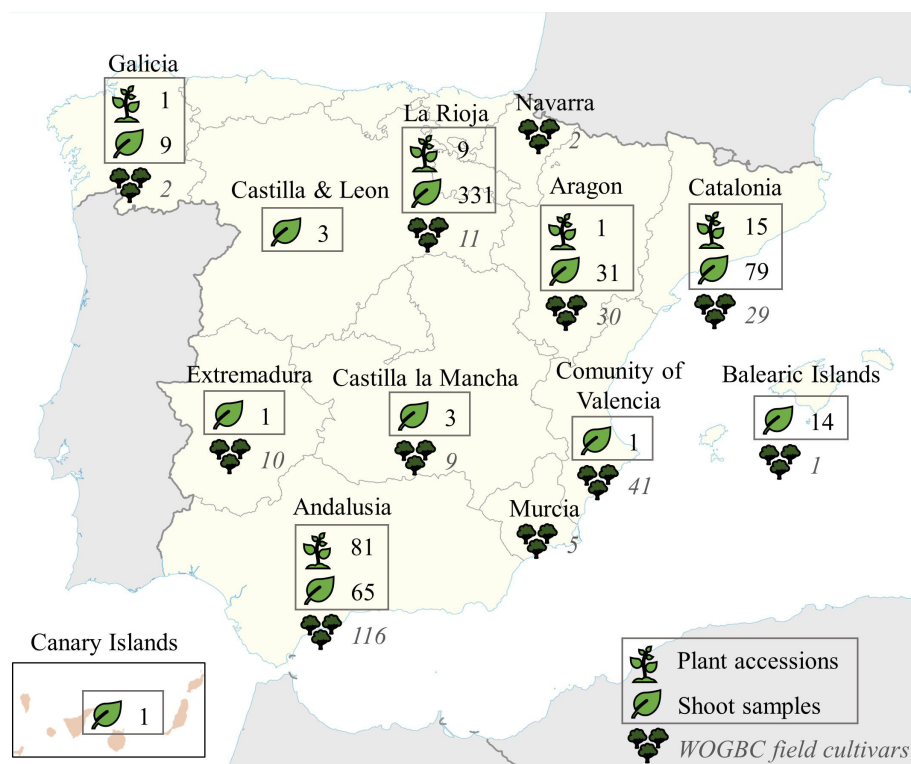


FIGURE 1

Geographical distribution of the new plant material acquired recently and subjected to identification in the present study (in boxes), and geographical origin of the national cultivars already conserved *ex situ* in the WOGBC field.

Table 1) as well as in 107 new accessions recently incorporated to the propagation facilities of the WOGBC for their *ex situ* conservation (each accession composed of 1 to 4 olive plants).

2.2 EST-SNP genotyping

DNA was extracted from fresh leaves and sprouts following de la Rosa et al. (2002) and quantitatively and qualitatively evaluated using spectrophotometry (Nanodrop 2000, Thermo Scientific, Wilmington, DE). Genotyping was conducted in the Sequencing and Genotyping Unit of the University of the Basque Country, by means of a core set of 96 EST-SNPs and following the Fluidigm method as in Belaj et al. (2022). Briefly, two preamplification primers (Locus-Specific Primer (LSP) and Specific Target Amplification (STA) primer) amplified the target region containing the SNP to be genotyped. All 96 SNPs were preamplified simultaneously in one multiplex PCR, for each sample separately, on a Veriti Thermal Cycler (Applied Biosystems by ThermoFisher, Waltham, MA, USA). Afterwards, an additional PCR amplified a portion of the target SNP region, using the LSP and two fluorescently labelled allele-specific internal primers ASP1 and ASP2, containing either the first or the second allele, respectively. The second PCR was performed on a Fluidigm 96.96 Dynamic Array IFC (Integrated Fluidic Circuit), where

reactions were performed in separate nano-wells for each SNP and sample combination, allowing simultaneous genotyping of 94 samples (+2 negative test controls - NTCs) at 96 SNP loci. This PCR was performed on a BioMark HD System (Fluidigm, South San Francisco, CA, USA). Finally, SNP genotypes were then determined by measuring the fluorescence intensity of both alleles normalised with respect to NTCs values, using SNP Genotyping Analysis Software (Fluidigm, South San Francisco, CA, USA). Two reference cultivars ('Picual' and 'Frantoio') were included in all PCR reactions as controls. Only samples with less than eight EST-SNP missing data were included for further analysis.

2.3 Data analysis

For the identification analysis of the plant material under study, the genetic profiles of all the samples included in this study were compared among them and with the WOGBC database that includes 668 different cultivars from around the world, 254 of them Spanish (Belaj et al., 2022). In addition, EST-SNP genotyping data were checked with passport data as well as *in situ* morphological and agronomical data from prospecting and collecting sites. Both the shoot samples and the new accessions maintained at different propagation facilities were considered as redundant or duplicates when they shared the same EST-SNP

profiles within them and/or with WOGBC cultivars. For each redundancy group, a representative profile was selected, and the rest of redundant samples were excluded for further analysis.

For further diversity and population structure analysis, the genetic profiles of all the non-redundant Spanish genotypes identified were considered, i.e. including both the different genotypes newly identified together with the rest of national cultivars previously identified and maintained at the WOGBC (254 different genotypes; Belaj et al. (2022)) (Figure 1; Supplementary Table 1).

Key genetic parameters were calculated for the set of 96 EST-SNPs genotyped in the whole set of non-redundant Spanish genotypes, as well as for *a priori* groups defined according to their regional origin. The genetic differentiation between groups was calculated by AMOVA and fixation index (F_{ST}) GenAlex 6.5 (Peakall and Smouse, 2012) and Cervus (Marshall et al., 1998) software were used to calculate the diversity parameters.

Population structure was first explored using the Bayesian-based approach implemented in the software package STRUCTURE v.2.2.4 (Pritchard et al., 2000) following the settings described in Belaj et al. (2022). The non-redundant Spanish genotypes were assigned to a specific cluster if their value of the corresponding Q-value (i.e., proportion of membership) were higher than 0.80, otherwise they were considered mosaic/admixed.

Discriminant Analysis of Principal Components (DAPC) was implemented as complementary clustering method to further explore the pattern of population structure (Jombart et al., 2010). In contrast to STRUCTURE, the DAPC is a multivariate method that uses a non-hierarchical approach for defining genetic clusters. The DAPC was performed in R 4.2.2 (R Core Team, 2021) using wrapper functions of the R package SambaR (de Jong et al., 2021) (<https://github.com/menodejong1986/SambaR>). Since a preselected set of 96 markers was used (Belaj et al., 2022) and data were curated right after genotyping (genotypes with less than 8 missing data retained), the 'filterdata' function, mandatory when following the SambaR workflow, was run with settings that allowed to retain all EST-SNPs and samples under study. Functions 'find.clusters' and 'dapc' were run inside the pipeline of SambaR, and the number of principal components, clusters, and discriminant functions were considered according to Jombart and Collins (2015) tutorial.

3 Results

3.1 Genotyping

The EST-SNP genotyping revealed a considerable level of redundant germplasm within the plant material prospected and collected in the present study. About 60% of the 538 DNA samples taken from olive tree shoots conserved *in situ* shared the same EST-SNP profile with at least another collected sample, being identified 215 different genotypes among them (Table 1). In the case of the 107 olive accessions maintained *ex situ*, only ~25% showed redundancy among them, with 80 different genotypes identified. Additional

redundant germplasm was found when expanding the comparison with the WOGBC EST-SNP database (668 different cultivars, 254 of them Spanish). Thus, 46.5% of the 215 different genotypes identified *in situ*, and 27.5% of the 80 different genotypes identified *ex situ* were found to be redundant with cultivars already maintained at WOGBC.

It is worth mentioning that all the redundancies were detected among cultivars previously identified in the Spanish territory, although some of them also presented synonymies with cultivars from other Mediterranean countries. The only exception was found in the sample 'Olivo de Mallorca-2', identified *in situ*, that matched with the French cultivar 'Aglandau'.

The EST-SNP genotyping of the plant material under study, enabled the identification of new synonymy cases (identical fingerprints but different naming). Thus, a total of 65 new synonymy cases belonging to 39 different genotypes were identified. Most of the new synonymies belonged to cultivars already found in the WOGBC collection. However, seven new synonymy groups were identified for the first time in the present study and were detected among the new plant material included through collection or prospecting surveys (Table 2).

Despite the high number of redundant accessions found in this study, a high number of new cultivars was identified. Thus, more

TABLE 2 New synonymy cases detected in the present study.

New synonymies detected	Representative cultivar*
Verdial	Acebuche de Autol
Manzanilla Castúa, Redondilla de Cuevas del Becerro	Alameño de Cabra
Arnellidero, Bermejuela, Serranilla	Arroniz
Aceituno, Calahorrana	Bodoquera
Picudillo, Silvestre	Bolvino
Vidrial de Cuevas	Buidiego
Cabacena	Llei del Bessó
Cerruda de Artasona	Cerruda de Olvena
Grossal del Pallars, Grossal de Cadaqués	Safrawi (SYR); syn of spanish cv Cirujal
Negral	Cirujal de Préjano
Llangueta	Corbella
Olivo de los Pozos	Corralones de Andujar
Alcarreño	Changlot Real
Solimar	Farga
Millarenca	Gorda Limocillo
Aceituno, Pla	Gordal Sevillana
Campiñesa, Coloraillo de Cortijo Nuevo	Hojiblanca
Minuera	Lechín de Granada

(Continued)

TABLE 2 Continued

New synonymies detected	Representative cultivar*
Plans	Lloma
Casta Cortijuelos	Loaime
Manzanal de Ráfales	Mançanal d'Arnes
Santa Lucia	Manzanella del Mezquin
Menya de Vila-rodona	Menya
Verdial de Setenil	Morona
Marons	Morona de Castellon
Bonany, Carrasqueny, Torres Mil-leni	Morruda de Segorbe
Mas de Bot	Morrut
Aceitunero, Pardo, Poncho, Redondilla, Sevillano	Negral de Préjano
Corraleña	Negrillo de Arjona
Acebuche de la Hoya, Olivo de Vilares	Olivo de Los Prados
Sevilli	Palomar
Aceitunero, Vidrial, Cirujal	Picalaceña de Cornago
Aceitunero, Negrillas	Picudillo
Picalaceña, Navarrillo, Racimuda, Coloradillo	Redondilla de Logroño
Rojal de Cabacés	Verdal de Bovera
Carrasqueña, Machona, Macho, Machazo, Pardo, Tempranillo	Royal de Calatayud
Desmayo	Verdal d'Arnes
Vera del Vallès	Verdal de Manresa
Casta Dilareña	Zorzaleño de Granada

*Representative cultivars are selected according to historical identification and passport data at WOGBC collection. The new synonymy groups identified in this study for the first time are indicated in bold, with representative cultivars chosen according to higher occurrence or relevant information obtained when collecting.

than half of the genotypes identified within the DNA samples (115 out of 215) as well as more than 70% of genotypes identified within the new accessions (58 out of 80) were found to be different to any of the WOGBC cultivars. In spite of their local distribution, most of these new genotypes have been identified in more than one collecting site or in different trees within the same orchard, thus evidencing a conscious vegetative propagation by farmers, that is considered a hallmark of cultivated olive germplasm. Besides, unique genotypes (i.e. identified only at one collecting site and/or tree) with passport data evidencing likely cultivated status (large fruits and stones, rough stone surfaces, high productivity, regular planting density, etc.) could be cultivars for which evidences of clonal propagation are yet to be found. For instance, a centennial tree sampled in 2016 in Canary Islands shared the same EST-SNP-genotype with another tree prospected 5 years later in an abandoned olive orchard in the south-east of Andalusia. Overall, a total of 173 new and distinct genotypes have been identified in the present study and will be progressively incorporated to the collection. This increases up to 427 the number of national

Spanish genotypes identified to date (Supplementary Table 1). When considering the prospecting areas, the regions of Andalusia, Catalonia and La Rioja were the ones where the highest number of new local cultivars were identified (57, 49 and 44, respectively). In the case of Andalusia, 33 of the new cultivars were obtained from a total of 65 centennial trees surveyed (data not shown). Finally, up to 24 homonymy groups (i.e., a common denomination referring to different cultivars) were detected, being five of them reported for the first time in the present work: “Casta/Castizo”, “Colorado/Coloradillo”, “Llei”, “Cerruda”, and “Vidrial”. The denomination based on the greenish colour of fruits (“Verde”, “Verdal”, “Verdial”, etc) was found to include 16 different cultivars (Supplementary Table 2).

3.2 Genetic diversity analysis

The 96 EST-SNP markers showed a relatively wide diversity in the 427 distinct Spanish genotypes under study (Supplementary Table 3). Minor allele frequency (MAF) values ranged from 0.192 to 0.498, with an average value of 0.376, and the proportion of markers with MAF <0.3 and >0.3 accounted for 20.8% and 79.2%, respectively. Shannon’s information index (I) values ranged from 0.49 to 0.69, with the mean value of 0.65. The observed heterozygosity (H_O) values ranged from 0.30 to 0.79, averaging 0.53, whereas the mean expected heterozygosity (H_E) was 0.46, ranging from 0.31 to 0.50. All but five EST-SNPs showed polymorphic information content (PIC) values over 0.30.

The genetic diversity was also estimated for groups defined *a priori* according to their regional origin or sampling sites (Table 3). The H_O and H_E ranged from 0.45 to 0.59, and from 0.41 to 0.46, respectively, depending on the region of origin. The group of genotypes from Balearic Islands showed more similarity between H_O and H_E , reporting therefore the highest fixation index (Table 3). According to the one-way AMOVA, the region of origin explained a low percentage of variance (4%), although ϕ_{ST} values among regions were significant ($p \leq 0.001$; Table 4). The EST-SNP pairwise differentiation among Spanish olive cultivars at regional level showed that the ones from northern regions of Aragon and La Rioja were the most genetically differentiated from the rest (Table 5); the highest differentiation values were observed between genotypes from Aragon and Extremadura. Interestingly, the Andalusian genotypes showed high similarities with those sampled in various regions such as Murcia, Extremadura, Castilla La Mancha, Galicia and Balearic Islands.

3.3 Population structure

The highest ΔK value (332,23) detected by STRUCTURE software was for $K = 3$, while the second-best solution was $K = 2$ ($\Delta K = 266.52$) (Supplementary Figure 1). The proportion of membership of each individual in each gene cluster was calculated (Supplementary Table 1; Supplementary Figure 2). At $K = 3$, the cluster A was predominant mostly in Northern (La Rioja) and North-eastern (Aragon and Catalonia) accessions, being most

TABLE 3 Summary of genetic diversity parameters estimated for the 427 different Spanish-genotypes grouped by region of origin*.

		N	Na	Ne	I	H _O	H _E	uH _E	F
Origin	Andalusia	173	2.00	1.84	0.64	0.56	0.45	0.45	-0.24
	Aragon	39	1.99	1.73	0.59	0.48	0.41	0.41	-0.16
	Balearic Islands	11	2.00	1.87	0.65	0.45	0.46	0.48	0.03
	Catalonia	77	2.00	1.83	0.64	0.49	0.45	0.45	-0.10
	Castilla La Mancha	11	1.98	1.76	0.60	0.58	0.42	0.44	-0.37
	Community of Valencia	41	2.00	1.85	0.65	0.52	0.45	0.46	-0.13
	Extremadura	10	1.98	1.79	0.62	0.59	0.43	0.45	-0.35
	Galicia	5	1.93	1.73	0.57	0.51	0.40	0.44	-0.25
	La Rioja	55	2.00	1.72	0.59	0.51	0.40	0.41	-0.25
	Murcia	5	2.00	1.77	0.61	0.57	0.42	0.47	-0.33
	Total	427	2.00	1.85	0.65	0.53	0.46	0.46	-0.15

*N, Number of distinct genotypes; Na, Average number of observed alleles; Ne, Number of effective alleles; I, Shannon's Information Index; H_O, Observed Heterozygosity; H_E, Expected Heterozygosity; uH_E, Unbiased Expected Heterozygosity; F, Fixation Index.

of them assigned with high membership values. Thus, 87 accessions were assigned to this cluster with $Q \geq 0.8$, and among them, the cultivar 'Negral de Bierge', displayed the highest values of membership (98.5%). The cluster B was found in some accessions from Catalonia (North-East), Andalusia (South), and the Eastern regions of Valencia and Balearic Islands, being the cluster with the lowest number of genotypes with $Q \geq 0.8$. The third cluster, C, was mainly represented by olive genotypes from Andalusia; i.e., 121 out of 155 genotypes with $Q \geq 0.8$ were Andalusian. Besides, olive accessions from the regions of Galicia (North-west), Extremadura (West), Castilla la Mancha (Center) and Murcia (South-east), were also assigned to this cluster. In general, very high values of membership were found for the accessions assigned to this cluster, being the well-known cultivar 'Hojiblanca' its highest representative ($Q = 0.97$). Finally, a large number of genotypes ($n = 174$; about 41% of the total) showed membership values lower than $Q \geq 0.8$ in any of the three clusters. Within the set of genotypes newly identified in this study, 53 genotypes were assigned to cluster A, 10 to cluster B, 35 to cluster C, and 75 remained as intermixed genotypes with $Q < 0.80$. The regions that showed a higher prevalence of admixture were Catalonia and, especially, the Community of Valencia, with 75% and 82% of their genotypes showing intermix.

STRUCTURE analysis revealed a certain geographic clustering of the Spanish olive accessions under study (Supplementary Figure 2, Supplementary Table 1). Such tendency could be seen

clearly when the membership coefficient inferred for each cluster was averaged by region of origin (Figure 2). In this sense, a separation of Northern and North-eastern accessions (mainly assigned to cluster A) from the Southern, South-eastern, Central, Western and North-western accessions (mainly assigned to cluster C) could be discerned. Likewise, the Community of Valencia and the Balearic Islands, positioned in the intermediate zone, were the regions that showed more admixture, with a proportional distribution in the 3 clusters.

The DAPC analysis performed without prior information on the accessions, identified five most likely genetic groups as indicated by BIC value (Supplementary Figure 3A). The first four Linear Discriminants functions and the first 80 Principal Components were retained for the analysis, representing more than 90% of total variability (Supplementary Figure 3B, C). Among the five groups identified, Group 1 included mainly genotypes from North-eastern (Catalonia, 44%) and Eastern (Valencia, 30.5%) regions; Group 2 comprised most of the genotypes sampled in Northern and North-eastern Spain, mainly in La Rioja (55%), Aragon (20.5%) and Catalonia (13%). Group 3 was represented by the lowest number of genotypes (38) and comprised mainly Andalusian (47.4%) and Catalanian (26.3%) genotypes. Group 4 consisted mainly of genotypes from Catalonia (42%) and Aragon (29%), while group 5 comprised mainly genotypes coming from the southern region of Andalusia (87%) (Figure 3; Supplementary Table 1). Thus, in total agreement with STRUCTURE, DAPC analysis revealed a clear differentiation between the genotypes mainly assigned to cluster A

TABLE 4 Analysis of molecular variance for 427 Spanish olive genotypes grouped by region of origin*.

Source	df	SS	MS	Est. Var.	%	ϕ_{ST}	$P(\phi)$
Among Pops defined by origin	9	787	87.4	0.960	4%	0.043	0.001
Within Indiv	427	10756	25.2	25.2	96%		

*df, degree of freedom, SS, sum of squares, MS, mean squares, Est. var., estimate of variance, %, Percentage of total variation, ϕ_{ST} , Phi statistic, $P(\phi)$ ϕ_{ST} probability level after 999 permutations.

TABLE 5 EST-SNP pairwise differentiation among Spanish olive cultivars at regional level*.

	Andalusia	Aragon	Balearic Islands	Catalonia	Castilla La Mancha	Community of Valencia	Extremadura	Galicia	La Rioja	Murcia
Andalusia	–	0.001	0.069	0.001	0.235	0.002	0.281	0.240	0.001	0.348
Aragon	0.071	–	0.006	0.007	0.004	0.001	0.001	0.012	0.027	0.038
Balearic Islands	0.022	0.050	–	0.014	0.033	0.010	0.024	0.022	0.008	0.388
Catalonia	0.051	0.020	0.091	–	0.003	0.006	0.002	0.028	0.001	0.123
Castilla La Mancha	0.005	0.076	0.073	0.066	–	0.023	0.233	0.181	0.007	0.171
Community of Valencia	0.025	0.042	0.167	0.017	0.038	–	0.029	0.135	0.001	0.296
Extremadura	0.002	0.102	0.103	0.069	0.010	0.036	–	0.191	0.004	0.214
Galicia	0.010	0.086	0.202	0.052	0.031	0.022	0.030	–	0.024	0.384
La Rioja	0.059	0.018	0.059	0.043	0.054	0.050	0.087	0.087	–	0.035
Murcia	0.002	0.055	0.000	0.023	0.034	0.004	0.024	0.000	0.061	–

*Values of FST are given below the diagonal (bold indicates significant differences), and corrected P-values are given above the diagonal.

(Group 1, 2, 4) from the ones predominant to cluster B (Group 3) and C (Group 5).

4 Discussion

In the present study, the use of appropriate strategies for exploring, incorporation and management of olive genetic resources by means of EST-SNP markers together with an intensive collaborative network, made possible the collection and identification of 173 new Spanish cultivars. It is important to highlight that most of the new germplasm identified belongs to

local and unknown cultivars. In this regard, in agreement with recent studies performed in olive (Hmam et al., 2018; Debbabi et al., 2020; Atrouze et al., 2021; Valeri et al., 2022), the identification of a high number of local cultivars indicates that the olive crop still has a high local genetic variability that needs to be recovered before its disappearance. The ongoing incorporation of this untapped local diversity into WOGBC will contribute to fulfil its main goal, that is, to acquire, maintain, document, assess and make available as much genetic diversity of the crop as possible (Belaj et al., 2016; Belaj et al., 2022).

It is expected that chances of preserving and finding untapped diversity in olive is higher in those areas with less pressure of

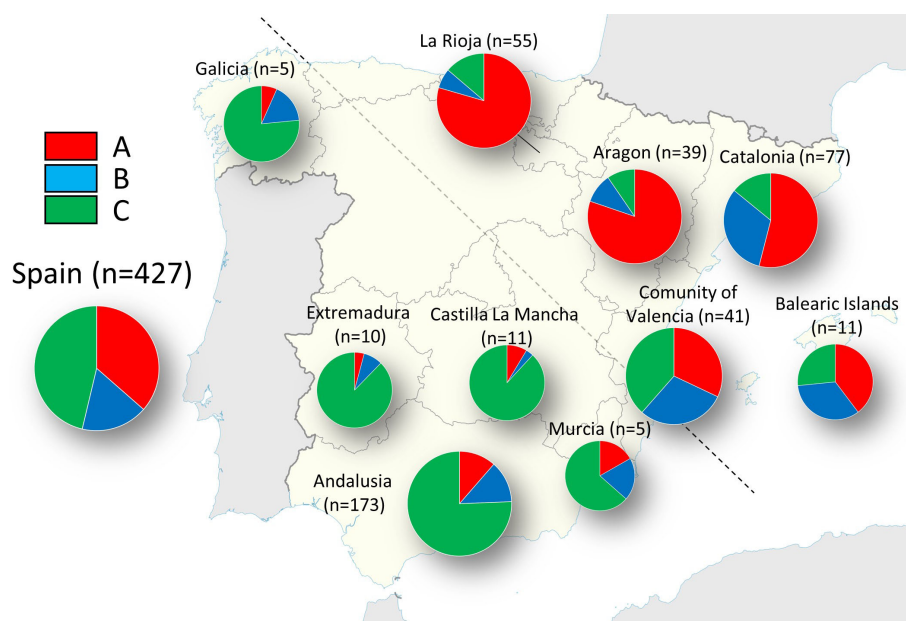


FIGURE 2
Pie charts representing the STRUCTURE clusters averaged by region for the 427 distinct genotypes.

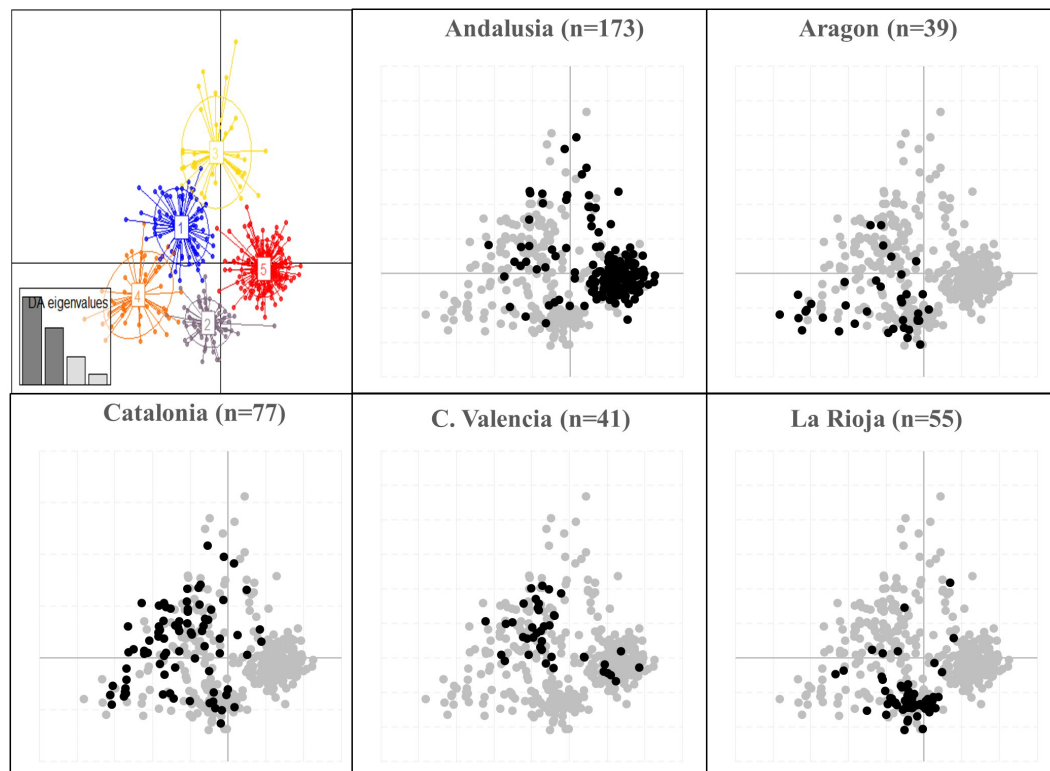


FIGURE 3

Scatter-plot of the discriminant analysis of principal components (DAPC) on a set of 427 Spanish olive genotypes identified by means 96 EST-SNP markers. Numbers and colors represent the five genetic groups found by the K-means method (see Jombart et al., 2010 for details). Regional differentiation is depicted in detail for those regions with $n > 30$ genotypes.

cultivar turnover and productivity (Belaj et al., 2022). In this sense, most of the new cultivars identified in northern areas (especially, Catalonia, La Rioja and Aragon) were probably neglected in previous national surveys, thus remaining uncatalogued throughout time, up to date. It is noteworthy the case of La Rioja, a region mainly known for its wine production, which, as shown here, has also a wide range of olive cultivars that have yet to be catalogued. In the case of Andalusia, in accordance with previous studies by means of SSR markers (Díez et al., 2011), an important number of the new cultivars identified were found in ancient trees growing in remote areas with complex topography and low productivity, i.e., areas with a low pressure of cultivar turnover. Other works that genetically characterized ancient olives trees in Mediterranean countries like Cyprus (Anestiadou et al., 2017), Israel (Barazani et al., 2014), Italy (Baldoni et al., 2006; Erre et al., 2010; Marchese et al., 2023), Malta (Valeri et al., 2022), Montenegro (Lazović et al., 2016), or Morocco (El Bakkali et al., 2013), also reported that only a small proportion of them matched to known olive cultivars. In agreement with the conclusion of these authors, our findings support that the ancient olives trees deserve a careful consideration and conservation measures as *in-situ* reservoir of olive genetic diversity.

The efficient identification of redundant germplasm prior to its introduction into a collection is as important as verifying and safeguarding as much diversity as possible. And it certainly contributes to an efficient management of olive genetic resources.

In this regard, as already seen in previous studies in olive (Díez et al., 2011; Atienza et al., 2013; Lazović et al., 2016; Ninot et al., 2018; Belaj et al., 2022), our result indicate that prospecting surveys may constitute a gauge of redundant genotypes in the same or close olive growing areas, at both local and regional scale. In addition, EST-SNP genotyping of the plant material under study made possible the detection of 65 new synonymies that were not recorded so far. In this sense, our findings reinforce the need of *a priori* identification of the new plant material prospected to avoid the inclusion of duplicates into *ex situ* germplasm collections contributing thus to their cost-effective management.

Besides, this work enabled the identification of five new homonymy groups, as well as the enlargement of well-known homonymy groups with new members. For example, “Manzanilla” denomination, which refers to “apple fruit shape” and which constitutes the greatest group of homonymies documented in Spain (Barranco et al., 2005; Belaj et al., 2022), was enlarged with 7 additional, phenotypically different, cultivars: ‘Mançanal d’Arnes’ (=‘Manzanal de Ráfales’), ‘Mançanenca d’Albagés’, ‘Mançanenca de Batea’, ‘Manzanella del Mezquín’, ‘Manzanilla de Alfarnatejo’, ‘Manzanilla Baquetera’, ‘Manzanilla Castúa’ (=‘Alameño de Cabra’).

Although various diversity studies have been conducted on olive germplasm at the national and regional level in Spain (Belaj et al., 2010; Díez et al., 2011; Trujillo et al., 2014; Fernández i Martí et al., 2015; Ninot et al., 2018), the present study constitutes the largest one performed with such a large number of Spanish

genotypes and using EST-SNPs markers. The genetic variability displayed by the set of 96 EST-SNPs on the 427 nonredundant Spanish genotypes was very similar to that obtained when using the same set of markers on 668 nonredundant Mediterranean genotypes maintained in the WOGBC (Belaj et al., 2022). This confirms the wide diversity of cultivated olive germplasm in Spain.

When evaluating the genetic variability by region, it was observed that genotypes from the northern ones were the most genetically different with respect to the rest of the Spanish regions. This could possibly indicate local adaptation of these genotypes to colder and wetter local environmental conditions (some genotypes in the pre-Pyrenees were localized above 800 m.a.s.l.) than those found in the rest of Spain (Fernández i Martí et al., 2015).

In total accordance, STRUCTURE and DAPC analysis revealed a certain geographic clustering of Spanish cultivars. Thus, the olive accessions under study clustered in three main gene pools, being the ones from North-North-eastern and Southern Spanish provinces, the most clearly differentiated. This regional differentiation is in agreement with previous studies conducted with other molecular markers in olive (Sanz-Cortés et al., 2001; Belaj et al., 2010; Díez et al., 2015; Jiménez-Ruiz et al., 2020). Some authors have suggested that this separation might be due to different routes of expansion of olive growing from the Eastern Mediterranean Basin along the South and the North coasts (Jiménez-Ruiz et al., 2020; Julca et al., 2020). In addition, a possible local selection, specifically adapted to particular environmental conditions and fulfilling agronomic expectations, may explain some of the differences found between northern and southern Spanish olive accessions. Besides, human displacement of olive cultivars under harsh agroclimatic events, might have shaped the spread and diversification of the olive tree as documented in historic literature. The Andalusian agronomist Al-Tignari documented an exuberant importation of olive trees brought in ships from northern Africa to repopulate the Al-Andalus olive grove devastated by a long drought occurred at the end of the Visigothic kingdom (mid-6th to early 8th century) (Guzmán Álvarez, 2004; Guzman Álvarez, 2007; Orlandis, 2011). Also, there are some notes about frosts and disease outbreaks in different regions of Spain that served as an incentive for olive growers to replant, renew or abandon their main cultivars during the 19th and 20th centuries (Guzman Álvarez, 2007). Finally, the high level of admixture found in Notheastern (Catalonia) and especially in Eastern (Valencia and Balearic Islands) accessions may indicate that higher interchange/flowing of plant material, could have occurred in this area, probably due to human displacement within and outside the peninsula territories (Belaj et al., 2004; Guzman Álvarez, 2007; Tous and Franquet i Bernis, 2019).

This work represents a significant enlargement of the conserved germplasm of cultivated olive in Spain. The fact that, through the surveys conducted here, there was an increase of more than 70% in the national olive germplasm accurately identified, indicates that there may still be endangered minor cultivars yet to be discovered in other olive-producing countries. And that giving the extension of the modern olive growing with very few cultivars, this cryptic germplasm is in great danger of disappearance. In our case, the new cultivars identified showed a high level of genetic diversity among and within Spanish regions, locating a new hot spot of

diversity in northern regions of the country. Some of the minor cultivars recovered were well adapted to particular environmental conditions and could harbour agronomical traits with a great potential for future national olive breeding programs aimed to mitigate climate change impact on the country. Also, local cultivars could be a very useful source of genes with great potential against new and unforeseen biotic and abiotic stresses, outburst of new pests and diseases, like the case of *Xylella fastidiosa*, as well as for improving oil quality or adapting to new market trends. In addition, the broadening of the collection may play an important role in enlarging the knowledge about olive genetic structure and relationships, which may be of interest in future genome-wide association studies and genitors selection in olive breeding programs. The set of 96 EST-SNPs markers here used proved to be an efficient tool for the identification and recovery of those minor endangered cultivars and is available to those researchers willing to perform similar works in other olive growing countries.

Data availability statement

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

Author contributions

FG-G: Data curation, Formal analysis, Methodology, Resources, Software, Visualization, Writing – original draft. AN: Data curation, Resources, Validation, Visualization, Writing – review & editing. JR: Resources, Writing – review & editing. SC: Resources, Writing – review & editing. JA: Resources, Writing – review & editing. JR: Resources, Writing – review & editing. IA: Resources, Writing – review & editing. JV-A: Resources, Writing – review & editing. JC-G: Resources, Writing – review & editing. ZS: Data curation, Formal analysis, Methodology, Software, Writing – review & editing. IL: Data curation, Formal analysis, Software, Writing – review & editing. RR-N: Conceptualization, Resources, Formal analysis, Funding acquisition, Methodology, Validation, Writing – review & editing, Writing – original draft. AB: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was financially supported by the regional IFAPA projects PR.CRF.CRF201900.004 and PR.CRF.CRF202200.004, partially funded by European Agricultural Fund for Rural Development (EAFRD). The conservation and management of WOGBC IFAPA Córdoba has been financially supported by INIA (RFP 2013-00005; RFP 2017-00007) and IFAPA (PP.PEI.IDF201601.2.; PR.CRF.CRF201900.004) projects.

Acknowledgments

The authors are grateful to all the farmers, associations, researchers, and collaborators at regional, national, and international levels for their help during prospecting and/or reception of new accessions. The authors are also grateful for the EST-SNP genotyping support of staff at UPV/EHU—Scientific Park Maria Goyri Biotechnology Center (Bizkaia, Spain). FJ G-G thanks the Programme of Grants for the Recruitment, Incorporation and Mobility of R+D+i Human Capital within the Andalusian Research, Development and Innovation Plan (PAIDI2020).

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.

References

- Anestiadou, K., Nikoloudakis, N., Hagidimitriou, M., and Katsiotis, A. (2017). Monumental olive trees of Cyprus contributed to the establishment of the contemporary olive germplasm. *PLoS One* 12 (11), e0187697. doi: 10.1371/journal.pone.0187697
- Atienza, S. G., de la Rosa, R., Domínguez-García, M. C., Martín, A., Kilian, A., and Belaj, A. (2013). Use of DArT markers as a means of better management of the diversity of olive cultivars. *Food Res. Int.* 54, 2045–2053. doi: 10.1016/j.foodres.2013.08.015
- Atrouz, K., Bousba, R., Marra, F. P., Marchese, A., Conforti, F. L., Perrone, B., et al. (2021). Algerian olive germplasm and its relationships with the central-western mediterranean varieties contributes to clarify cultivated olive diversification. *Plants-Basel* 10 (4), 678. doi: 10.3390/plants10040678
- Baldoni, L., Tosti, N., Ricciolini, C., Belaj, A., Arcioni, S., Pannelli, G., et al. (2006). Genetic structure of wild and cultivated olives in the central mediterranean basin. *Ann. Bot.* 98 (5), 935–942. doi: 10.1093/aob/mcl178
- Barazani, O., Westberg, E., Hanin, N., Dag, A., Kerem, Z., Tugendhaft, Y., et al. (2014). A comparative analysis of genetic variation in rootstocks and scions of old olive trees - a window into the history of olive cultivation practices and past genetic variation. *BMC Plant Biol.* 14, 146. doi: 10.1186/1471-2229-14-146
- Barranco, D. (2010). "Varieties and rootstocks," in *Olive growing*, vol. 757. Eds. D. Barranco, R. Fernández-Escobar and L. Rallo (Australia: RIRDC, Mundi-Prensa and Junta de Andalucía).
- Barranco, D., Cimato, A., Fiorino, P., Rallo, L., Touzani, A., Castañeda, C., et al. (2000). *World Catalogue of Olive Varieties* (Madrid, Spain: International Olive Council).
- Barranco, D., Trujillo, I., and Rallo, P. (2005). "Elaiografía Hispánica," in *Variedades de Olivo en España*. Eds. C. L. Rallo, D. Barranco, J. M. Caballero, J. Tous and I. Trujillo (Sevilla: Junta de Andalucía. Ministerio de Agricultura, Pesca y Alimentación. Ediciones Mundi-Prensa), 80–231.
- Belaj, A., Cipriani, G., Testolin, R., Rallo, L., and Trujillo, I. (2004). Characterization and identification of the main Spanish and Italian olive cultivars by simple-sequence-repeat markers. *Hortscience* 39 (7), 1557–1561. doi: 10.21273/HORTSCI.39.7.1557
- Belaj, A., de la Rosa, R., Lorite, I. J., Mariotti, R., Cultrera, N. G. M., Beuzón, C. R., et al. (2018). Usefulness of a new large set of high throughput EST-SNP markers as a tool for olive germplasm collection management. *Front. Plant Sci.* 9 (1320). doi: 10.3389/fpls.2018.01320
- Belaj, A., Domínguez-García, M., Gustavo Atienza, S., Martín Urdiroz, N., de la Rosa, R., Satovic, Z., et al. (2012). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* 8 (2), 365–378. doi: 10.1007/s11295-011-0447-6
- Belaj, A., Gurbuz Veral, M., Sikaoui, H., Moukhlí, A., Khadiri, B., Mariotti, R., et al. (2016). "Olive Genetic Resources," in *The Olive Tree Genome, Compendium of Plant Genomes*. Eds. E. Rugini, L. Baldoni, R. Muleo and L. Sebastiani (Cham, Switzerland: Springer International Publishing), 27–54.
- Belaj, A., Muñoz-Díez, C., Baldoni, L., Satovic, Z., and Barranco, D. (2010). Genetic diversity and relationships of wild and cultivated olives at regional level in Spain. *Scientia Hort.* 124, 323–330. doi: 10.1016/j.scienta.2010.01.010
- Belaj, A., Ninot, A., Gómez-Gálvez, F. J., El Riachy, M., Gurbuz-Veral, M., Torres, M., et al. (2022). Utility of EST-SNP markers for improving management and use of olive genetic resources: A case study at the worldwide olive germplasm bank of Córdoba. *Plants* 11 (7), 921. doi: 10.3390/plants11070921
- Besnard, G., Terral, J.-F., and Cornille, A. (2018). On the origins and domestication of the olive: a review and perspectives. *Ann. Bot.* 121 (3), 385–403. doi: 10.1093/aob/mcx145
- Debbabi, O. S., Miazzi, M. M., Elloumi, O., Fendri, M. F., Ben Amar, F., Savoia, M., et al. (2020). Recovery, assessment, and molecular characterization of minor olive genotypes in Tunisia. *Plants-Basel* 9 (3), 382. doi: 10.3390/plants9030382
- de Jong, M. J., de Jong, J. F., Hoelzel, A. R., and Janke, A. (2021). SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. *Mol. Ecol. Resour.* 21 (4), 1369–1379. doi: 10.1111/1755-0998.13339
- de la Rosa, R., James, C. M., and Tobutt, K. R. (2002). Isolation and characterization of polymorphic microsatellites in olive (*Olea europaea* L.) and their transferability to other genera in the Oleaceae. *Mol. Ecol. Notes* 2 (3), 265–267. doi: 10.1046/j.1471-8278.2002.00217.x
- de la Rosa, R., León, L., Guerrero, N., Rallo, L., and Barranco, D. (2007). Preliminary results of an olive cultivar trial at high density. *Aust. J. Agric. Res.* 58 (5), 392–395. doi: 10.1071/ar06265
- Díez, M. J., de la Rosa, L., Martín, I., Guasch, L., Cartea, M. E., Mallor, C., et al. (2018). Plant genebanks: present situation and proposals for their improvement. The case of the Spanish network. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01794
- Díez, C. M., Trujillo, I., Barrio, E., Belaj, A., Barranco, D., and Rallo, L. (2011). Centennial olive trees as a reservoir of genetic diversity. *Ann. Bot.* 108 (5), 797–807. doi: 10.1093/aob/mcr194
- Díez, C. M., Trujillo, I., Martínez-Urdiroz, N., Barranco, D., Rallo, L., Marfil, P., et al. (2015). Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* 206 (1), 436–447. doi: 10.1111/nph.13181
- El Bakkali, A., Essalouh, L., Tollon, C., Rivallan, R., Mournet, P., Moukhlí, A., et al. (2019). Characterization of Worldwide Olive Germplasm Banks of Marrakech (Morocco) and Córdoba (Spain): Towards management and use of olive germplasm in breeding programs. *PLoS One* 14 (10), e0223716. doi: 10.1371/journal.pone.0223716
- El Bakkali, A., Haouane, H., Hadiddou, A., Oukabli, A., Santoni, S., Udupa, S. M., et al. (2013). Genetic diversity of on-farm selected olive trees in Moroccan traditional olive orchards. *Plant Genet. Res.: Character. Utilization* 11 (2), 97–105. doi: 10.1017/s1479262112000445
- Erre, P., Chessa, I., Muñoz-Díez, C., Belaj, A., Rallo, L., and Trujillo, I. (2010). Genetic diversity and relationships between wild and cultivated olives (*Olea europaea* L.) in Sardinia as assessed by SSR markers. *Genet. Resour. Crop Evol.* 57 (1), 41–54. doi: 10.1007/s10722-009-9449-8
- FAOSTAT. (2020). Available at: <http://www.fao.org/faostat> (Accessed 06.11.2022).
- Fernández Escobar, R., de la Rosa, R., León, L., Gómez, J. A., Testi, F., Orgaz, M., et al. (2013). Evolution and sustainability of the olive production systems. In: Arcas, N., Arroyo López, F. N., Caballero, J., D'Andria, R., Fernández, M., Fernández Escobar, R., et al. Present and future of the Mediterranean olive sector. (Zaragoza: CIHEAM / IOC) 11-42 (Options Méditerranéennes: Série A. Séminaires Méditerranéens; n. 106).
- Fernández i Martí, A., Font i Forcada, C., Socías i Company, R., and Rubio-Cabetas, M. (2015). Genetic relationships and population structure of local olive tree accessions

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.2023.1267601/full#supplementary-material>

from Northeastern Spain revealed by SSR markers. *Acta Physiol. Plant.* 37 (1), 1–12. doi: 10.1007/s11738-014-1726-2

Guzmán Álvarez, J. R. (2004). *El Palimpsesto Cultivado. Historia de los paisajes del olivar andaluz*. Junta de Andalucía. Consejería de Agricultura y Pesca (Sevilla, España: Servicio de Publicaciones y Divulgación). Available at: http://www.juntadeandalucia.es/export/drupaljda/1337165052El_Palimpsesto_cultivado.pdf. Colección el arado y la red.

Guzmán Álvarez, J. R. (2007). “La génesis de los paisajes olivareros: siglos XVI-XX,” in *Tierras del Olivo* (Jaén, Spain: Editorial El Legado Andalusi), 185–197.

Hmmam, I., Mariotti, R., Ruperti, B., Cultrera, N., Baldoni, L., and Barcaccia, G. (2018). Venetian olive (*Olea europaea*) germplasm: disclosing the genetic identity of locally grown cultivars suited for typical extra virgin oil productions. *Genet. Resour. Crop Evol.* 65 (6), 1733–1750. doi: 10.1007/s10722-018-0650-5

Íñiguez, A., Paz, S., and Illa, F. J. (2001). *Variedades de olivo cultivadas en la Comunidad Valenciana* (Valencia, Spain: Generalitat Valenciana. Conselleria de Agricultura, Pesca y Alimentación).

Jiménez-Ruiz, J., Ramírez-Tejero, J. A., Fernández-Pozo, N., Leyva-Pérez, M. O., Yan, H., Rosa, R., et al. (2020). Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome* 13 (1), e20010. doi: 10.1002/tpg2.20010

Jombart, T., and Collins, C. (2015). A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.0.0, 1–43. Available at: <https://adegenet.r-forge.r-project.org/files/adegenet-dapc.pdf>.

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

Julca, I., Marcet-Houben, M., Cruz, F., Gómez-Garrido, J., Gaut, B. S., Diez, C. M., et al. (2020). Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biol.* 18 (1), 148. doi: 10.1186/s12915-020-00881-6

Kaniewski, D., Van Campo, E., Boiy, T., Terral, J.-F., Khadari, B., and Besnard, G. (2012). Primary domestication and early uses of the emblematic olive tree: palaeobotanical, historical and molecular evidence from the Middle East. *Biol. Rev. Cambridge Philos. Soc.* 87 (4), 885–899. doi: 10.1111/j.1469-185X.2012.00229.x

Lazović, B., Adakalić, M., Pucci, C., Perović, T., Bandelj, D., Belaj, A., et al. (2016). Characterizing ancient and local olive germplasm from Montenegro. *Scientia Hort.* 209, 117–123. doi: 10.1016/j.scientia.2016.06.022

León, L., de la Rosa, R., and Arriaza, M. (2021). Prioritization of olive breeding objectives in Spain: Analysis of a producers and researchers survey. *Spanish J. Agric. Res.* 19 (4):e0701. doi: 10.5424/sjar/2021194-18203

Lorite, I. J., Cabezas, J. M., Ruiz-Ramos, M., de la Rosa, R., Soriano, M. A., Leon, L., et al. (2022). Enhancing the sustainability of Mediterranean olive groves through adaptation measures to climate change using modelling and surfaces. *Agric. For. Meteorol.* 313. doi: 10.1016/j.agrformet.2021.108742

Marchese, A., Bonanno, F., Marra, F. P., Trippa, D. A., Zelasco, S., Rizzo, S., et al. (2023). Recovery and genotyping ancient Sicilian monumental olive trees. *Front. Conserv. Sci.* 4. doi: 10.3389/fcsc.2023.1206832

Marshall, T. C., Slate, J., Kruuk, L. E. B., and Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* 7 (5), 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

Medina-Alonso, M. G., Navas, J. F., Cabezas, J. M., Weiland, C. M., Rios-Mesa, D., Lorite, I. J., et al. (2020). Differences on flowering phenology under Mediterranean and

Subtropical environments for two representative olive cultivars. *Environ. Exp. Bot.* 180. doi: 10.1016/j.envexpbot.2020.104239

Ninot, A., Howad, W., Aranzana, M. J., Senar, R., Romero, A., Mariotti, R., et al. (2018). Survey of over 4,500 monumental olive trees preserved on-farm in the northeast Iberian Peninsula, their genotyping and characterization. *Scientia Hort.* 231, 253–264. doi: 10.1016/j.scientia.2017.11.025

Orlandis, J. (2011). *Historia del reino visigodo español* (Madrid, España: Ediciones RIALP, S.A.).

Peakall, R., and Smouse, P. E. (2012). GenAEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28 (19), 2537–2539. doi: 10.1093/bioinformatics/bts460

Pérez, A. G., León, L., Pascual, M., de la Rosa, R., Belaj, A., and Sanz, C. (2019). Analysis of olive (*Olea europaea* L.) genetic resources in relation to the content of vitamin E in virgin olive oil. *Antioxidants* 8 (242), 242. doi: 10.3390/antiox8080242

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155 (2), 945–959. doi: 10.1093/genetics/155.2.945

Rallo, L., Barranco, D., Castro-García, S., Connor, D. J., del Campo, M. G., and Rallo, P. (2013). High-Density Olive Plantations. *Horticultural Reviews* (41), 303–384. doi: 10.1002/9781118707418.ch07

R Core Team (2021). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: <https://www.R-project.org/>.

Sanz-Cortés, F., Badenes, M. L., Paz, S., Íñiguez, A., and Llácer, G. (2001). Molecular characterization of olive cultivars using RAPD markers. *J. Am. Soc. Hortic. Sci.* 126 (1), 7–12. doi: 10.21273/JASHS.126.1.07

Serrano, A., de la Rosa, R., Sanchez-Ortiz, A., and Leon, L. (2020). Genetic and environmental effect on volatile composition of extra virgin olive oil. *Eur. J. Lipid Sci. Technol.* 122 (12). doi: 10.1002/ejlt.202000162

Terral, J. F., Alonso, N., Buxó, R., Chatti, N., Fabre, L., Fiorentino, G., et al. (2004). Historical biogeography of olive domestication (*Olea europaea* L.) as revealed by geometrical morphometry applied to biological and archaeological material. *J. Biogeogr.* 31 (1), 63–77. doi: 10.1046/j.0305-0270.2003.01019.x

Terral, J. F., and Arnold-Simard, G. (1996). Beginnings of olive cultivation in eastern Spain in relation to holocene bioclimatic changes. *Quaternary Res.* 46 (2), 176–185. doi: 10.1006/qres.1996.0057

Tous, J., and Franquet i Bernis, J. M. (2019). *Olivo y Aceites de Calidad*. (Benicarló, España: Onada Ediciones).

Trujillo, I., Ojeda, M., Urdiroz, N., Potter, D., Barranco, D., Rallo, L., et al. (2014). Identification of the Worldwide Olive Germplasm Bank of Córdoba (Spain) using SSR and morphological markers. *Tree Genet. Genomes* 10 (1), 141–155. doi: 10.1007/s11295-013-0671-3

Valeri, M. C., Mifsud, D., Sammut, C., Pandolfi, S., Lilli, E., Bufacchi, M., et al. (2022). Exploring olive genetic diversity in the Maltese islands. *Sustainability* 14 (17), 10684. doi: 10.3390/su141710684

Viñuales-Andreu, J. (2007). *Variedades de olivo del Somontano* (Huesca, España: Huesca).

Yilmaz-Duzyaman, H., de la Rosa, R., and Leon, L. (2022). Seedling selection in olive breeding progenies. *Plants-Basel* 11 (9), 1195. doi: 10.3390/plants11091195



OPEN ACCESS

EDITED BY

Axel Diederichsen,
Agriculture and Agri-Food Canada (AAFC),
Canada

REVIEWED BY

Photini V. Mylona,
Hellenic Agricultural Organisation (HAO),
Greece
Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

*CORRESPONDENCE

Jochen C. Reif

✉ reif@ipk-gatersleben.de

RECEIVED 31 July 2023

ACCEPTED 31 October 2023

PUBLISHED 11 January 2024

CITATION

Berkner MO, Weise S, Reif JC and
Schulthess AW (2024) Genomic prediction
reveals unexplored variation in grain
protein and lysine content across a vast
winter wheat genebank collection.
Front. Plant Sci. 14:1270298.
doi: 10.3389/fpls.2023.1270298

COPYRIGHT

© 2024 Berkner, Weise, Reif and Schulthess.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). 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.

Genomic prediction reveals unexplored variation in grain protein and lysine content across a vast winter wheat genebank collection

Marcel O. Berkner¹, Stephan Weise², Jochen C. Reif^{1*}
and Albert W. Schulthess¹

¹Breeding Research Department, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany, ²Genebank Department, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany

Globally, wheat (*Triticum aestivum* L.) is a major source of proteins in human nutrition despite its unbalanced amino acid composition. The low lysine content in the protein fraction of wheat can lead to protein-energy-malnutrition prominently in developing countries. A promising strategy to overcome this problem is to breed varieties which combine high protein content with high lysine content. Nevertheless, this requires the incorporation of yet undefined donor genotypes into pre-breeding programs. Genebank collections are suspected to harbor the needed genetic diversity. In the 1970s, a large-scale screening of protein traits was conducted for the wheat genebank collection in Gatersleben; however, this data has been poorly mined so far. In the present study, a large historical dataset on protein content and lysine content of 4,971 accessions was curated, strictly corrected for outliers as well as for unreplicated data and consolidated as the corresponding adjusted entry means. Four genomic prediction approaches were compared based on the ability to accurately predict the traits of interest. High-quality phenotypic data of 558 accessions was leveraged by engaging the best performing prediction model, namely EG-BLUP. Finally, this publication incorporates predicted phenotypes of 7,651 accessions of the winter wheat collection. Five accessions were proposed as donor genotypes due to the combination of outstanding high protein content as well as lysine content. Further investigation of the passport data suggested an association of the adjusted lysine content with the elevation of the collecting site. This publicly available information can facilitate future pre-breeding activities.

KEYWORDS

genebank genomics, genomic prediction, grain quality, lysine content, protein content, wheat

1 Introduction

Worldwide, 410 Mt of consumable plant-based proteins are provided by agriculture, with soybean (*Glycine max* (L.) Merr.), maize (*Zea mays* L.), and wheat (*Triticum aestivum* L.) contributing the largest quantities (Leinonen et al., 2019). Unlike the first two crops, wheat is mostly used directly in human nutrition (OECD/FAO, 2021). Thus, it is not surprising that wheat provides on average 19% of the proteins consumed by humans, with some regional peaks reaching more than one third in North Africa as well as in West and Central Asia (Erenstein et al., 2022). Remarkable ratios were also found in some regions of South Asia: wheat consumed as flat bread accounts for about three-fifths of the daily protein consumption in Pakistani households (Hussain et al., 2004). Undoubtedly, the associated dominance in the diets are partially due to the prevalent cultivation in the respective regions but also wheat's widespread availability on a global market (Shewry and Hey, 2015). Moreover, the preference for wheat can also be assigned to the specific characteristics of the protein fraction of the wheat grain which lead to the unique baking and processing quality of wheat flour (Shewry, 2009). This is one of the reasons for wheat being processed to a diversity of breads, pastries and noodles (Shewry, 2009) and as such forms a key aspect of the cuisine in many regions.

Despite the large quantity of consumed wheat protein, the nutritional quality of this protein is rather inadequate due to the unbalanced amino acid composition. In particular, shortcomings in the lysine content are the limiting factor (Leinonen et al., 2019) which is especially problematic since the essential amino acid lysine cannot be produced by the human organism itself and thus, must be obtained from the diet (Ufaz and Galili, 2008). On the one hand, these shortcomings can be leveled out in a diverse diet which comprises lysine-rich protein sources such as legumes, meat, fish or dairy products (Ritchie et al., 2018; Leinonen et al., 2019). On the other hand, a considerable number of people, especially in developing countries, does not have the purchasing power to diversify their diet with, for example, animal-based products (Hussain et al., 2004; Pellett and Ghosh, 2004; Muleya et al., 2022). An unbalanced wheat-rich diet may result in lysine deficiency (Meybodi et al., 2019). Such a deficiency is known to cause severe physical underdevelopment in children (Batool et al., 2015). Moreover, an inadequate supply with high quality protein can affect physiological processes, the immune system as well as the cognitive development (Batool et al., 2015). Impact on elderly adults is also widely reported and for this group, deficiency results in severe impairment of health including symptoms such as anemia and fatigue (Meybodi et al., 2019). Overall, the symptoms associated with inadequate protein supply are summarized under the name protein-energy malnutrition (Meybodi et al., 2019) and affect millions of people in developing countries (Batool et al., 2015).

Some strategies have already been proposed to increase the lysine content of staple foods. For example, artificial fortification of wheat flour with ground legumes, pseudo cereals or synthesized amino acids (Hussain et al., 2004) has been shown to be effective, but may have adverse effects on the processing quality or taste of end products (Meybodi et al., 2019). Another promising strategy

could be to breed wheat varieties which combine an overall high protein content with an enrichment of lysine in the grain. In general, the potential of developing cereal crops with such characteristics has been demonstrated in maize. Naturally occurring maize mutants, such as *opaque2* and *floury2*, have been reported with a significantly elevated lysine content (Morton et al., 2016). In a case study, Muleya and collaborators (2022) concluded that the use of varieties with such a mutation reduces the risk of lysine deficiency by 21% for the poorest quintile of households in Malawi. To the best of our knowledge, an analog wheat variety has however not been developed so far. Lysine content has generally not been of interest in commercial wheat breeding programs and therefore, the potential of a breeding-based approach might be particularly high for such an orphan trait. Moreover, the naturally occurring lysine content of wheat grains is mainly influenced by the genotype and depends only to a small extent on environmental factors (Lawrence, 1976). Both arguments advocate for a breeding-based approach such as outlined: Firstly, the variation in lysine content of a large quantity of genotypes needs to be analyzed which is very laborious in the field and laboratory. The first step is followed by the identification of donor genotypes with a high lysine content in the protein fraction. Lastly, the favorable genetics of donor genotypes would be considered in pre-breeding activities and selectively transferred into the elite gene pool of modern breeding programs. While the latter step is mostly rather foreseeable, the first two are the bottlenecks for increasing the lysine content because they are time-consuming, demand resources and the result largely depends on the variation available for analysis.

Genebank collections for wheat are known to harbor large genetic diversity (Sansaloni et al., 2020; Schulthess et al., 2022) and phenotypic variation (Philipp et al., 2018; Schulthess et al., 2022). Diversity is however trait-specific and thus, identification of potential donor genotypes with a high content of lysine and protein rely on phenotypic investigation. Earlier attempts to screen genebank collections of wheat for both traits date back to the early 1970s. Vogel and collaborators evaluated 12,613 wheat accessions from the World Wheat Collection of the United States Department of Agriculture (USDA) (Vogel et al., 1973). In the same decade, both traits were measured for 9,706 *Triticum* accessions at the predecessor institution of the Federal *ex situ* Genebank of Agricultural and Horticultural Crops which is today hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research in Gatersleben (IPK Genebank) (Lehmann et al., 1978). The aim of the aforementioned study was to screen the entire collection once and to identify accessions with a strong deviation from the population mean. The deviating accessions were re-evaluated in another year in order to account for an overestimation due to environmental effects. Until the mid-1980s, further successions were successively investigated in a structured manner (Müntz and Lehmann, 1987). Despite the sheer amount of work reflected by the work from Lehmann and collaborators (1978), this data has not been mined in depth according to today's standards and possibilities. Since then, developments in biostatistics and genomics urge the need for a re-evaluation of this historical dataset. This includes the connection of phenotypic data to genotypic data derived by next generation sequencing, which becomes more and more available for large

parts of the cereal collections at the IPK Genebank (Schulthess et al., 2022). Combining and analyzing data will undoubtedly become more important for the work of genebank curators in the future. Since the evaluation in the 1970s, the IPK Genebank has increased in size. With more than 27 thousand genebank accessions of *Triticum* species (Oppermann, 2023), the IPK Genebank preserves nowadays the 9th largest collection of plant genetic resources of wheat and its crop wild relatives (FAO, 2010). Genomic prediction could be used to characterize these new non-phenotyped parts of the collection as well as those parts without reliable phenotypic data. The power of targeted genomic prediction has recently been shown by many studies in the context of genebanks (Yu et al., 2016; Gonzalez et al., 2021; Berkner et al., 2022; Schulthess et al., 2022). Finally, informing the interested public on the newly generated information according to the FAIR (Findable, Accessible, Interoperable and Reusable) (Wilkinson et al., 2016) principles will further activate genebanks. This strategy could enable breeders to specifically select suitable donor genotypes and eventually, it may contribute to a future with less malnutrition in developing countries.

The main aim of this study was to activate historical records of the nutritional quality of wheat accessions stored at the IPK Genebank for their use in plant breeding and research. In more detail, we targeted (1) to curate the raw historical records for protein and lysine content which were generated between 1970 and 1986, (2) to analyze the data across years in order to generate outlier-corrected best linear unbiased estimates (BLUEs) for genebank accessions, (3) to apply a most suited model for genomic prediction in order to predict phenotypes for the majority of genebank accessions and (4) to suggest a set of well characterized suitable donor genotypes to breeders and the interested public.

2 Materials and methods

2.1 Curation of historical records

Historical data on protein and lysine content were compiled and curated. Some of the data originally recorded on punched tapes was unlocked; other data was recorded manually from paper files. All records were checked for accuracy and linked to the currently used accession numbers. This data originates from a large screening of the *Triticum* collection of the IPK Genebank. Between 1970 and 1986, 4,971 accessions were cultivated in 11 almost consecutive years (Figure S1), seeds were harvested and analyzed in the laboratory for protein content and lysine content. Detailed description of the procedure has been given by Lehmann and collaborators (1978). BLUEs for thousand grain weight (TGW) were used as published by Philipp and collaborators (2019).

2.2 Origin and curation of genomic data

This study relied on a genomic dataset which has been generated by Schulthess and collaborators (2022). Briefly, the authors requested 7,651 accessions from the *Triticum* collection

of the IPK Genebank and developed 7,745 isolate lines from them. From here onward, these isolates are referred to as accession samples. All 7,745 accession samples were genotyped by following a genotyping-by-sequencing approach. Reads were aligned to the first version of the reference genome var. Chinese Spring (IWGSC, 2018). After alignment, markers were rejected based on homozygosity of either the reference or alternative allele. In the next step, information of markers was omitted based on missing values (> 10%), a minimum homozygous allele count of < 10% and a maximum heterozygosity of > 1%. Later, imputation was done based on the dominant allele. Afterwards, further filtering based on a minor allele frequency of 1% led to a final matrix with 17,118 markers which was used for downstream analysis.

2.3 Outlier correction and analysis of phenotypic data

The raw data for protein and lysine content was trimmed to ensure that the data could be analyzed. Per trait, accessions were excluded from further analysis if they were represented by a single datapoint. Phenotypic values of 561 accessions remained after this trimming. Furthermore, all records of a year were omitted if no overlap with records of other years could be found. Outlier correction and calculation of BLUEs was done as described by Philipp and collaborators (2018). Briefly, the following linear mixed model was fitted to the data:

$$y_{ij} = \mu + g_i + a_j + e_{i(j)} \quad (1)$$

where y_{ij} is the protein content (or lysine content) measured on seeds of the accession i which were harvested in the year j . Accordingly, μ is the general fixed population average effect, while g_i and a_j represent the effects of the genotype and the year, respectively. The term $e_{i(j)}$ refers to the error of the model of which the variance is modelled as specific for each year. For the identification of outliers and the estimation of BLUEs, the term g_i was modelled as fixed while a_j was modelled as random. In contrast, both terms were considered as random for the calculation of heritability. Outliers were identified based on standardized residuals and with a correction for multiple testing (Holm, 1979; Nobre and Singer, 2011) as implemented by Philipp and collaborators (2018).

Heritabilities (h^2) of both traits were estimated as described by Philipp and collaborators (2018),

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_e^2}{\bar{N}_Y}} \quad (2)$$

where σ_G^2 and σ_e^2 refer to the genetic variance and the average of year-specific error variances, respectively. \bar{N}_Y is the average number of years in which an accession was tested. In addition, above explained variances components were used to compute plot-based heritabilities as:

$$h_{pb}^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2} \quad (3)$$

Lysine content was adjusted for protein content and TGW, because lysine content was strongly correlated with both other traits. The adjustment approach was a derivative of the approach applied by Vogel and collaborators (1975). Briefly, a multiple linear regression model was fitted on the BLUEs of lysine content in dependence on protein content and TGW. Afterwards, lysine content was adjusted genotype-wise based on the partial regression coefficients and the mean-centered protein content and TGW values as follows:

$$\text{Lysine}_{adj} = \text{Lysine} - b_{\text{Protein}}(\text{Protein} - \overline{\text{Protein}}) - b_{\text{TGW}}(\text{TGW} - \overline{\text{TGW}}) \quad (4)$$

2.4 Analysis of population structure

The relatedness of the 7,745 genotyped accession samples was studied based on a principal coordinate (PCo) analysis (Gower, 1966). For this, pair-wise Rogers' distances (Rogers, 1972) were calculated between genomic profiles of all accession samples and compiled into a distance matrix; the complexity of the distance matrix was reduced by deriving PCos (Gower, 1966). First and second PCos, which retain the highest amount of variation, were plotted against each other to graphically portray possible patterns resulting from population structure.

2.5 Genomic prediction models and their evaluation

In the present study, four different genomic prediction models, namely G-BLUP, EG-BLUP, Bayes A, and Bayesian Lasso, were compared based on their performance. The G-BLUP model (VanRaden, 2008) predicts phenotypic values based on additive genetic effects. These effects are explained by the relationship among the genotypes. The prediction model for n genotypes has the following matrix notation (Henderson, 1985):

$$\mathbf{y} = \mathbf{1}_n\boldsymbol{\mu} + \mathbf{g} + \mathbf{e} \quad (5)$$

where the phenotypic values (BLUEs), given by the vector \mathbf{y} , are a function of the general mean ($\boldsymbol{\mu}$) and the n -dimensional vectors \mathbf{g} and \mathbf{e} , which account for the genotypic values and the model's residuals, respectively. The n -dimensional vector of ones ($\mathbf{1}_n$) assigns $\boldsymbol{\mu}$ to each element of \mathbf{y} . The vectors \mathbf{g} and \mathbf{e} follow multivariate normal distributions $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ which depend on the genomic-estimated additive relationship matrix \mathbf{G} and the genetic variance (σ_g^2) or \mathbf{I} and the residual variance (σ_e^2), respectively. The $n \times n$ matrix \mathbf{G} was calculated based on the first method described in VanRaden (2008) while \mathbf{I} is an n -dimensional identity matrix.

EG-BLUP accounts for additive-by-additive epistasis (Jiang and Reif, 2015) and can be seen as an extension to the G-BLUP model as follows:

$$\mathbf{y} = \mathbf{1}_n\boldsymbol{\mu} + \mathbf{g} + \mathbf{g}_1 + \mathbf{e} \quad (6)$$

In equation (6), the terms \mathbf{y} , $\mathbf{1}_n$, $\boldsymbol{\mu}$, \mathbf{g} , and \mathbf{e} are as defined in equation (5). The n -dimensional vector \mathbf{g}_1 accounts for the additive-by-additive effect among genotypes. This effect follows a multivariate normal distribution $\mathbf{g}_1 \sim N(\mathbf{0}, \mathbf{H}\sigma_{g_1}^2)$, where $\mathbf{H} = \mathbf{G}\#\mathbf{G}$, with $\#$ being the Hadamard product operator (Jiang and Reif, 2015).

BayesA (Meuwissen et al., 2001) was used as the base Bayesian model for genomic prediction. In this prediction approach, priors of the regression parameters are assumed to follow a scaled-t distribution. Genomic prediction with the Bayesian Lasso was applied according to Park and Casella (2008). In this approach, the regression parameters have a double-exponential prior.

The four genomic prediction models were compared based on their ability to accurately predict phenotypes. In this comparison, the unit of quality was the prediction ability, which was defined as the correlation between the BLUEs and the predicted phenotypes. The comparison was established by means of five-fold cross-validation. All accession samples with known BLUEs were assigned to one of five equally sized groups. Four of these groups were incorporated in the prediction model as training set in order to predict the phenotypes in the remaining group, known as test set. The prediction was repeated in such a manner that each group has once been the test set and four times part of the training set. Thereafter, predicted phenotypes of all test sets were combined and the Pearson correlation coefficient with the respective BLUEs was calculated. This whole process was independently repeated 100 times. For an unbiased comparison, all models were tested based on the same training and test set. The best performing prediction model was used to predict the phenotypes of all accession samples with genomic data. In the latter case, all accession samples having available phenotypic data were used as training set.

All computational calculations, analysis as well as the creation of figures was implemented in the R environment (R v. 4.0.2). Solving the linear mixed model for data curation, BLUEs computation and the estimation of variance components from phenotypic data was done by engaging ASReml-R 4 v. 4.1.0.110 (Butler et al., 2018). Genomic prediction models were implemented with the R package BGLR v. 1.0.8 (Pérez and De Los Campos, 2014).

3 Results

3.1 Curated data with high quality

The data curation resulted in a comprehensive dataset for protein content and lysine content which comprised 11 years of experimental trials. In total, the resulting raw dataset included 5,952 records for protein content and 5,940 records for lysine content from a total of 4,971 accessions. Across years, the raw data did not only display differences in the traits' distributions; but moreover, the number of recorded data points differed strongly with a clear dominance for the year 1970 in which 3,442 records were taken per trait (Figure S1). In contrast, only six measurements were reported

for 1983; these were excluded due to the absent overlap with any other year. Despite the large amount of data, the dataset was rather incomplete with an unbalanced structure and most accessions tested only in one year (Table S1): Of all 4,971 accessions, 4,410 accessions were grown and characterized once without any replication. These records were excluded from further analysis to ensure that reliable BLUEs can be obtained for the remaining accessions. After this step, remaining accessions were evaluated in up to seven years, with an average number of 2.41 and 2.39 for protein and lysine content, respectively.

The quality of the data can be reviewed based on heritability for the two traits (Table 1). For protein content, the heritability before outlier correction reached 0.77 and could only be slightly improved due to the correction. The quality of the data for lysine content improved by 23.14% due to the removal of 17 outlier data points, leading to an increase in heritability from 0.47 to 0.58. The

estimated plot-based heritabilities behaved accordingly, as also evidenced by the negligible amount of rejected data points.

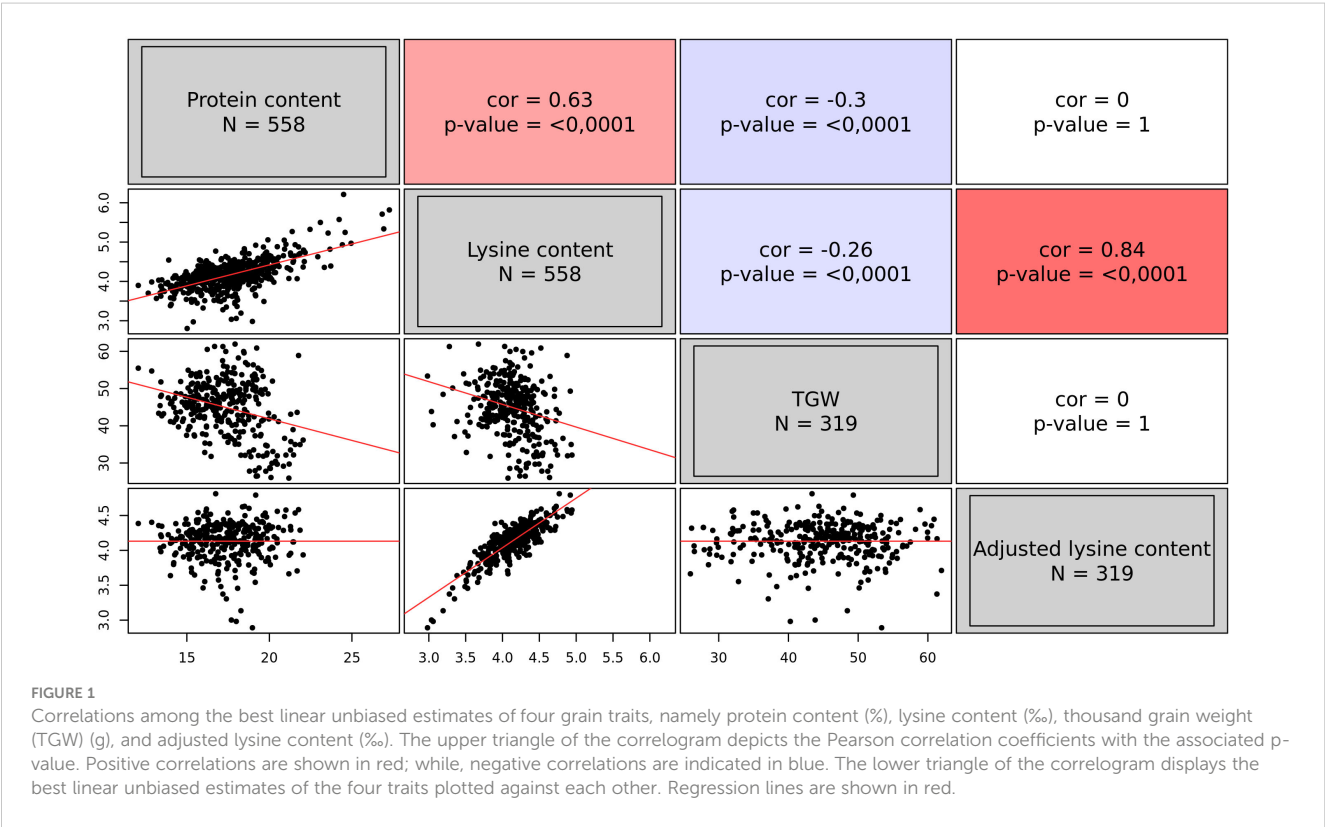
3.2 Estimated average phenotypic performance of accessions and associations of the traits

The analysis resulted in BLUEs of 558 accessions for both traits, namely protein content and lysine content. On average, accessions had a protein content of 17.61% and a lysine content of 4.17%. However, some accessions were found with a very positive deviation from the average (Figure 1): 12 accessions exhibited a lysine content of more than 5.0‰. Protein content and lysine content were highly correlated ($r = 0.63$, $p < 0.01$). Out of the 558 accessions, only 319 accessions additionally had BLUEs for TGW. TGW was negatively

TABLE 1 Description of the dataset for protein content and lysine content before and after outlier correction.

Trait	Outlier correction	μ_N	σ_g^2	σ_e^2	h^2	h_{pb}^2
Protein content	before	2.410	3.894	2.808	0.77	0.58
	after	2.407	4.017	2.679	0.78	0.60
Lysine content	before	2.394	0.087	0.235	0.47	0.27
	after	2.385	0.096	0.168	0.58	0.36

Depicted are the average number of datapoints per accession (μ_N), the genetic variance (σ_g^2), the average of year-specific error variances (σ_e^2) as well as the heritability (h^2) and the plot-based heritability (h_{pb}^2). Accessions with a single datapoint per trait were disregarded here.



correlated with both, protein content and lysine content (Figure 1). With this information, the adjusted lysine content of these 319 accessions was calculated. The adjustment completely broke the correlations of lysine content with TGW and protein content but retained the strong correlation with lysine content *per se*.

3.3 Comparison of different genomic prediction approaches

Both genotypic data and phenotypic data, which were available for 337 accession samples, were used for further genomic analysis and prediction of protein and lysine content. The number of genotyped accession samples with phenotypic information was only slightly lower for adjusted lysine content. In addition, no

clear relationship pattern was observed between the distribution of the 337 accession samples along the PCos of the genomic distances and the phenotypic variation. Thus, no subpopulations were found with substantially higher or worse performing accessions compared to the population average. All in all, the available training set with reliable phenotypes corresponds to a representative sample of the whole winter wheat collection (Figure 2). Therefore, despite its limited size and provided high cross-validated prediction abilities, reliable predictions should be expected for both, phenotypic and non-phenotypic accessions.

Four different genome-wide prediction approaches were implemented and compared based on the correlation between BLUEs and predicted phenotypes. EG-BLUP outperformed G-BLUP and both Bayesian methods for the prediction of protein content, lysine content, and adjusted lysine content (Figure 3). In

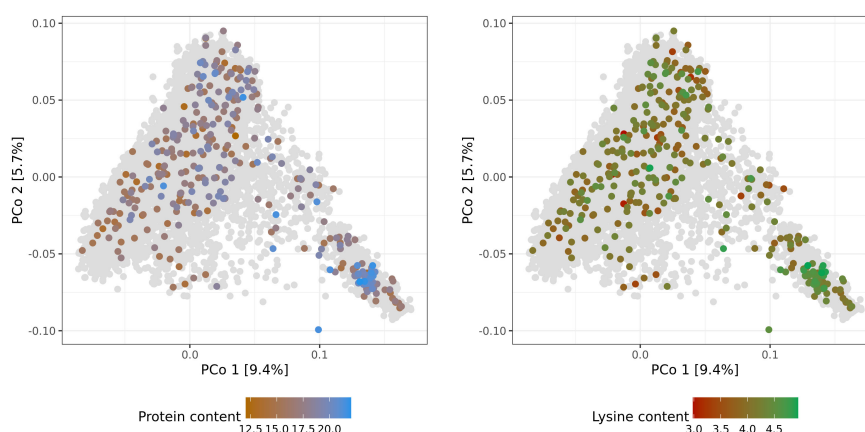


FIGURE 2

Molecular diversity of the IPK winter wheat collection covered by accession samples with best linear unbiased estimates of protein (%) and lysine (%) content. Distributions of the 337 phenotypic values are depicted via colorcoding and shown separately per trait. Biplots are based on the first and second principal coordinates (PCo) from the Rogers' distances between 7,745 accession samples characterized with genotyping-by-sequencing. Gray dots represent genotyped accession samples lacking best linear unbiased estimates.

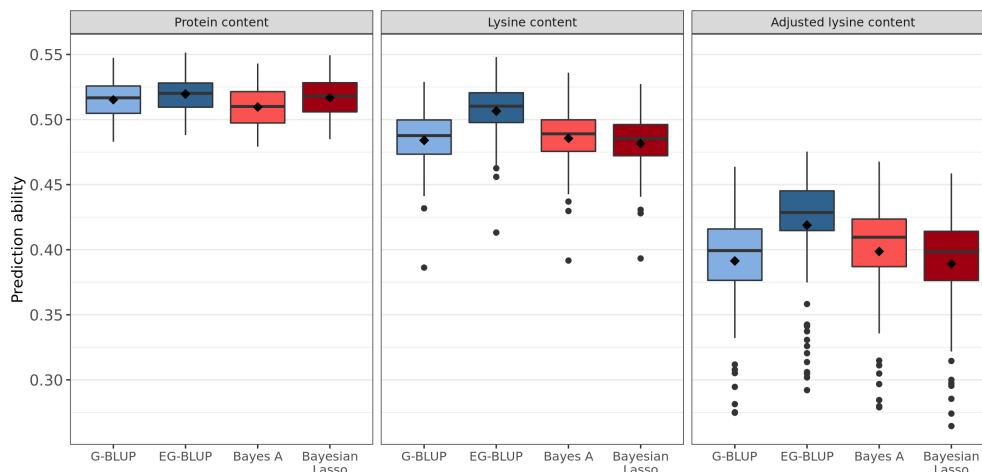


FIGURE 3

Distribution of genomic prediction abilities estimated in 100 five-fold cross-validation runs for protein, lysine, and adjusted lysine content. Four genomic prediction models were considered: G-BLUP, EG-BLUP, Bayes A, and Bayesian Lasso. Boxes enclose 50% of the central data, including median (horizontal black bold line) and mean (black diamonds), while whiskers are $\pm 1.5 \times$ interquartile range and dots represent extreme values.

terms of average cross-validated prediction abilities, EG-BLUP was 5.05% superior than the best of the three other alternative approaches for the prediction of adjusted lysine content. In addition, two different approaches were compared for the prediction of adjusted lysine content (Figure S2). The separate prediction of lysine content, protein content and TGW in order to calculate the derived trait based on these predictions was marginally less accurate than using the derived trait for the genomic prediction.

3.4 Predicted phenotypes of 7,745 accession samples

Protein content, lysine content, and adjusted lysine content were predicted for 7,745 accession samples by applying EG-BLUP - the most-accurate prediction model in cross-validations. For all accession samples, predicted protein content, lysine content, and adjusted lysine content averaged 16.85%, 4.08‰, and 4.13‰, respectively, with associated standard deviations of 0.74%, 0.11‰, and 0.09‰, correspondingly (Figure 4). Some accessions had outstanding values for the three traits with highest predicted values of protein content, lysine content, and adjusted lysine content amounting to 20.92%, 4.8‰, and 4.64‰, respectively. Interestingly, we found few accessions that had high values for both protein content and adjusted lysine content.

For all three traits, the size of the training set was rather small compared with the test set. For example, information of only 329 accession samples was used in order to predict the adjusted lysine content for 7,416 accession samples (Figure 4). Interestingly, the mean and median were both lower in the test set compared with the

training set. This deviation was most dominant for the predicted protein content where the means were 17.31% and 16.82% in training set and test set, respectively.

3.5 Definition of promising donor genotypes

The newly explored information can be used to select germplasm for pre-breeding programs with yet unexploited genetic diversity. To motivate future germplasm usage, favorable accessions were preliminary selected based on culling levels for predicted protein content and predicted adjusted lysine content in parallel. For both traits, the more stringent threshold was set to 99.9% of the normal distribution and five accessions could preliminary be selected (Figure 5). The respective accessions had not only favorable predicted phenotypes; moreover, these accessions were also recorded with particularly high BLUEs for both traits. Thus, the prediction can be seen as a confirmation of the genetic superiority of these accessions. Additional 19 accessions were identified with a more relaxed threshold for the culling levels selection ($z = 0.99$) (Table S2).

3.6 Association of lysine with altitude

The five accessions which were preliminary selected based on the stringent culling levels selection originated from Nepal and Afghanistan (Table S2). A characteristic of these accessions is the high altitude of their collecting sites. The altitudes of the collecting sites ranged between 1,925 m and 2,975 m above sea level, as

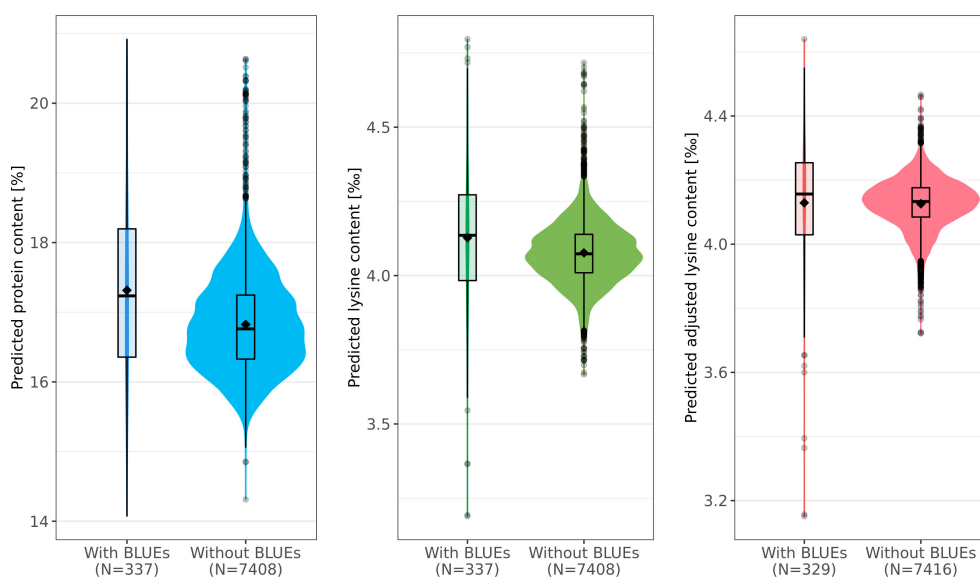


FIGURE 4

Predicted phenotypes of 7,745 wheat accession samples for protein (%), lysine (‰), and adjusted lysine (‰) content. For each trait, distributions of the prediction are shown separately for those accessions with best linear unbiased estimates (BLUEs, left) and those for which only genotypic data was present (right). Boxes enclose 50% of the central data, including median (horizontal black bold line) and mean (black diamonds), while whiskers are $\pm 1.5 \times$ interquartile range and dots represent extreme values.

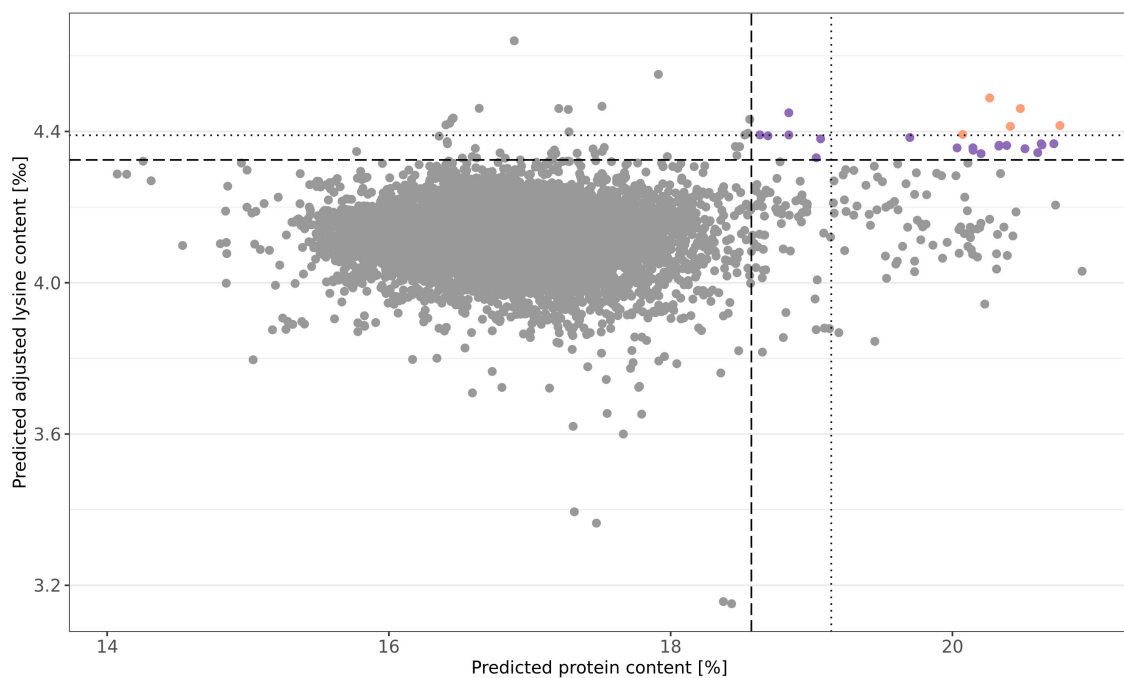


FIGURE 5

Culling levels selection based on genomic predicted protein (%) and adjusted lysine (‰) content. Shown are the predicted phenotypes of 7,745 wheat accession samples. Selection was performed with two intensities: a stringent threshold was defined by 0.999 of the normal distribution (dotted line) while a more relaxed selection threshold resulted from 0.99 of the normal distribution (dashed line). Orange and purple dots represent accessions samples which were selected based on the stringent and relaxed threshold, respectively.

indicated by the genebank catalog (Oppermann, 2023) and the description of the collecting expedition (Witcombe, 1975). The altitude of the collecting site is only known for 927 accession samples which are 12.12% of the examined winter wheat collection. The correlation between the altitude and the predicted traits was analyzed despite the incomplete data. Predicted protein and lysine contents were positively correlated with the altitude of the collecting site; the correlation coefficient amounted to 0.23 and 0.50, respectively ($p < 0.01$). The predicted adjusted lysine content was also positively correlated with the altitude of the collecting site ($r = 0.45$, $p < 0.01$) (Figure S3).

4 Discussion

The analysis of the present historical data was affected by their non-orthogonal structure. Most of the data trace back to the year 1970, with the vast majority of the accessions tested without replication. Lehmann and collaborators (1978) planned to screen all accessions once with the aim of identifying accessions with a strong positive deviation from the mean of the total collection. In order to reduce cost and workload, a repetition of the analysis was conducted only for accessions with high protein content ($> 18.5\%$) and lysine content ($> 4.5\%$). This strategy has two major disadvantages: First, the potential of some accessions may have been underestimated in a single testing year which would then have resulted in an erroneous rejection of this accession. Second, the total phenotypic variation cannot be estimated correctly without reliable

data from the underperforming accessions. However, precise estimates of variation are a prerequisite for optimized allocation of resources in breeding. On the one hand, this bias in the database for prediction should lower the quality of the prediction especially for the underperforming accessions (Zhao et al., 2012). On the other hand, this should be less relevant in our case because the selection decision in the first stage was based on data collected in one year coupled with the moderate plot-based heritabilities (Table 1). To account for the described shortcomings, we excluded accessions which were represented by only a single datapoint, as unreplicated data would provoke large uncertainty of the phenotypic data.

The exclusion of accessions with unreplicated data had strong consequences for the generated BLUEs. Trimming of the dataset reduced the number of accessions in the dataset to 11% and due to the multiple-stage testing strategy, resulted in higher mean values of protein content (15.92% to 17.66%) and lysine content (3.72‰ to 4.10‰). During the time of data collection, the multiple-stage evaluation resulted in optimized selection gain. From today's perspective, the shortcomings in the dataset however highlight the need to systematically plan screenings in a way which already consider the proper statistical evaluation. Especially with limited resources, repeated phenotyping of a well-chosen subset of accessions should be favored since missing phenotypic information can be determined by genomic predictions that rely on cheap genotyping of whole genebank collections (Yu et al., 2016). Given the selection strategy elaborated above, it was important to investigate whether selection decisions stood in the way of a representative training population. Inspection of the PCos

and distribution of phenotypic values suggests that we found this to be the case to a limited extent (Figure 2).

4.1 Strong associations between seed traits

The results showed a strong positive correlation between the BLUEs for protein and lysine content (Figure 1). An association of these traits has already been reported based on a large screening of the USDA World Wheat Collection (Vogel et al., 1975). The authors reported an even stronger correlation of 0.804 and 0.871 for the years 1972 and 1973, respectively. Furthermore, these authors reported a slightly negative correlation of TGW with protein content ($r = -0.278$) and lysine content ($r = -0.266$), respectively, in the year 1972. Thus, these correlations are in the same order of magnitude as the correlations found in the present study. In conclusion, these correlations indicate that accessions with a higher protein content do also have a higher overall lysine content. This is hardly surprising, because lysine is part of many groups of proteins even though not in equal abundance. Furthermore, lighter grains were identified to have an overall higher protein and lysine content. Arguably, this is due to the heterogeneous distribution of both components in the wheat grain. The storage proteins in the endosperm have a significantly lower lysine content than the embryo and bran (Vogel et al., 1976). In line with this, the relative lysine content of wheat grains decreases during the grain filling and maturation of the seed, thus, when the endosperm increases in size (Molino et al., 1988). Arguably, the fraction of the endosperm on the whole grain is larger in accessions with a high TGW. This suggests that the proportion of tissues with low lysine content increases in heavier grains. On the other hand, Vogel and collaborators (1976) also found a strong correlation of 0.91 between the lysine content of the endosperm and of the whole grain and concluded that the whole grain trait values are sufficiently reliable for selection. The data set analyzed in the present study was generated based on whole grain samples (Lehmann et al., 1978), which thus represent a mixture of embryo, bran, and endosperm. Unfortunately, modern milling processes white flour which contains exclusively endosperm tissue (Yu and Tian, 2018). Therefore, accessions with beneficial characteristics could hypothetically rely on an elevation of the lysine content in seed tissues rarely used in human nutrition. In this regard, the distribution of amino acids should be further investigated in the future, especially in outperforming accessions. In the present study, the intention was to identify genotypes that have a high proportion of lysine in the protein fraction independently of the seed size; thus, the adjustment of lysine content was important to account for the described associations.

4.2 EG-BLUP with high potential for genomic prediction

The comparison of genomic prediction models has shown that EG-BLUP outperforms the other models in terms of prediction ability (Figure 3). This is consistent with the findings of previous

genomic prediction studies in wheat. EG-BLUP resulted in more accurate predictions compared with G-BLUP for the prediction of TGW, plant height, and yellow rust resistance (Berkner et al., 2022). The particular advantage of this model is its ability to account for additive effect but also for additive-by-additive epistasis (Jiang and Reif, 2015). These results highlight therefore the importance of additive-by-additive epistasis and are thus in line with previous finding in wheat (Jiang et al., 2017; Raffo et al., 2022). The superiority across many different traits, demonstrates the robustness of the EG-BLUP model when confronted with different genetic architectures.

According to our cross-validated comparison, genomic prediction of derived traits such as adjusted lysine content can be performed accurately (Figure S2), but requires however some careful attention. In the present case, the implemented adjustment method relied on the availability of phenotypic values for three traits per accession. This restriction reduces the size of the training set and thus, the information which can be used for genomic prediction in a multiple-trait context (Schulthess et al., 2016). Moreover, the adjustment method relies on associations between lysine content and the two associated traits. These associations are, however, only valid for the examined set of genotypes and can differ between subsets of accessions such as for region-specific subpopulations. Even though the prediction of the derived trait based on the adjusted lysine content itself was most accurate (Figure S2), it might be more appropriate to predict the basis traits separately if subsampling is planned later, if the traits are biased by subpopulations, or if the availability of data is very unbalanced across different traits.

4.3 Enrichment of the genebank catalog facilitates new strategies for breeding programs

The study presents three types of data, namely, curated raw data, BLUEs, and predicted phenotypes for interested stakeholders in breeding and research. Without any doubt, the estimated (BLUEs) and genomic predicted phenotypic performance can be used for targeted selection of accessions. Although all the above-mentioned data has now become publicly available, we wanted to examine specifically accessions with high protein content and adjusted lysine content. With high selection intensities (culling levels of $z = 0.999$), we selected five promising accessions: While one of the preliminary selected accessions came from Afghanistan, four accessions originate in the Arun valley in Nepal (Oppermann, 2023). The latter ones derived from a collecting expedition in 1971. Considering that the collecting sites of all four accessions were located in neighboring villages (Witcombe, 1975), they arguably share one common mechanism of upregulated synthesis and storage of protein and lysine. At reduced selection intensity with a culling level of $z = 0.99$, 19 additional accession were identified and 14 of these originate from the very same expedition to Nepal. Witcombe and Rao (1976) evaluated the accessions of that collecting journey based on 39 traits and clustered plant material based on phenotypic characteristics but ignoring the geographic

proximity of the collecting sites. Interestingly, 14 accessions from the preliminary selected 18 accessions with Nepalese origin derived from the same phenotypic cluster. Witcombe and Rao (1976) identified the high altitude as one factor which leads to the common characteristics of this cluster.

4.4 Potential rationales for lysine's association with high altitudes

Adjusted lysine content was associated with the altitude of the collecting sites of accessions. On the one hand, it could be argued that the clustering of accessions with high adjusted lysine content collected at high altitudes in Nepal was caused by a spontaneous and rare mutation unrelated to selective advantages. On the other hand, high adjusted lysine content was significantly associated ($r = 0.45$; $p < 0.01$) (Figure S3) with the altitude of the collecting site in our study for a larger sample of 927 accessions for which altitude information of the collecting site was available. We thus advance the hypothesis that lysine content could play a role in the adaptation to the altitude at which these landraces have been grown continuously for many cropping seasons. High-altitude environments share common features, such as low temperatures, strong exposure to wind, drought due to lower humidity, high ultraviolet radiation, and hypoxia (Tranquillini, 1963; Jinqu et al., 2021). All of these characteristics can cause abiotic stress to plants, but only the latter two are specific to high altitudes. Therefore, high lysine content could be relevant for the adaptation to high ultraviolet radiation or hypoxia.

The involvement of lysine in the tolerance to various abiotic stresses, such as drought and salinity, has been summarized by Kishor and collaborators (2020). Yadav and collaborators (2019) investigated the impact of drought on the metabolite profile of wheat plants in glasshouse experiments. Drought-resilient wheat lines showed an increase in the lysine content of the vegetative tissue as well as impacts on other amino acids such as serin and asparagin. Additionally, Ding and collaborators (2016) reported a more than twofold increase in lysine content under hypoxic conditions in seedlings of rice (*Oryza sativa* L.). In contrast to this finding, the high adjusted lysine content in the present study is however not just a reaction of the wheat plant to abiotic stress. The high adjusted lysine content reflects a permanent adaptation which also results in higher lysine contents when these accessions are not facing the stresses of high altitudes. The present study relies on field trials conducted at 110 m above sea level. Moreover, the current data reflects the lysine content of mature seeds but not of vegetative plant tissue such as seedling. For seed tissue, we can only speculate about a possible interplay of lysine with stresses such as ultraviolet radiation or hypoxia.

Abiotic stress resistance of seeds is thought to be partially mediated via proteins of the late embryogenesis abundant (LEA) protein families (Zan et al., 2020). For instance, drought, extreme temperatures and ultraviolet radiation are associated with the expression of members of this protein family (Wang et al., 2008; Zan et al., 2020). In wheat, Liu and collaborators (2019) identified 179 genes encoding such proteins in the genome of var. Chinese

Spring. These proteins cluster into eight groups with distinct characteristics. Dehydrins are one of these groups and their protective characteristics relies on the K-segment which is specifically enriched in lysine (Yang et al., 2015; Zan et al., 2020). In line with this, Bhattacharya and collaborators (2019) found that the amino acid composition of LEA proteins in wheat can largely rely on lysine. In the case of one analyzed protein, lysine accounted for more than one quarter of all amino acids. If abundance of such proteins is causal for high lysine contents remain however speculation and this urges the need for further investigation.

The present study has not only outlined a strategy to mine historical data but also to leverage the data by genomic prediction. Moreover, this study equipped breeders and researchers with data for protein, lysine, and adjusted lysine content of in total 7,651 accession which can serve breeders to select suitable accessions for their pre-breeding programs. This might build the starting point of varieties which are not just high in protein but which further have a more favorable composition of amino acids and might help to overcome protein-energy malnutrition in future.

Data availability statement

This study comprises the publication of three different types of information. These are namely the raw data in ISA-Tab format, the R code for the calculation of BLUEs and genomic prediction with all input files, and the most important output files of the analysis. The output files include BLUEs of protein and lysine content as well as the predictions of protein content, lysine content and adjusted lysine content. The aforementioned information is available via the e!DAL (Arend et al., 2014) online repository (<https://dx.doi.org/10.5447/ipk/2023/20>).

Author contributions

MB: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. SW: Data curation, Writing – review & editing. JR: Conceptualization, Methodology, Supervision, Writing – review & editing. AS: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the German Federal Ministry of Education and Research as part of the Project GeneBank2.0 [grant no. FKZ031B0184A to AS] and by the AGENT project that is financed by the European Union's Horizon 2020 research and innovation program [grant agreement no. 862613 to MB]. Open access publishing received financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [grant no. 491250510].

Acknowledgments

This paper is dedicated to Andreas Graner, who recognized very early the value of the historical quality data of the IPK wheat collection according to the motto “the value of a collection increases with the associated information density” and thus initiated a pillar to transform the IPK Genebank into a bio-digital resource center. This research is available as a preprint on bioRxiv (Berkner et al., 2023).

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.

References

- Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D., et al. (2014). e!DAL - a framework to store, share and publish research data. *BMC Bioinform.* 15, 214. doi: 10.1186/1471-2105-15-214
- Batool, R., Butt, M. S., Sultan, M. T., Saeed, F., and Naz, R. (2015). Protein–energy malnutrition: A risk factor for various ailments. *Crit. Rev. Food Sci. Nutr.* 55, 242–253. doi: 10.1080/10408398.2011.651543
- Berkner, M. O., Schulthess, A. W., Zhao, Y., Jiang, Y., Oppermann, M., and Reif, J. C. (2022). Choosing the right tool: Leveraging of plant genetic resources in wheat (*Triticum aestivum* L.) benefits from selection of a suitable genomic prediction model. *Theor. Appl. Genet.* 135, 4391–4407. doi: 10.1007/s00122-022-04227-4
- Berkner, M. O., Weise, S., Reif, J. C., and Schulthess, A. W. (2023). Genomic unveiling of the diversity in grain protein and lysine content throughout a genebank collection of winter wheat. *bioRxiv*. [Preprint]. doi: 10.1101/2023.07.05.547805
- Bhattacharya, S., Dhar, S., Banerjee, A., and Ray, S. (2019). Structural, functional, and evolutionary analysis of late embryogenesis abundant proteins (LEA) in *Triticum aestivum*: A detailed molecular level biochemistry using in silico approach. *Comput. Biol. Chem.* 82, 9–24. doi: 10.1016/j.compbiolchem.2019.06.005
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2018). “ASReml-R reference manual version 4,” (Hemel Hempstead: VSN International Ltd).
- Ding, J., Yang, T., Feng, H., Dong, M., Slavin, M., Xiong, S., et al. (2016). Enhancing contents of γ -aminobutyric acid (GABA) and other micronutrients in dehulled rice during germination under normoxic and hypoxic conditions. *J. Agric. Food Chem.* 64, 1094–1102. doi: 10.1016/j.jultsonch.2017.08.029
- Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., and Braun, H.-J. (2022). “Chapter 4 global trends in wheat production, consumption and trade” in *Wheat Improvement*. Eds. M. P. Reynolds and H.-J. Braun (Cham: Springer), 47–66.
- FAO (2010). “Chapter 3 The state of ex situ conservation,” in *The second report on the state of the world's plant genetic resources for food and agriculture* (Rome: FAO), 53–90.
- Gonzalez, M. Y., Zhao, Y., Jiang, Y., Stein, N., Habekuss, A., Reif, J. C., et al. (2021). Genomic prediction models trained with historical records enable populating the German ex situ genebank bio-digital resource center of barley (*Hordeum* sp.) with information on resistances to soilborne barley mosaic viruses. *Theor. Appl. Genet.* 134, 2181–2196. doi: 10.1007/s00122-021-03815-0
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325. doi: 10.2307/2333639
- Henderson, C. R. (1985). Best linear unbiased prediction using relationship matrices derived from selected base populations. *J. Dairy Sci.* 68, 443–448. doi: 10.3168/jds.S0022-0302(85)80843-2
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.
- Hussain, T., Abbas, S., Khan, M. A., and Scrimshaw, N. S. (2004). Lysine fortification of wheat flour improves selected indices of the nutritional status of predominantly cereal-eating families in Pakistan. *Food Nutr. Bull.* 25, 114–122. doi: 10.1177/156482650402500202
- IWGSC (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, 1–13. doi: 10.1126/science.aar7191
- Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907
- Jiang, Y., Schmidt, R. H., Zhao, Y., and Reif, J. C. (2017). A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* 49, 1741–1746. doi: 10.1038/ng.3974
- Jinqiu, Y., Bing, L., Tingting, S., Jinglei, H., Zelai, K. L., Lu, L., et al. (2021). Integrated physiological and transcriptomic analyses responses to altitude stress in oat (*Avena sativa* L.). *Front. Genet.* 12. doi: 10.3389/fgene.2021.638683
- Kishor, P. B. K., Suravajhala, R., Rajasheker, G., Marka, N., Shridhar, K. K., Dhulala, D., et al. (2020). Lysine, lysine-rich, serine, and serine-rich proteins: link between metabolism, development, and abiotic stress tolerance and the role of ncRNAs in their regulation. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.546213
- Lawrence, J. M. (1976). Environmental influences on wheat lysine content. *J. Agric. Food Chem.* 24, 356–358. doi: 10.1021/jf60204a064
- Lehmann, C. O., Rudolph, A., Hammer, K., Meister, A., Müntz, K., and Scholz, F. (1978). Eiweißuntersuchungen am Getreide- und Leguminosen-Sortiment Gatersleben - Teil 1: Gehalt an Rohprotein und Lysin von Weizen sowie von Weizen-Art- und -Gattungsbastarden. *Die Kulturpflanze* 26, 133–161. doi: 10.1007/BF02146158
- Leinonen, I., Iannetta, P. P. M., Rees, R. M., Russell, W., Watson, C., and Barnes, A. P. (2019). Lysine supply is a critical factor in achieving sustainable global protein economy. *Front. Sustain. Food Syst.* 3. doi: 10.3389/fsufs.2019.00027
- Liu, H., Xing, M., Yang, W., Mu, X., Wang, X., Lu, F., et al. (2019). Genome-wide identification of and functional insights into the late embryogenesis abundant (LEA) gene family in bread wheat (*Triticum aestivum*). *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-49759-w
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genet.* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819
- Molino, I. M. M., Rojo, B., Martinez-Carrasco, R., and Pérez, P. (1988). Amino acid composition of wheat grain. 1: changes during development. *J. Sci. Food Agric.* 42, 29–37. doi: 10.1002/jsfa.2740420105
- Meybodi, N. M., Mirmoghtadaie, L., Sheidaei, Z., and Mortazavian, A. M. (2019). Wheat bread: potential approach to fortify its lysine content. *Curr. Nutr. Food Sci.* 15, 1–8. doi: 10.1038/s41598-019-49759-w
- Morton, K. J., Jia, S., Zhang, C., and Holding, D. R. (2016). Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins. *J. Exp. Bot.* 67, 1381–1396. doi: 10.1093/jxb/erv532
- Muleya, M., Tang, K., Broadley, M. R., Salter, A. M., and Joy, E. J. M. (2022). Limited supply of protein and lysine is prevalent among the poorest households in Malawi and exacerbated by low protein quality. *Nutrients* 14, 1–12. doi: 10.3390/nu14122430
- Müntz, K., and Lehmann, C. O. (1987). Reserveproteinforschung und genbank. *Die Kulturpflanze* 35, 25–52. doi: 10.1007/BF02163328
- Nobre, J. S., and Singer, J. M. (2011). Leverage analysis for linear mixed models. *J. Appl. Stat.* 38, 1063–1072. doi: 10.1080/02664761003759016
- OECD/FAO (2021) *Data about the production and use of wheat, maize, and soybean*. Available at: https://stats.oecd.org/Index.aspx?DataSetCode=HIGH_AGLINK_2021 (Accessed January 18, 2023).
- Oppermann, M. (2023). *Data from: IPK genebank accessions passport data snapshot 2023-02-17* (e!DAL - Plant Genomics & Phenomics Research Data Repository). doi: 10.5447/ipk/2023/6
- Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337
- Pellet, P. L., and Ghosh, S. (2004). Lysine fortification: Past, present, and future. *Food Nutr. Bull.* 25, 107–113. doi: 10.1177/156482650402500201

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.2023.1270298/full#supplementary-material>

- Pérez, P., and De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442
- Philipp, N., Weise, S., Oppermann, M., Börner, A., Graner, A., Keilwagen, J., et al. (2018). Leveraging the use of historical data gathered during seed regeneration of an ex situ genebank collection of wheat. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00609
- Philipp, N., Weise, S., Oppermann, M., Börner, A., Keilwagen, J., Kilian, B., et al. (2019). Historical phenotypic data from seven decades of seed regeneration in a wheat ex situ collection. *Sci. Data* 6, 1–9. doi: 10.1038/s41597-019-0146-y
- Raffo, M. A., Sarup, P., Guo, X., Liu, H., Andersen, J. R., Orabi, J., et al. (2022). Improvement of genomic prediction in advanced wheat breeding lines by including additive-by-additive epistasis. *Theor. Appl. Genet.* 135, 965–978. doi: 10.1007/s00122-021-04009-4
- Ritchie, H., Reay, D. S., and Higgins, P. (2018). Beyond calories: A holistic assessment of the global food system. *Front. Sustain. Food Syst.* 2. doi: 10.3389/fsufs.2018.00057
- Rogers, J. S. (1972). "Measures of genetic similarity and genetic distance," in *Studies in Genetics VII*, ed. Wheeler, M. (Austin, TX: University of Texas), 145–153.
- Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18404-w
- Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., et al. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nat. Genet.* 54, 1544–1552. doi: 10.1038/s41588-022-01189-7
- Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., and Zhao, Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor. Appl. Genet.* 129, 273–287. doi: 10.1007/s00122-015-2626-6
- Shewry, P. R. (2009). Wheat. *J. Exp. Bot.* 60, 1537–1553. doi: 10.1093/jxb/erp058
- Shewry, P. R., and Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food Energy Secur.* 4, 178–202. doi: 10.1002/fes3.64
- Tranquillini, W. (1963). The physiology of plants at high altitudes. *Annu. Rev. Plant Physiol.* 15, 345–362. doi: 10.1146/annurev.pp.15.060164.002021
- Ufaz, S., and Galili, G. (2008). Improving the content of essential amino acids in crop plants: goals and opportunities. *Plant Physiol.* 147, 954–961. doi: 10.1104/pp.108.118091
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1973). Results of systematic analyses for protein and lysine composition of common wheats (*Triticum aestivum* L.) in the USDA World Collection. *Historical Res. Bulletins Nebraska Agric. Experiment Station* 258, 1–27.
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1975). Reevaluation of common wheat from the USDA world wheat collection for protein and lysine content. *Historical Res. Bulletins Nebraska Agric. Experiment Station* 272, 1–36.
- Vogel, K. P., Johnson, V. A., and Mattern, P. J. (1976). Protein and lysine content of grain, endosperm, and bran of wheats from the USDA world wheat collection. *Crop Sci.* 16, 655–660. doi: 10.2135/cropsci1976.0011183X001600050014x
- Wang, B., Wang, Y., Zhang, D., Li, H., and Yang, C. (2008). Verification of the resistance of a LEA gene from *Tamarix* expression in *Saccharomyces cerevisiae* to abiotic stresses. *J. For. Res.* 19, 58–62. doi: 10.1007/s11676-008-0010-y
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18
- Witcombe, J. R. (1975). *University College Bangor NEPAL Expedition 1971* (Bangor: University College of North Wales).
- Witcombe, J. R., and Rao, A. R. (1976). The genecology of wheat in a Nepalese centre of diversity. *J. Appl. Ecol.* 13, 915–924. doi: 10.2307/2402266
- Yadav, A. K., Carroll, A. J., Estavillo, G. M., Rebetzke, G. J., and Pogson, B. J. (2019). Wheat drought tolerance in the field is predicted by amino acid responses to glasshouse-imposed drought. *J. Exp. Bot.* 70, 4931–4948. doi: 10.1093/jxb/erz224
- Yang, W., Zhang, L., Lv, H., Li, H., Zhang, Y., Xu, Y., et al. (2015). The K-segments of wheat dehydrin WZY2 are essential for its protective functions under temperature stress. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00406
- Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., et al. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants* 2, 1–7. doi: 10.1038/nplants.2016.150
- Yu, S., and Tian, L. (2018). Breeding major cereal grains through the lens of nutrition sensitivity. *Mol. Plant* 11, 23–30. doi: 10.1016/j.molp.2017.08.006
- Zan, T., Li, L., Li, J., Zhang, L., and Li, X. (2020). Genome-wide identification and characterization of late embryogenesis abundant protein-encoding gene family in wheat: Evolution and expression profiles during development and stress. *Gene* 736, 1–14. doi: 10.1016/j.gene.2020.144422
- Zhao, Y., Gowda, M., Longin, F. H., Würschum, T., Ranc, N., and Reif, J. C. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theor. Appl. Genet.* 125, 707–713. doi: 10.1007/s00122-012-1862-2



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Hanna Bolibok-Bragoszewska,
Warsaw University of Life Sciences, Poland
Javaid Akhter Bhat,
Nanjing Agricultural University, China

*CORRESPONDENCE

Monica Carvajal-Yepes

✉ m.carvajal@cgjar.org

Peter Wenzl

✉ p.wenzl@cgjar.org

RECEIVED 14 November 2023

ACCEPTED 29 December 2023

PUBLISHED 18 January 2024


CITATION

Carvajal-Yepes M, Ospina JA, Aranzales E,
Velez-Tobon M, Correa Abondano M,
Manrique-Carpintero NC and Wenzl P (2024)
Identifying genetically redundant accessions
in the world's largest cassava collection.
Front. Plant Sci. 14:1338377.
doi: 10.3389/fpls.2023.1338377

COPYRIGHT

© 2024 Carvajal-Yepes, Ospina,
Aranzales, Velez-Tobon, Correa Abondano,
Manrique-Carpintero and Wenzl. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/)
(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.

Identifying genetically redundant accessions in the world's largest cassava collection

Monica Carvajal-Yepes*, Jessica A. Ospina, Ericson Aranzales,
Monica Velez-Tobon, Miguel Correa Abondano ,
Norma Constanza Manrique-Carpintero and Peter Wenzl*

Genetic Resources Program, Alliance Bioversity International and International Center for Tropical Agriculture (CIAT), Cali, Colombia

Crop diversity conserved in genebanks facilitates the development of superior varieties, improving yields, nutrition, adaptation to climate change and resilience against pests and diseases. Cassava (*Manihot esculenta*) plays a vital role in providing carbohydrates to approximately 500 million people in Africa and other continents. The International Center for Tropical Agriculture (CIAT) conserves the largest global cassava collection, housing 5,963 accessions of cultivated cassava and wild relatives within its genebank. Efficient genebank management requires identifying and eliminating genetic redundancy within collections. In this study, we optimized the identification of genetic redundancy in CIAT's cassava genebank, applying empirical distance thresholds, and using two types of molecular markers (single-nucleotide polymorphism (SNP) and SilicoDArT) on 5,302 *Manihot esculenta* accessions. A series of quality filters were applied to select the most informative and high-quality markers and to exclude low-quality DNA samples. The analysis identified a total of 2,518 and 2,526 (47 percent) distinct genotypes represented by 1 to 87 accessions each, using SNP or SilicoDArT markers, respectively. A total of 2,776 (SNP) and 2,785 (SilicoDArT) accessions were part of accession clusters with up to 87 accessions. Comparing passport and historical characterization data, such as pulp color and leaf characteristic, we reviewed clusters of genetically redundant accessions. This study provides valuable guidance to genebank curators in defining minimum genetic-distance thresholds to assess redundancy within collections. It aids in identifying a subset of genetically distinct accessions, prioritizing collection management activities such as cryopreservation and provides insights for follow-up studies in the field, potentially leading to removal of duplicate accessions.

KEYWORDS

cassava, genebank, genetic redundancy, curators, diversity

Abbreviations: CIAT, International Center for Tropical Agriculture; QC, Quality control; IBS, Identity-By-State; MLGs, Multilocus genotypes.

1 Introduction

In the 1960s, plant scientists began coordinated collection and conservation efforts to reverse the decline of traditional landraces of essential food crops associated with the Green Revolution. After more than six decades, they have preserved over seven million germplasm accessions from more than 16,500 plant species (FAO, 2010). These invaluable genetic resources are now safeguarded in 1,750 genebanks worldwide (Engels, 2003; FAO, 2010). Most of these plant samples are conserved away from their natural habitats, forming ex-situ collections (Cohen et al., 1991). These collections play a crucial role in developing superior crop varieties, enhancing yields, improving nutrition, adapting to climate change, and bolstering resilience against pests and diseases (Westengen et al., 2018; Sheat et al., 2019; Baptista et al., 2022; Mba and Ogonnaya, 2022). They therefore contribute to enhancing agriculture, bolstering food security, and sustaining livelihoods.

Cassava (*Manihot esculenta* Crantz) is believed to have been domesticated in the Amazon Basin over 6,000 years ago (Olsen and Schaal, 1999; Olsen, 2004; Carvalho et al., 2017). Cassava provides nutrition to more than 500 million people in Africa and other continents around the world (Howeler et al., 2013). Presently, there are more than 13,832 cassava accessions conserved ex situ in genebanks across at least nine countries globally. Colombia, Brazil, and Nigeria are the countries conserving the largest collections with 5,963, 3,620, and 3,234 accessions, respectively (genesys-pgr.org/, Genesys, 2023). Colombia hosts an international genebank at the International Center for Tropical Agriculture (CIAT) that conserves the world's largest *in-vitro* collection of cassava and its wild relatives. The collection consists of 5,577 accessions of the cultivated species and 386 wild relatives belonging to 23 *Manihot* species from south-western North America to northern Argentina (Allem, 1994; Duputié et al., 2011). The cultivated species can be propagated either by seed or vegetatively from stems, the latter being the most common practice for commercial production (Howeler et al., 2013). Vegetative propagation, ensures that the new plants are genetically identical to the parent plant, preserving desired traits such as disease resistance, high yield, and nutritional quality in the next generation (ibid).

In most crop species, true seeds serve as the primary method for germplasm conservation (Engels, 2003). However, cassava is a highly heterozygous and clonally propagated crop and cannot rely on true seeds for conservation efforts (Mafla et al., 1993; Qi et al., 2022). Until the 1990s, CIAT's cassava germplasm collection was maintained both in the field and *in-vitro* (Mafla et al., 1993). However, due to increasing difficulties in managing the field collection, only the *in-vitro* collection was retained under slow-growth conditions (Hershey, 2008). *In-vitro* cultures serve as sources of disease-free materials for distribution, multiplication, and as explants for cryo-preservation (FAO, 2014).

Conserving the cassava collection *in-vitro* is costly. A 2011 study estimated a conservation cost of US\$71 per accession, which in 2023 would be equivalent to US\$97 (CGIAR Genebanks Consortium 2011). Considering that plant genetic resources for food and agriculture are conserved in perpetuity, identifying

genetically redundant accessions or potential duplicates within the collection is crucial for optimizing physical storage, reducing maintenance costs, enhancing characterization, and ensuring collections' accessibility and usability.

Over the past decades, various methods have been employed to differentiate between genotypes in genebank collections. Initially, biochemical markers were employed (Lefèvre and Charrier, 1992), followed by the use of molecular markers such as random amplified polymorphic DNA, amplified fragment length polymorphism, and microsatellites (Virk et al., 1995; Dean et al., 1999; Kisha and Cramer, 2011; Motilal et al., 2013). In recent years, the advent of high-throughput sequencing technologies has revolutionized the characterization of genebank crop collections, such as wheat, barley, bean, cassava, and rice (Nadeem et al., 2018; Ferguson et al., 2019; Milner et al., 2019; Sansaloni et al., 2020; Tanaka et al., 2021). These techniques examine genome-wide natural variation patterns (Orek et al., 2023) and enable the comprehensive assessment of genetic distinctness and redundancy across entire genomes (Fu, 2023; Orek et al., 2023). The genetic distinctness and redundancy of cassava clones grown by farmers or conserved in genebanks has been evaluated using a variety of methods, revealing genetic redundancy levels of 20–50 percent within and across collections (Chavarriaga-Aguirre et al., 1999; Albuquerque et al., 2019; Orek et al., 2023; Soro et al., 2023).

We hypothesize that establishing empirically defined genetic-distance thresholds will enable the effective identification of genetically redundant accessions within the vast cassava collection at CIAT. The aim of this study was thus to (a) empirically define genetic-distance thresholds to identify genetically redundant accessions and (b) identify genetically redundant accessions within the largest global cassava collection conserved at CIAT. The results of this study provide valuable insights for efficient collection management, while generating genetic characterization data that will enable a more targeted use of accessions supporting cassava crop improvement for the current and future crop challenges.

2 Methods

2.1 Plant materials and genebank accessions

A subset of 21 accessions were randomly selected from the CIAT cassava core collection (core collection defined by Hershey et al., 1994) conserved in Palmira, Colombia, to establish the thresholds for identifying genetically redundant accessions. Leaf tissue was collected from plantlets conserved *in vitro* under slow-growing conditions to extract DNA, generating various biological and technical replicates. These replicates included: (i) different individuals from the same accession obtained from different conservation units, referred to here as “Ind-Reps”; (ii) distinct DNA samples extracted from the same individual, referred to here as “Extract-Reps”, and (iii) the same DNA sample analyzed twice, labeled as “DNA-Reps” (Supplementary Figure 1). Among the 21 accessions, a total of 141 samples were analyzed, consisting of 84 samples from 42 pairs of DNA-Reps, 74 samples from 37 pairs of

Extract-Reps, and 62 samples from 21 trios of Ind-Reps (Supplementary Table 1). These accessions were originally collected from 12 countries (Brazil, Colombia, Cuba, Ecuador, Fiji, Guatemala, Mexico, Panama, Paraguay, Peru, Thailand, and USA). After evaluating these 21 accessions and their corresponding replicates, DNA extractions were performed on another 5,414 accessions from the cultivated cassava collection, originally from 28 different countries (Table 1). Among these accessions, 614 are declared as breeding lines while 4,800 are landraces. Including the 141 replicates, this study involved a total of 5,555 samples.

2.2 DNA extractions and genotyping

DNA extractions were conducted between 2016 and 2021 as funding resources became available. Approximately 10 mg of lyophilized leaf tissue, obtained from *in-vitro* plantlets was used for DNA extraction on 96 well plates. Samples were homogenized at 12,000 rpm for 1 min using the Geno/Grinder 2010 (Spex SamplePrep LLC, NJ) and subsequently lysed with CTAB extraction buffer (containing 2M NaCl, 0.25M EDTA pH 8.0, 0.1M Tris-HCl pH 8.0, 2% CTAB, 2% PVP, and 0.2% 2-Mercaptoethanol) (Dellaporta et al., 1983). After mixing by vortex, samples were incubated at 65°C for 30 min, followed by the addition of an equal volume of chloroform:Isoamyl alcohol 24:1 (Sigma-Aldrich, USA). The resulting mixture was carefully mixed for 5 min and then centrifuged for 20 min at 3,000 rpm at 4°C. The aqueous phase was transferred to another tube and mixed with equal volume of cold isopropanol, followed by an incubation at -20°C for 1 h. After incubation, samples were centrifuged for 20 min at 3,000 rpm at 4°C. Upon removal of the supernatant, the pellets were washed with 80% cold ethanol and centrifuged again. Pellets were air-dried and resuspended in TE buffer (pH 8.0, Alpha Teknova, USA) with 40 µg of RNase (QIAGEN, Germany). The samples were then incubated at 37°C for 30 min and stored at -20°C. Samples were handled using multichannel pipettes. The concentration and purity of DNA was estimated by calculating the absorbance at 260/280 nm using Bio Teak Synergy H1m (Agilent Technologies, USA) and quality was assessed on a 0.8% agarose gel.

Fifty microliters of genomic DNA, with a concentration of 50 ng/µl for each sample, were submitted to Diversity Array Technology in Canberra, Australia, for genotyping by sequencing. DArTseq™ technology was employed, utilizing a combination of *MseI* and *PstI* restriction enzymes to prepare the genomic representation and subsequent next-generation sequencing as described by Kilian et al. (2012). Marker identification and allele-calling were performed with DS14 software (Diversity Arrays Technology P/L). In this study, two types of markers were utilized: codominant SNP markers and presence/absence dominant SilicoDArT markers.

2.3 Filters for high-quality marker and high-quality sample selection

A series of quality filters were carefully reviewed and applied in this study to select the most informative and high-quality markers, while

excluding genomic representations prepared from low-quality DNA samples. For SNP markers, the selection process involved considering two parameters reflecting the markers' information content, minor allele frequency (maf), estimated as the frequency at which the less-common allele of a genetic variant occurs within a population, and call rate, estimated as the proportion of samples for which the genotype is called and there is no missing value (Hernandez et al., 2019). Additionally, three parameters related to the technical aspects of marker quality were used, AvgMarkerCount, CVMarkerCount, and RepAvg. Average marker count (AvgMarkerCount) denotes the average number of sequence-tag copies of a marker, calculated by averaging the mean number of sequence-tag copies of the two SNP alleles. CVMarkerCount represents the coefficient of variation of AvgMarkerCount, utilized to minimize the chance of erroneously matched paralogue alleles from different loci. Additionally, RepAvg — an estimate calculated by DArT P/L — assessed the proportion of technical replicate assay pairs for which the calls of a given marker were consistent. A final filter was applied for markers that mapped to the cassava reference genome v7.1 (Bredeson et al., 2016). These filters were sequentially applied in the following order: $\text{maf} \geq 0.001$, $\text{callrate} \geq 0.8$, $\text{AvgMarkerCount} \geq 12$, $\text{CVMarkerCount} \leq 0.6$, $\text{RepAvg} \geq 0.98$, and markers mapped to the reference genome v7.1.

For SilicoDArT dominant markers, a similar approach was used, with consideration given to two parameters reflecting the information content of markers call rate and OneRatio, and similar parameters related to the technical aspects of marker quality AvgReadDepth, CVReadDepth, and Reproducibility. OneRatio indicates the proportion of samples for which the genotype score was “1” (present). Average read depth (AvgReadDepth) denotes the average tag read count, calculated as the total sum of tag read counts across all samples divided by the number of samples' score as “1”. CVReadDepth, which represents the coefficient of variation of AvgReadDepth, was utilized to remove markers with high variability in the number of tag read counts, potentially reflecting issues during PCR amplification. The filters were applied consecutively in the following order: $\text{call rate} \geq 0.95$, $\text{OneRatio} \geq 0.05$, $\text{AvgReadDepth} \geq 12$, $\text{CVReadDepth} \leq 0.7$, and $\text{Reproducibility} \geq 0.98$. The impact of each filter on the number of markers and the genetic distances between pairs of replicates (DNA-Reps, Extract-Reps, and Ind-Reps) was thoroughly studied before settling on the selected values.

To eliminate low-quality samples that could potentially affect genetic-distance calculations, we assessed a range of parameters across 5,555 samples, including the 141 samples of the technical and biological replicates, and the 5,414 accessions from the cultivated cassava collection (Table 1). The parameters assessed included target quality control (QC), a categorical parameter provided by DArT P/L (Canberra, Australia), classifying library quality as “good”, “downshifted” or “weak”; total read count (tagcounttotal), unique read count (tagcountunique), individual SNP callrate, observed heterozygosity of individuals (Ho), individual SilicoDArT callrate, and individual SilicoDArT OneRatio, the latter represents the proportion of markers within one sample called as “1” (present). The total read count represents the total number of reads obtained per sample from sequencing each library. The three categories of target QC are evaluated on an agarose gel. A library categorized as ‘good’ exhibits DNA within the expected size

TABLE 1 Summary of the 5,414 accessions from the cultivated cassava collection used in this study.

Country of origin	Number of Breeders' Lines	Number of Landraces	Total number of accessions	Number acc. In core	Region
ARG	6	105	111	5	Eastern South America
BOL		7	7	3	Eastern South America
BRA	86	1158	1244	93	Eastern South America
CHN		2	2	2	Asia
COL	439	1804	2243	155	Western South America
CRI		96	96	15	Central/North America & Caribbean
CUB	2	76	78	18	Central/North America & Caribbean
DOM		4	4	4	Central/North America & Caribbean
ECU		105	105	28	Western South America
FJI		5	5	1	Asia
GTM		78	78	10	Central/North America & Caribbean
HND		35	35	1	Central/North America & Caribbean
IDN	21	218	239	6	Asia
JAM		21	21	2	Central/North America & Caribbean
MEX	4	95	99	18	Central/North America & Caribbean
MYS	8	55	63	13	Asia
NGA	18		18	3	Africa
NIC		4	4		Central/North America & Caribbean
PAN		48	48	8	Central/North America & Caribbean
PER		395	395	72	Western South America
PHL	2	4	6	2	Asia
PRI		16	16	8	Central/North America & Caribbean
PRY	2	198	200	38	Eastern South America
SLV		11	11		Central/North America & Caribbean
THA	26	9	35	3	Asia
unknown		1	1		NA
USA		7	7	2	Central/North America & Caribbean
VEN		234	234	52	Western South America
VNM		9	9		Asia
Total	614	4,800	5,414	562	

range. ‘Downshifted’ libraries show DNA on a gel with a broader size range, shifting to smaller sizes, and may be associated with predigested DNA due to poor DNA quality. Libraries categorized as ‘weak’ are likely to have poor amplification due to uneven DNA concentrations. Each of these parameters was plotted to identify a suitable threshold for selecting high-quality samples, considering the target QC of samples. The thresholds used for each parameter to retain samples were tagcounttotal > 1,500,000, tagcountunique > 230,000, individual SNP callrate > 0.73, Ho > 0.05 and < 0.16, individual SilicoDart callrate > 0.996 and, individual SilicoDart OneRatio > 0.2. Subsequent analysis excluded samples failing to meet these criteria ([Supplementary Table 2](#)).

2.4 Calculation of genetic distances for SNP and SilicoDart markers

For SNP markers, we calculated Identity-By-State (IBS) distances using the 1-IBS function in PLINK v1.0 ([Purcell et al., 2007](#)), released on June 29, 2007. The IBS calculation is based on counting alleles that are identical by state (IBS) between pairs of individuals at each genotyped marker. PLINK computes the IBS sharing proportion by comparing the number of shared alleles (IBS count) to the total non-missing alleles for each pair of individuals. This proportion is then subtracted from 1 to determine the distance between each pair of samples. IBS distances were independently calculated for several datasets. The first dataset comprised 141 samples, encompassing both technical and biological replicates from 21 accessions. IBS distances were calculated using these 141 samples and SNP marker sets before and after applying marker-quality filters, resulting in 22,840 to 7,001 SNPs (without excluding low-quality samples). The second dataset included a subset of 131 samples of technical and biological replicates, obtained after excluding 10 low-quality samples, and 6,987 SNP markers obtained after applying marker-quality filters. The third dataset consisted of 5,302 accessions, representing 95 percent of the cultivated cassava collection conserved at CIAT, and 7,180 SNP markers obtained after applying consecutive filters to the six marker quality parameters as described in section 2.3.

To calculate genetic distances for SilicoDart markers, we used the gl.dist.ind function with the Jaccard method from the dartR package v2.0.4 ([Gruber et al., 2018](#); [Mijangos et al., 2022](#)). The Jaccard distance matrix was calculated from the dataset of 131 samples, including technical and biological replicates, and 29,456 or 13,715 SilicoDart markers obtained before and after marker-quality filters, respectively. Additionally, the Jaccard distance was calculated for 5,302 samples and 8,186 SilicoDart markers after applying filters to marker quality parameters.

2.5 Genetic-distance threshold to identify genetically redundant accessions

To identify genetic distinctness and redundancy within the cassava collection, we initially estimated the average genetic distances among the three types of pairs of replicates (DNA-Reps, Extract-Reps, and Ind-Reps) from 21 accessions from the core collection. This step aimed

to evaluate the efficacy of selected marker-quality filters in retaining a high number of markers while achieving minimal (close to zero) genetic-distance estimates for replicate pairs. To assess the threshold’s ability to detect genetically redundant accessions, we utilized the mlg.filter function from the poppr package v2.9.3 ([Kamvar et al., 2014](#)), in R program v4.2.2 ([R Core Team, 2022](#)), following the methodology outlined by [Albuquerque et al. \(2019\)](#). This function collapses multilocus genotypes (MLGs) falling below a specific threshold based on genetic distance.

In our analysis using the SNP dataset, we conducted two tests from the 21 subset accessions: one test, comprised a dataset with 141 samples and 7,001 SNP markers. We empirically set the minimum genetic distance at 0.06, informed by the maximum observed IBS distance among replicates. A second test, with the 21 accessions, excluded 10 low-quality samples, resulting in a dataset of 131 samples and 6,987 SNP markers. Here, the minimum genetic distance was set at 0.015, determined from new estimations of average genetic distance between pairs of replicates after removing the 10 low-quality samples ([Supplementary Figure 2](#)). Subsequently, we performed the final analysis to detect the number of MLGs using the IBS distance matrix calculated across 5,302 accessions from the cultivated cassava collection and 7,180 SNP markers. Additionally, we explored the impact of varying the threshold for cluster identification, considering minimum IBS thresholds ranging from 0.000 to 0.06.

In parallel, we detected MLG using SilicoDart markers and Jaccard distance, using the 131 samples and 13,715 SilicoDart markers, exploring different thresholds (0.012 and 0.025). Furthermore, we detected MLG using the Jaccard distance, estimated for the larger dataset comprising 5,302 accessions and 8,186 SilicoDart markers.

2.6 Clustering analyses

We performed agglomerative clustering using the complete linkage method to assess and validate the MLGs identified across the 141 and 131 samples. Additionally, we conducted agglomerative clustering employing the ward.D2 linkage method with the 5,302 accessions to discern potential patterns of genetic redundancy within/across regions of origin, and to identify discrepancies between the results obtained with the two marker types ([Murtagh and Legendre, 2014](#)). The clustering algorithm utilized either IBS or Jaccard genetic distances and was implemented through the hclust function in the stats R package v4.2.2 ([R Core Team, 2022](#)). The resulting hierarchical clusters were visualized directly or transformed into Newick trees using the cluster-to-tree conversion function hc2Newick from the ctc R package v1.72.0. The Newick trees were customized and annotated using the Interactive Tree of Life (iTOL) online tool ([Letunic and Bork, 2021](#)).

2.7 Comparing and contrasting datasets and data types

To compare the results obtained from the MLG detection analysis using the two types of genetic markers and the derived

genetic distances (Jaccard and IBS distances), we generated Venn diagrams to compare the accessions detected as unique or as redundant within the MLGs detected. For this purpose, we utilized the `venn.diagram` function from the `VennDiagram` R package v1.7.3.

Additionally, a subset of 20 MLGs of varying sizes (collapsing 2, 3, 4, 6, 7, 12, 28, 29, 31, and 84 accessions) were used to review passport and historical characterization data with clusters. The reviewed passport data included: biological status (landrace or breeding germplasm), country and region of origin, common names, and collection date. The historical characterization variables included color root pulp, shape of central leaf, number of leaf lobes, petiole color, and color of the first expanded leaf. The historical characterization data and images were extracted from Genesys, 2023 (<https://www.genesys-pgr.org/datasets/>; <https://www.genesys-pgr.org/a/images/>).

3 Results

3.1 Defining empirical thresholds for detecting genetically redundant accessions

A total of 22,840 polymorphic SNPs were detected across 141 samples from 21 cassava accessions selected from the core collection, including their corresponding biological and technical replicates. The application of six consecutive filters to the marker-quality parameters reduced the number of markers from 22,840 to 7,001 SNPs. The estimated average genetic distance for the 42 DNA-Reps pairs decreased from 0.008 (± 0.004 SD) to 0.002 (± 0.003 SD) (see F0 to F6 in Table 2; Figure 1A). The average genetic distance of the 37 Extract-Rep and 21 Ind-Rep pairs remained relatively higher (0.004 ± 0.011 and 0.003 ± 0.012 , respectively; Table 2; Figure 1A). Because the highest genetic distance value for Ind-Reps was 0.0576, the initial minimum distance to distinguish genetically unique accessions within this validation set of 141 samples was set at 0.06. The MLG analysis detected 21 different MLGs, representing 18 genetically unique accessions (each with their respective technical and biological replicates), one group of two genetically redundant accessions with their technical and biological replicates (CUB74, PAN70; MLG 10, Table 3), and two groups with replicates of accession USA4 (MLGs 20 and 21, Table 3; Figure 1B). The division of USA4 samples into two distinct clusters points to potential issues with DNA sample quality.

3.2 Sample quality assessment

We assessed seven different quality parameters across 5,555 samples from the cultivated cassava collection, which included the 5,414 accessions described in Table 1, as well as 21 accessions along with their 141 technical and biological replicates (Supplementary Table 1). These parameters included target QC, `tagcounttotal`, `tagcountunique`, individual SNP call rate, Ho of individuals, individual SilicoDArT call rate, and individual SilicoDArT OneRatio. Among the 5,555 samples, 5,532 targets (commonly

known as sequenced DNA libraries) were categorized as “good”, 21 were classified as “downshifted”, and two as “weak”. The assessment of target QC was conducted by DArT P/L and was based on agarose gel evaluation (specific data not provided). The statistical distribution of other associated sample quality parameters revealed that the mean values of `tagcounttotal` and `tagcountunique` were 2,306,247 (ranging from 1,007,722 to 3,995,430) and 385,875 (ranging from 92,395 to 1,136,722) reads, respectively. Seven samples exhibited remarkably low levels of `tagcounttotal`, falling below 1,500,000 read counts, while 27 samples displayed `tagcountunique` values lower than 230,000 (Supplementary Table 2). Regarding sample quality of SNP-related parameters, such as call rate and Ho, the mean values were 0.83 (ranging from 0.11 to 0.91) and 0.09 (ranging from 0 to 0.2), respectively. Notably, most of the downshifted samples had call rates below 0.73 (a total of 72 samples), and/or Ho values either below 0.05 or above 0.16 (79 samples) (Supplementary Table 2; Figure 2).

The sample's SilicoDArT-related parameters, including call rate and OneRatio, had mean values of 0.97 (ranging from 0.86 to 1) and 0.27 (ranging from 0.04 to 0.35), respectively. Notably, 30 samples had a call rate above 0.996, and 21 samples exhibited OneRatio values below 0.2. By visually representing these parameters and employing distinct colors for the three target QC categories, we observed that a majority of the downshifted samples (18 samples) fell below or above the designated thresholds (Figure 2; Supplementary Table 2). In total, 101 samples fell below or above the threshold for at least one of these parameters, while 42 samples exhibited disparities in 2 to 6 parameters (Supplementary Table 3). A total of 143 samples out of the initial 5,555 were excluded from further analysis as they were identified with low-quality. Out of these 143, ten samples were drawn from the set of 141 biological and technical replicates (Supplementary Table 3). This resulted in a dataset comprising 131 samples and 6,987 SNP markers, which was used for establishing the genetic distance threshold and validating the detection of MLGs. Consequently, after removing low-quality samples a total of 5,302 accessions from the cultivated cassava genebank collection were retained for the MLG detection analysis.

3.3 Verification of minimum genetic distances for MLG detection after removing low-quality samples

Comparing the genetic distances among the dataset of 141 or 131 technical and biological replicates (before and after filtering samples), showed a significant reduction in the genetic distance variation within replicates of four accessions, COL148, CUB74, MEX86, and notably in the case of USA4 (Supplementary Figure 2). Consequently, this reduction in the genetic distance after removing low quality samples allowed us to set the minimum genetic distance for MLG detection to 0.015. Using this threshold and the dataset of 131 samples, we repeated the MLG detection procedure. This process validated the accuracy of the approach to identify groups of distinct genotypes for the replicates derived from each accession. The analysis identified 20 MLGs, as documented in Table 3,

TABLE 2 Combining multiple marker-quality thresholds.

Order	Parameter used SNPs	DNA-Reps							Extract-Reps	Ind-Reps
		F0	F1	F2	F3	F4	F5	F6	F6	F6
1	maf		≥ 0.001	≥ 0.001	≥ 0.001	≥ 0.001	≥ 0.001	≥ 0.001	≥ 0.001	≥ 0.001
2	Callrate Loc			≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8
3	AvgMarkerCount				≥ 12	≥ 12	≥ 12	≥ 12	≥ 12	≥ 12
4	CVMarkerCount					≤ 0.6	≤ 0.6	≤ 0.6	≤ 0.6	≤ 0.6
5	RepAvg						≥ 0.98	≥ 0.98	≥ 0.98	≥ 0.98
6	Mapped genome v7							yes	yes	yes
SNP	Number of SNPs	22,840	22,840	17,221	8,768	8,400	7,717	7,001	7,001	7,001
	mean dist	0.008	0.008	0.0077	0.0035	0.0029	0.0022	0.0022	0.0041	0.0037
	SD dist	0.0048	0.0048	0.0053	0.0041	0.0041	0.0036	0.0036	0.0113	0.0127
	max. dist	0.0212	0.0212	0.0043	0.0161	0.0157	0.0132	0.0127	0.0507	0.0576
	min. dist	0.0047	0.0047	0.021	0.0013	0.0006	0.0003	0.0002	0.0002	0.0004
Order	Parameter* used	DNA-Reps							Extract-Reps	Ind-Reps
		F0	F1	F2	F3	F4	F5		F5	F5
1	Callrate Loc		≥ 0.95	≥ 0.95	≥ 0.95	≥ 0.95	≥ 0.95		≥ 0.95	≥ 0.95
2	OneRatio			≥ 0.05	≥ 0.05	≥ 0.05	≥ 0.05		≥ 0.05	≥ 0.05
3	AvgReadDepth				≥ 12	≥ 12	≥ 12		≥ 12	≥ 12
4	CVReadDepth					≤ 0.7	≤ 0.7		≤ 0.7	≤ 0.7
5	Reproducibility						≥ 0.98		≥ 0.98	≥ 0.98
SilicoDART	Number of SilicoDART	29,456	26,015	23,674	16,664	14,089	13,715		13,715	13,715
	mean dist	0.0011	0.0008	0.0009	0.0004	0.0003	0.0003		0.0002	0.0006
	SD dist	0.0027	0.0024	0.0027	0.0017	0.0016	0.0014		0.0007	0.0022
	max. dist	0.0131	0.0117	0.0127	0.0078	0.0063	0.0063		0.0034	0.0094
	min. dist	0.0001	0.0000	0.0000	0.0000	0	0		0	0.0000

*Minor allele frequency (maf), frequency at which the less-common allele of a genetic variant occurs within a population; Call rate, estimated as the proportion of samples for which the genotype is call and there is no missing value; AvgMarkerCount, the average number of sequence-tag copies of a marker; CVMarkerCount, the coefficient of variation of AvgMarkerCount; RepAvg, the proportion of technical-replicate assay pairs for which the calls of a given marker were consistent; OneRatio, the proportion of samples for which the genotype score was "1" (present); AvgReadDepth, the average tag read count, calculated as the total sum of tag read counts across all samples divided by the number of samples score as "1"; CVReadDepth, the coefficient of variation of AvgReadDepth.

List of parameters for SNP and SilicoDART markers, order of application, starting from no filters applied (F0) to all filters applied (F6). Number of markers, the mean of genetic distance (IBS for SNPs and Jaccard for SilicoDART), standard deviation (SD) of pairs of DNA, Extract-Reps, and Ind-Reps.

collapsing all replicates from each accession into distinct MLGs, including replicates from USA4. As previously observed, two accessions (CUB74 and PAN70) were identified within the same MLG.

Moreover, we assessed the number of MLGs by calculating Jaccard genetic distance from the SilicoDART markers to the selected 131 high-quality samples. The initial number of SilicoDART markers for this set of samples was 29,456. Similarly, as with the SNPs, a series of marker quality parameters were assessed including call rate, OneRatio, AvgMarkerCount, CVMarkerCount and Reproducibility. Applying filters to these parameters reduced markers from 29,456 to 13,715 as shown in Table 2. The estimated average genetic distance of the 42 pairs of DNA-Reps decreased from 0.0011 ± 0.002 to 0.0003 ± 0.0014 (see F0 to F5 in Table 2; Figure 1A). Moreover, the mean genetic

distance for the 37 Extract-Rep and 21 Ind-Rep pairs measured 0.0002 ± 0.0007 and 0.0006 ± 0.0022 , respectively (Table 2; Figure 1A). When utilizing a minimum Jaccard genetic distance threshold of 0.012 for MLG detection, a total of 23 unique MLGs were detected. Notably, two of these MLGs (15 and 23) corresponded to independent MLGs, including separately technical replicates from the MEX86 accession. Similarly, another three MLGs (20, 21, and 22) were identified for different technical replicates of the USA-4 accession (Table 3). These results suggested that a higher threshold for Jaccard distances needs to be used to allow all replicates from each accession to collapse in the same MLGs.

By increasing the minimum Jaccard genetic distance to 0.025, samples were collapsed into 20 MLGs, similarly to the SNP markers when using a minimum IBS genetic distance threshold of 0.015

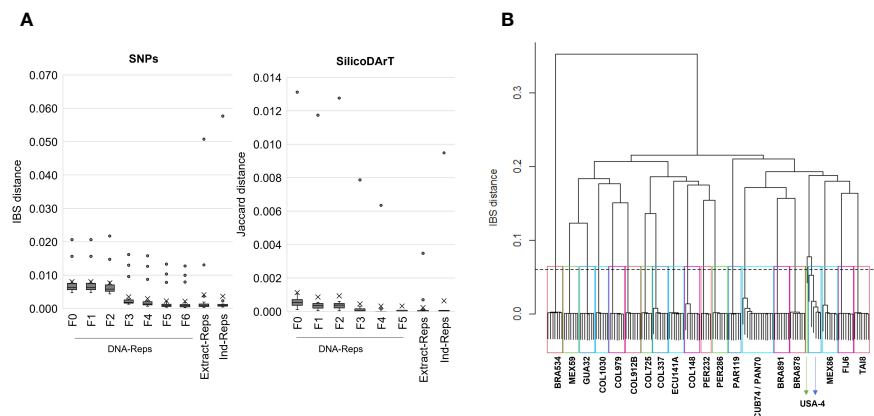


FIGURE 1

(A) The cumulative impact of implementing exemplary thresholds for each marker-quality parameter on genetic distances. The box plots display the distribution of the genetic distance of technical and biological replicates at various stages of filter application, from F0 (unfiltered data) to F6. F0 to F6 indicate distances of DNA-Rep pairs across filtering steps. Extract-Rep and Ind-Rep pairs distances were estimated with filters in F6 for SNP or F5 for SilicoDArT. (B) Dendrogram of hierarchical clustering using the “complete linkage” method and the IBS distance. Twenty-one rectangles highlight the 21 clusters formed under a genetic distance of 0.06 (dashed line). The IBS distance matrix was calculated from 141 samples and 7,001 SNP markers.

(Table 3). The hierarchical clustering analysis showed 20 clusters below a threshold of 0.015 for IBS and 0.025 Jaccard distances (Figure 3). Notably, all technical replicates from each accession clustered together, including those from MEX86 and USA4. Consistently, two accessions were grouped together (CUB74/PAN70), highlighting an instance of redundant accessions within the dataset.

3.4 Assessing genetic redundancy within the cultivated cassava genebank collection

We utilized a validated approach, carefully selecting high-quality samples and markers. We also established a minimum genetic distance threshold for IBS and Jaccard distances, allowing us to distinguish distinct and redundant accessions. With this approach, we evaluated genetic redundancy in 5,302 accessions, covering 95 percent of CIAT’s cultivated cassava collection and 88 percent of the entire collection, as shown in Table 1. Initially, we had 33,395 polymorphic SNPs and 39,103 SilicoDArTs. After subjecting these markers to quality assessments (refer to Supplementary Figure 3) and applying the same filtering thresholds as for the replicates’ dataset, we ended up with a refined set of 7,180 SNP and 8,186 SilicoDArT markers (see Supplementary Table 4). These markers were distributed across the 18 chromosomes of the cassava reference genome, maintaining proportional representation across chromosomes before and after filtering markers, as illustrated in Figure 4A. Initially, 29,040 unfiltered SNPs were mapped across the reference genome, while 4,355 remained unmapped. Following the application of filters, we obtained 7,180 SNPs mapped to the reference genome. In the case of SilicoDArTs, 26,932 unfiltered markers were mapped, leaving 12,171 unmapped. In the final filtering process, 5,883 SilicoDArT markers were mapped and 2,303 remained unmapped.

The MLG detection analysis, using the IBS genetic-distance matrix, identified a total of 1,567 distinct genotypes, each represented by only one accession per MLG (clusters with one accession or C1), and hereafter referred to as “single-accession MLGs”. A total of 3,735 accessions were collapsed within 951 MLGs, ranging from 2 to up to 84 accessions per cluster (denoted in Table 4 as C2 to C84 in the MLG size column) and hereafter referred to as “multi-accession MLGs”. In total, the analysis using SNPs identified 2,518 unique MLGs across the 5,302 accessions (Table 4) and 2,784 accessions were detected as redundant. The MLG detection analysis using the Jaccard genetic-distance matrix identified a total of 1,568 single-accession MLGs. A total of 3,734 accessions were collapsed within 958 multi-accession MLGs containing from 2 to 87 accessions (C2 to C87). SilicoDArT marker analysis identified a total of 2,526 unique MLGs across the 5,302 accessions, and 2,776 accessions were detected as redundant. The two genetic distance measures showed that 47 percent of the genotypes were distinct or unique (47.5 percent and 47.6 percent for SNPs and SilicoDArT, respectively), while 52 percent were redundant (Table 4). Furthermore, we examined the dispersion of genetic distances among accessions collapsed within “multi-accession MLGs” of varying sizes, ranging from 2 to 84/87 accessions (as indicated in Table 4, MLG size C2 to C87). Interestingly, pairs of accessions within these MLGs exhibited dispersion of genetic distances above the given threshold for MLG detection (Figure 4B).

Additionally, we investigated the effect of varying the threshold for MLG identification using the SNP dataset, by running a series of MLG analyses and using IBS distance thresholds ranging from 0.000 to 0.06. When using thresholds between 0.01 and 0.06, the number of clusters varied moderately between 2,648 and 2,343. However, for thresholds below 0.01, the number of clusters increased substantially to more than 5,000 MLGs, most likely because of genotyping errors artificially inflating genetic distances (Figure 5A).

TABLE 3 Summary of multilocus genotype (MLG) detection using SNP and SilicoDArT markers in 21 accessions, including biological and technical replicates.

Accessions	SNP–IBS distance				SilicoDArT–Jaccard distance					
	Threshold 0.06*		Threshold 0.015**		Threshold 0.012**		Threshold 0.020**		Threshold 0.025**	
	MLG size	MLG ID	MLG size	MLG ID	MLG size	MLG ID	MLG size	MLG ID	MLG size	MLG ID
BRA534	7	1	7	1	7	1	7	1	7	1
BRA878	7	2	7	2	7	2	7	2	7	2
BRA891	7	3	7	3	7	3	7	3	7	3
COL1030	6	4	5	4	5	4	5	4	5	4
COL148	7	5	5	5	5	5	5	5	5	5
COL337	7	6	5	6	5	6	5	6	5	6
COL725	5	7	5	7	5	7	5	7	5	7
COL912B	7	8	7	8	7	8	7	8	7	8
COL979	7	9	7	9	7	9	7	9	7	9
CUB74/PAN70	14	10	13	10	13	10	13	10	13	10
ECU141A	7	11	7	11	7	11	7	11	7	11
FJI6	7	12	6	12	6	12	6	12	6	12
GUA32	7	13	7	13	7	13	7	13	7	13
MEX59	7	14	7	14	7	14	7	14	7	14
MEX86	7	15	7	15	1	15	1	15	7	15
PAR119	6	16	6	16	6	16	6	16	6	16
PER232	5	17	5	17	5	17	5	17	5	17
PER286	7	18	6	18	6	18	6	18	6	18
TAI8	7	19	7	19	7	19	7	19	7	19
USA4	1*	20	5	20	1	20	5	20	5	20
USA4	6*	21	–	–	1	21	–	–	–	–
USA4	–	–	–	–	3	22	–	–	–	–
MEX86	–	–	–	–	6	23	6	21	–	–
TOTAL	141	21	131	20	131	23	131	21	131	20

*dataset with 141 samples; **dataset with 131 samples.
Minimum genetic distance thresholds are specified along with MLG size and MLG number. Accessions with (+) indicate multiple MLG detection for at least one marker or distance threshold.

3.5 Comparative analysis of genetic redundancy assessment approaches

When comparing the various MLGs detected among the 5,302 accessions using both IBS and Jaccard distances, both approaches (SNP and SilicoDArT, respectively) consistently identified 5,047 accessions as either distinct or redundant. The remaining 255 accessions were not detected by either method. Among the 5,302 accessions, 1,440 were commonly detected within single-accession MLGs (datasets: “Silico C1: 1,568acc” and “SNP C1: 1,567acc”), while 3,607 accessions were identified within multi-accession MLGs (datasets: “Silico C2-C87: 3,734acc”, “SNP C2-C84: 3,735acc”) (Table 4; Figure 5B). Within the group of accessions detected as

redundant by both approaches, 119 accessions were detected in different MLG sizes by each method (SNPs and SilicoDArT). Additionally, 128 accessions were detected as single-accession MLGs with SilicoDArT markers, but those accessions were detected within multi-accession MLGs by SNPs. Conversely, 127 accessions were detected as single-accession MLGs by SNPs but were detected within multi-accession MLGs by SilicoDArTs (Figure 5B). A total of 374 accessions showed discrepancies in the detection of MLGs by both types of markers (119 + 127 + 128 = 374).
We classified MLGs into four distinct groups based on their sizes: single-accession MLGs (C1) coded in green, two-accession MLGs (C2) coded in yellow, and two additional categories for multi-accession MLGs collapsing over two accessions. The first

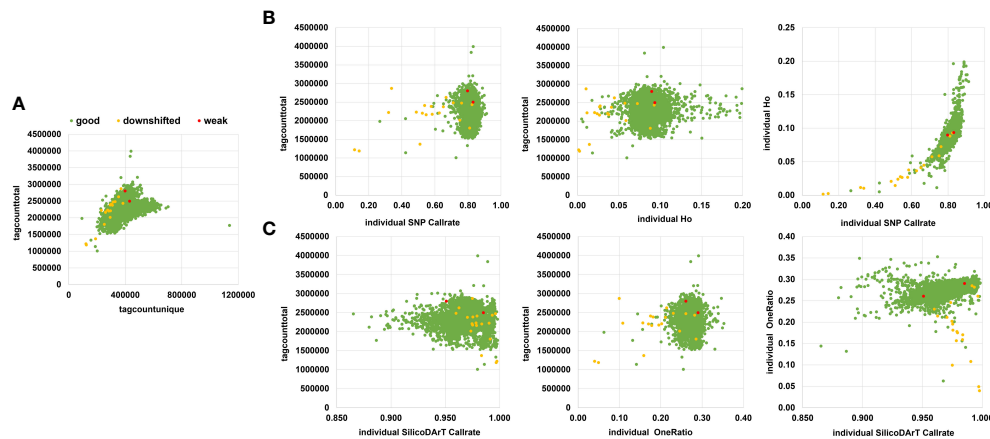


FIGURE 2

Sample quality parameters were assessed to filter out low-quality samples. (A) Scatter plot displaying tagtotalcounts vs. tagcountunique per sample. (B) Scatter plot showing the relationship between tagtotalcount vs. individual SNP call rate or individual observed heterozygosity (Ho), or individual call rate vs. Individual Ho. (C) Scatter plot using SilicoDarts, illustrating the relationship between tagtotalcount vs. individual call rate or individual oneRatio. Furthermore, it includes a comparison of call rate vs. oneRatio for SilicoDart markers. Samples are color-coded according to their target quality-control assessment on a gel, categorized as good (green), downshifted (yellow), or weak (red).

group comprised MLGs with 3–5 accessions (C3–C5) coded in orange, while the second group included MLGs with more than five accessions (C6 to C87) coded in red (refer to the last column in Table 4). Furthermore, we categorized accessions into five regions based on their countries of origin: western South America, eastern South America, Central/North America & Caribbean, Asia, and Africa (see Table 1). We then evaluated the number and percentage of accessions per region falling within the four distinct groups of MLG sizes (Table 5; Supplementary Figure 4). The most represented region within the collection is western South America (2,957 accessions) composed by Colombia, Venezuela, Ecuador, and Peru; followed by eastern South America (1,511 accessions) composed by Brazil, Paraguay, Bolivia and Argentina. Among the five regions, Africa and Asia showed the highest percentages of

distinctness, meaning that they are composed mostly of single-accession MLGs, with unique genotypes recorded for 68 and 39 percent of the accessions within each region, respectively. Central/North America & Caribbean, eastern South America and western South America were the regions showing the highest percentages of redundancy, with 83,73 and 68 percent, respectively (Table 5). On the other hand, western South America and eastern South America had the highest values of discrepancies across the results obtained from the two types of markers (SNP and SilicoDart).

The hierarchical clustering analysis of the 5,302 accessions, using IBS and Jaccard distances, revealed the presence of at least three major groups (Supplementary Figure 4). The first major group predominantly comprises accessions from western South America and Central/North America & the Caribbean, a second group

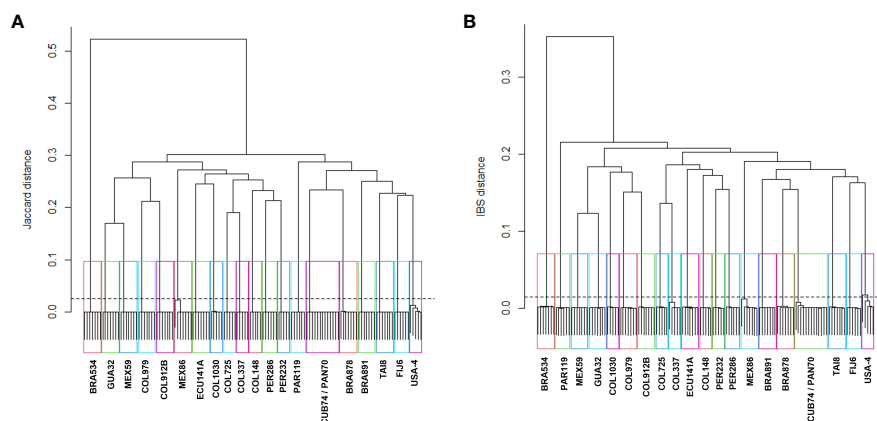


FIGURE 3

Dendrogram of hierarchical clustering using the "complete" method with the IBS distance for SNPs (A), and Jaccard distance for SilicoDarts (B). Twenty rectangles highlight the 20 clusters formed under a genetic distance of 0.015 or 0.025 using SNP and SilicoDart (dashed line) markers, respectively. Accession names representing samples within each group are shown. The distance matrices were calculated from 131 samples and 6,987 SNP or 13,715 SilicoDart markers.

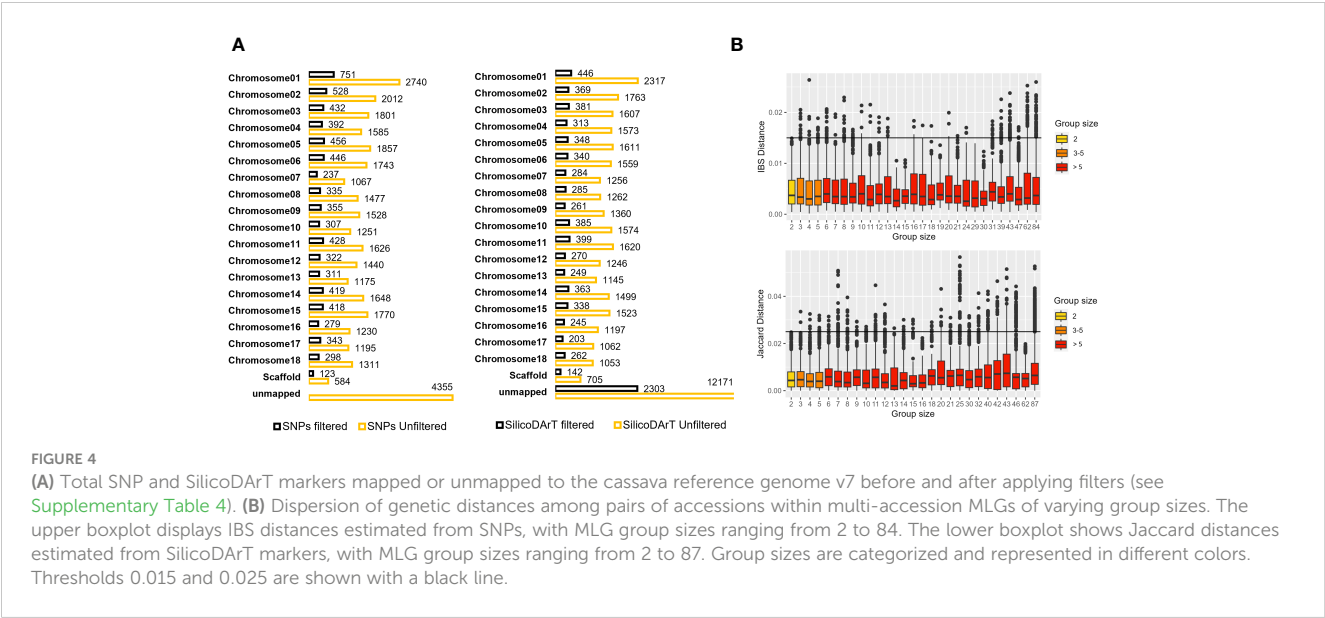


TABLE 4 Summary of MLG detection with SNP and SilicoDArT markers across 5,302 accessions using genetic distances thresholds of 0.015 and 0.025 for IBS and Jaccard distances, respectively.

MLG size	SNP		SilicoDArT		Color code
	Number of MLGs	Number of acc. in MLGs	Number of MLGs	Number of acc. in MLGs	
C1	1567	1567	1568	1568	Green
C2	472	944	495	990	Yellow
C3	191	573	174	522	Orange
C4	96	384	101	404	
C5	61	305	56	280	
C6	31	186	32	192	Red
C7	21	147	21	147	
C8	14	112	14	112	
C9	15	135	14	126	
C10	8	80	9	90	
C11	8	88	9	99	
C12	6	72	7	84	
C13	7	91	3	39	
C14	2	28	4	56	
C15	1	15	3	45	
C16	3	48	2	32	
C17	1	17	–	–	
C18	1	18	1	18	
C19	1	19	–	–	
C20	1	20	2	40	
C21	1	21	1	21	

(Continued)

TABLE 4 Continued

MLG size	SNP		SilicoDArT		Color code
	Number of MLGs	Number of acc. in MLGs	Number of MLGs	Number of acc. in MLGs	
C24	1	24	–	–	
C25	–	–	1	25	
C29	1	29	–	–	
C30	1	30	2	60	
C31	1	31	–	–	
C32	–	–	1	32	
C39	1	39	–	–	
C40	–	–	1	40	
C42	–	–	1	42	
C43	2	86	1	43	
C46	–	–	1	46	
C47	1	47	–	–	
C62	1	62	1	62	
C84	1	84	–	–	
C87	–	–	1	87	
Total	2518	5302	2526	5302	
Num. redundant	2784		2776		
% of distinct	47.49		47.64		
% of redundant	52.51		52.36		

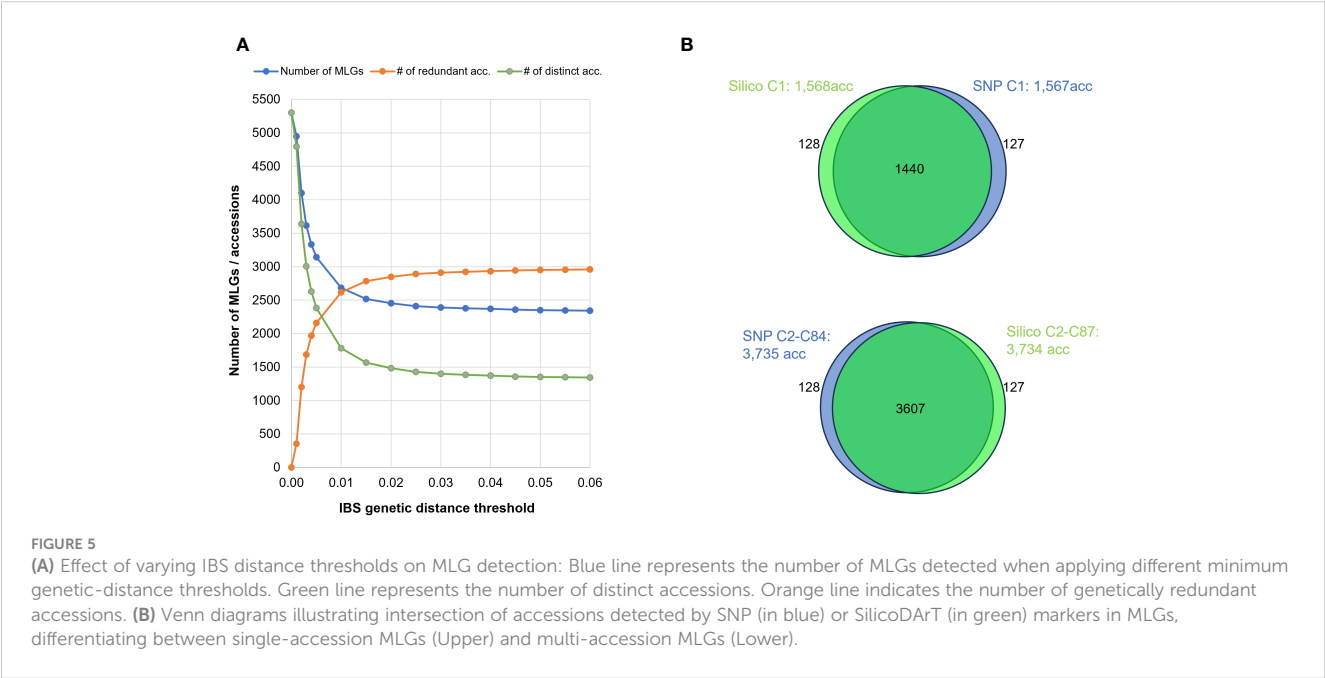


TABLE 5 Distribution of accessions categorized by region of origin and four distinct MLG size categories: single-accession MLGs (C1), two-accession MLGs (C2), MLGs with 3 to 5 accessions (C3–C5), and MLGs with more than 5 accessions (C6–C84/C87).

Type of marker	Regions		Western South America	Eastern South America	Central/North America & Caribbean	Asia	Africa	Total
	Parameter	Total	2967	1511	451	354	19	5302
SNP	Number of accessions	C1	943	399	74	138	13	1567
		C2	523	287	67	61	6	944
		C3–C5	606	413	108	81		1208
		C6–C84	841	412	202	74		1529
	Percentage	distinctness	31.8	26.4	16.4	39.0	68.4	
		redundancy	68.2	73.6	83.6	61.0	31.6	
SilicoDART	Number of accessions	C1	947	400	70	138	13	1568
		C2	546	301	71	66	6	990
		C3–C5	624	400	104	78		1206
		C6–C87	850	410	206	72		1538
	Percentage	distinctness	31.9	26.5	15.5	39.0	68.4	
		redundancy	68.1	73.5	84.5	61.0	31.6	
Both	Number of accessions	Discrepancy	214	114	30	16	0	374

encompasses a blend of accessions from eastern South America, western South America, Asia, and Africa. The third group predominantly comprises accessions from eastern South America (Table 1; Supplementary Figure 4). Both types of markers produced similar trees with slight differences. Although the purpose of this study is not to delve deeply into the population structure of the collection, the dendrogram helps visualize how single-accession MLGs and multi-accession MLGs are widely spread across the three major groups, with some differences across regions, as shown in Table 5. The discrepancies across both approaches (SNP and SilicoDART) are also widely spread across the three clusters (Supplementary Figure 4).

3.6 Examining passport and historical data in genetically redundant accessions

Twenty cases of multi-accession MLGs with varying sizes, ranging from 2 to 84, were selected to review the passport and historical characterization data. The biological status, country of origin, common names and collection dates were reviewed from the passport data. The analysis revealed that 10 cases shared the same biological status of Landrace or Breeding Line (refer to counts of ones across cases in Table 6), while another 10 cases have discrepancies. On the other hand, 10 cases included accessions from multiple countries ranging from 2 to 6, while the other 10 cases included accessions from only one country of origin (cases 1, 2, 3, 5, 7, 9, 10, 11, 12, and 17). Only one case among the 20 had accessions with identical common names within a single MLG (case 9). The remaining cases consist of MLGs with accessions known with multiple names ranging from 2 to 52 different names (Table 6).

Regarding the historical characterization data, five descriptors were reviewed, including shape of central leaf, petiole color, color of the first expanded leaf, number of leaf lobes and color of root pulp. Data were not available for all accessions. The percentage of accessions with available information for each descriptor varied from 20 percent for the number of leaf lobes, to 21 percent for color of the first expanded leaf, to 22 percent for petiole color, to 33 percent for shape of the central leaf, and to 77 percent for the color of root pulp (Table 6; Supplementary Table 5). We evaluated the variations within each MLG by assigning a value of 1 when all accessions shared the same descriptor and 2 or above when descriptors were different for at least one accession within an MLG. Additionally, we included the number of accessions with information for that particular descriptor, considering the total number of accessions within each MLG (Table 6).

When examining historical characterization descriptors individually, we observed that among all cases, only four MLGs (cases 5, 6, 7, and 8) had accessions with the same shape of the central leaf, either lanceolate or ovoid (Table 6; Figure 6). In all other instances, there were varied records for the central leaf shape within an MLG. Case 16, for example, encompasses 12 accessions, 10 of them documented with distinct central leaf shapes like oblong-lanceolate, linear-pandurate, lanceolate, and straight or linear (Supplementary Table 5, and Table 6). Note that cases where the MLG had only one accession with information for a particular descriptor were not counted as ones. In terms of petiole color, accessions within three MLGs (cases 1, 8, and 9) shared the same petiole color, while all other MLGs featured accessions with multiple petiole colors (green, green with some red, purple, red, red with some green or yellowish-green). Figure 6 showcases two accessions from case 16: AGR79 with linear-pandurate lobes and a red petiole, and

TABLE 6 Summary of passport and historical characterization data for selected multi-accession MLGs.

Cases		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Counts of ones
MLG	Size	2	2	2	2	2	2	2	2	2	3	3	4	6	7	7	12	28	29	31	84	
	Name (SNP/ SilicoDArT)	4/ 65	5/5	13/ 2977	25/ 2472	37/ 37	42/ 42	47/ 47	49/ 49	207/ 207	11/ 5157	35/ 4918	16/ 4775	2468 /3121	393/ 164	2525/ 2525	1168/ 1595	1182/ 4747	2313/ 2459	739/ 1273	4535/ 4313	
Passport Data	BiologicalStatus	1 (2/ 2)	1 (2/ 2)	1 (2/2)	2 (2/2)	2 (2/ 2)	2 (2/ 2)	1 (2/ 2)	2 (2/ 2)	1 (2/2)	1 (3/3)	1 (3/3)	1 (4/4)	2 (6/6)	2 (7/7)	1 (7/7)	1 (12/ 12)	2 (28/ 28)	2 (29/ 29)	2 (31/ 31)	2 (84/ 84)	10
	CountriesOrigin	1 (2/ 2)	1 (2/ 2)	1 (2/2)	2 (2/2)	1 (2/ 2)	2 (2/ 2)	1 (2/ 2)	2 (2/ 2)	1 (2/2)	1 (3/3)	1 (3/3)	1 (4/4)	5 (6/6)	2 (7/7)	3 (7/7)	3 (12/ 12)	1 (28/ 28)	6 (29/ 29)	5 (31/ 31)	5 (84/ 84)	10
	CommonNames	2 (2/ 2)	2 (2/ 2)	ND	2 (2/2)	ND	2 (2/ 2)	2 (2/ 2)	2 (2/ 2)	1 (2/2)	3 (3/3)	ND	2 (2/4)	5 (5/6)	6 (7/7)	5 (5/7)	10 (12/ 12)	11 (24/ 28)	18 (25/ 29)	24 (31/ 31)	52 (78/ 84)	1
	CollectionDates	ND	2	ND	ND	2	ND	ND	ND	2	1	ND	1	2	4		1	21	13	11	54	
Characterization Data	ShapeCentralLeaf	2 (2/ 2)	1 (2/ 2)	2 (2/2)	1 (1/2)	2 (2/ 2)	1 (2/ 2)	1 (2/ 2)	1 (2/ 2)	2 (2/2)	2 (3/3)	2 (3/3)	3 (4/4)	2 (6/6)	2 (6/7)	3 (7/7)	4 (10/ 12)	3 (26/ 28)	5 (27/ 29)	4 (18/ 31)	4 (77/ 84)	4
	PetioleColor	1 (2/ 2)	1 (1/ 2)	1 (1/2)	ND	ND	ND	1 (1/ 2)	1 (2/ 2)	1 (2/2)	ND	2 (3/3)	2 (4/4)	2 (3/6)	2 (3/7)	1 (1/7)	4 (8/12)	2 (3/28)	3 (7/29)	3 (7/ 31)	3 (4/84)	3
	Color1st ExpandedLeaf	1 (2/ 2)	1 (1/ 2)	1 (1/2)	ND	ND	ND	1 (1/ 2)	1 (2/ 2)	1 (2/2)	ND	1 (3/3)	2 (4/4)	1 (3/6)	1 (3/7)	1 (1/7)	2 (7/12)	1 (3/28)	1 (7/29)	1 (7/ 31)	2 (4/84)	10
	NumberLeafLobes	2 (2/ 2)	1 (1/ 2)	1 (1/2)	ND	ND	ND	1 (1/ 2)	2 (2/ 2)	1 (2/2)	ND	1 (3/3)	1 (4/4)	1 (3/6)	2 (3/7)	1 (1/7)	1 (7/12)	1 (3/28)	1 (5/29)	1 (5/ 31)	2 (4/84)	9
	ColorRootPulp	1 (2/ 2)	1 (2/ 2)	1 (2/2)	1 (1/2)	1 (2/ 2)	1 (2/ 2)	1 (2/ 2)	ND	1 (2/2)	1 (3/3)	2 (3/3)	2 (4/4)	1 (5/6)	3 (6/7)	1 (7/7)	1 (7/12)	1 (25/ 28)	1 (26/ 29)	1 (11/ 31)	3 (77/ 84)	14
	Counts of ones	5	7	6	2	2	2	7	3	7	4	4	4	3	1	2	4	4	3	3	0	
	Counts ND	1	0	2	4	4	3	1	2	0	3	2	0	0	0	0	0	0	0	0	0	

The name row indicates the ID of the MLG with the SNP and SilicoDArT analyses, respectively. The first number in the body of the table indicates the number of classes found within a given MLG and passport or characterization variable. The numbers in parentheses indicate the number of accessions with data available. ND indicates that data was not available. Cases and variables where only one category across the accessions included are summarized in the “counts of ones” row and column.

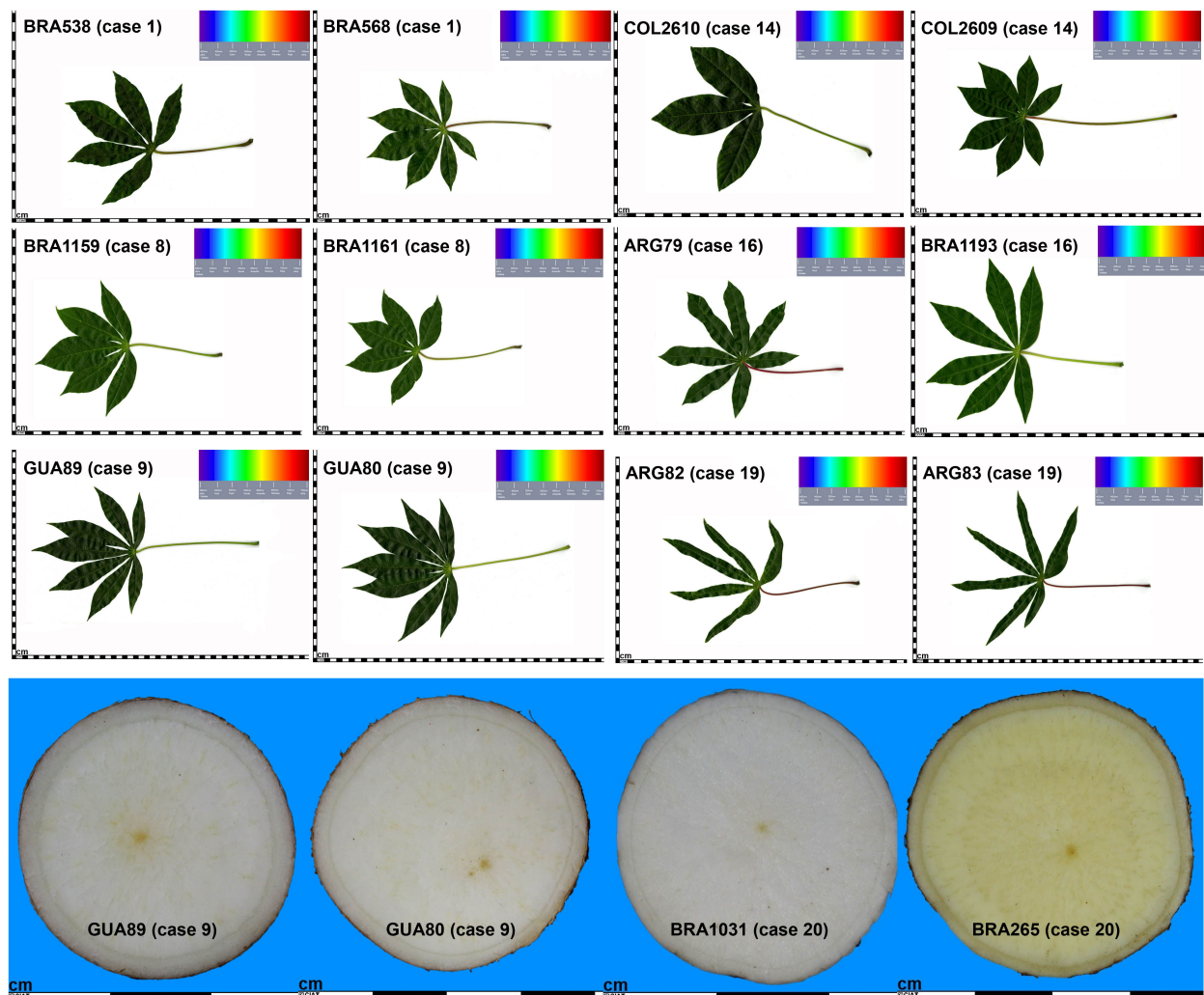


FIGURE 6

Selected cases of multi-accession MLGs, displaying historical leaf (top) and root (bottom) images from accessions within varied-sized MLGs, contrasting similarities, or differences in shape of central leaf, petiole color, number of leaf lobes, and color of root pulp.

BRA1193 with lanceolate central lobes and a yellowish-green petiole record. Regarding the color of the first expanded leaf and the number of leaf lobes, there were 10 and 9 cases, respectively, where the color and the number of lobules were consistent. For instance, cases 8 and 9 exhibited an equal number of leaf lobules. However, cases 1, 14, 16, and 19 contained accessions with varying numbers of lobes (Figure 6). Finally, among the cases examined, 13 showed consistency in the color of the root pulp, including instances like case 9 with accessions GUA80 and GUA89 (Figure 6). However, seven cases exhibited variations in the records for root pulp color. For example, case 20 consisted of 84 accessions, including BRA1031 and BRA265, which displayed differences in pulp color (Figure 6).

The examination of cases across the four passport variables and five historical characterization descriptors revealed that none of the 20 cases showed concordance across all nine reviewed descriptors. Nine cases had incomplete records, marked as “ND” (not determined), with the absence of information varying from 1 to 4 instances. Among these 20 cases, the highest number of coincidences observed was 7, found in only 3 MLGs (cases 2, 7, and 9). The

remaining cases of MLGs with genetically redundant accessions showed coincidences of 16 variables. Interestingly, the multi-accession MLGs with 84 accessions had zero coincidences across the nine variables under review (Table 6). The complete-linkage hierarchical agglomerative clustering of cases 1, 8, 9, 14, 19, and 20 is shown in Supplementary Figure 5A, displaying similarities in the clustering between SNPs and SilicoDARs. Similarly, as observed in Figures 4B, 3 or 5 out of the 20 cases of SNPs and SilicoDARs, respectively, displayed genetic distances across some of accession pairs higher than the given threshold (Supplementary Figure 5B).

4 Discussion

Cassava is currently the third-largest source of carbohydrates for human consumption in the world after rice and maize, playing a crucial role in ensuring food security, particularly in many low-income countries (Howeler et al., 2013). Global cassava production exhibits significant yields in Africa (64.8%), Asia (26.9%), and Latin America

(8.3%) (FAOSTAT, 2021), with a total production of 314 million tons in 2021. While productivity is influenced by factors such as agronomic practices, climate, and pest management, major challenges include susceptibility to pests and diseases, post-harvest losses, and the pressing need for improved varieties with higher yields and nutrition content (Montagnac et al., 2009; Zainuddin et al., 2018; Ntui et al., 2023).

To address these challenges, genebanks play a crucial role in conserving cassava diversity, serving as a foundation for breeding programs and facilitating the identification of sources of resistance to pests and diseases (Bellotti and Arias, 2001; Sheat et al., 2019). The operational procedures for seed storage and plant propagation have been in place for decades, enabling genebanks to establish, maintain, and conserve collections of plant genetic resources for crop improvement (Mascher et al., 2019). Nowadays, genebanks prioritize acquiring knowledge about their existing collections, focusing on enhancing the efficiency of genetic resources management and enhancing utilization, rather than expanding their collections (van Treuren and van Hintum, 2003). Research focused on comprehending and optimizing the composition of collections is therefore of particular interest (Nadeem et al., 2018; Mascher et al., 2019; Milner et al., 2019; Sansaloni et al., 2020). To optimize collection composition, it is essential for genebanks to prioritize the identification and elimination of redundancies, thereby considering both genetic and economic perspectives (van Treuren and van Hintum, 2003). Determining the appropriate thresholds to declare two genotypes as identical presents a challenging decision. Previous studies investigating genetic redundancy in cassava collections have defined their minimum genetic distance at 0.05, either arbitrarily (Albuquerque et al., 2019) or empirically (Orek et al., 2023; Soro et al., 2023). Empirical definitions have used the distribution of pairwise distances between duplicated DNAs as a 'calibration principle' (Noli et al., 2013; Rabbi et al., 2015). In our study, we took an empirical approach, optimizing the thresholds for identifying genetically redundant accessions at 0.015 and 0.025 genetic distance for SNP and SilicoDART markers, respectively (Table 3; Figure 3). To achieve this, we utilized a subset comprising 21 accessions from the core collection, along with technical and biological replicates. DNA-Reps were employed to reduce potential miscalling errors during genotyping, specifically for some heterozygous SNPs misidentified as homozygotes due to low sequencing read depth (Hamblin and Rabbi, 2014). Extract-Reps were used to address any traceability or contamination errors during DNA extraction. Additionally, Ind-Reps were considered to account for biological differences, potentially arising from traceability errors during routine multiplication of *in-vitro* plantlets.

The use of exemplary quality parameters for marker selection and sample exclusion had a notable impact on the estimated genetic distances (IBS and Jaccard) among replicates, resulting in reduced mean and standard deviation values (Table 2; Figures 1A, 2). Parameters such as call rate and maf were instrumental in selecting high-informative markers with minimal missing data (above 20 percent) and minor allele frequency greater than 0.001. Additionally, the choice of markers based on the average number of sequence tag copies (AvgMarkerCount) facilitated the removal of markers with insufficient sequencing depth. Furthermore, markers were selected with a low coefficient of variation of AvgMarkerCount

(CVMarkerCount) to exclude potential paralog sequences (McKinney et al., 2017). Crucially, sample quality played an important role in optimizing the minimum genetic distances for identifying redundant genotypes (Supplementary Figure 2; Table 3). Target QC is crucial for sequencing library suitability. While non-"good" categories may sometimes result in lower read counts, it is not always the case (Figure 2). Therefore, considering multiple sample quality parameters is vital for selecting high-quality samples. Ignoring this can overestimate genetic distances for redundant accessions (Supplementary Figure 2). This is evident in individual SNP call rate, showing a tendency for lower values, and in reviewing individual Ho and OneRatio of SilicoDART markers (Figure 2). Low individual Ho for SNP and OneRatio of SilicoDART marker values may result from low call rates, possibly due to low total read counts, while high values for these two parameters may indicate unintended DNA sample cross-contamination. However, caution is needed, as these parameters, if not used carefully, may introduce bias and eliminate samples with exceptional individual Ho and OneRatio values due to genetic background. Excluding samples with exceptionally low counts in total sequenced tags (tagcounttotal), unique reads (tagtotalunique), individual SNP call rate, individual Ho, and individual OneRatio (the latter for SilicoDARTs) proved effective in eliminating samples with probable technical issues, primarily derived from low-quality DNA or cross-contamination. Most of these problematic samples were identified as downshifted libraries. By implementing this set of parameters, it was possible to significantly reduce the minimum genetic distance thresholds, leading to the consolidation of all replicates from a single accession within individual MLGs.

The identification of MLGs with SNP markers has been employed in studies to assess genetic redundancy within cassava collections (Albuquerque et al., 2019; Soro et al., 2023). MLGs represent the combination of alleles at multiple genetic loci within an individual or group of individuals. This approach can be used to identify genetic redundancy (potential duplicate accessions) and to determine genetic distinctness (unique accessions) within a collection. In our study, we employed MLG detection to identify genetic distinctness and redundancy within CIAT's cassava collection, comprising 5,302 accessions. We utilized two marker types: co-dominant SNPs and dominant SilicoDART markers and compared results across the two approaches. The dominant markers, provide binary information, indicating the presence or absence of a specific allele at a particular locus, while the codominant markers provide more detailed information by distinguishing between different genotypes/allele combinations at a specific locus (Amiteye, 2021). The resulting genetic distance matrices of each genetic marker, IBS and Jaccard, exhibited high similarity in the number of MLGs detected, with 2,518 (47.4 percent) and 2,526 (47.6 percent) MLGs out of the 5,302 accessions, respectively (Table 4). Redundancy was observed across the five regions from which germplasm originates, but it was notably higher in accessions from Central/North America & the Caribbean (Table 5; Supplementary Figure 4). A total of 374 accessions exhibited discrepancies between both methods, with around 127 accessions being identified as unique by one method and redundant by the other. These accessions need to be further reviewed, especially in the cases where accessions have been

detected as unique by at least one approach. Interestingly, upon examining the dispersions of genetic distances for various MLG sizes, it becomes evident that for certain pairs of accessions within multi-accession MLGs of different sizes, the distance exceeds the minimum genetic distance thresholds used (Figure 4B).

In this study, we do not delve into the population structure of the collection. Instead, the focus of the hierarchical clustering analysis conducted on the 5,302 accessions is to compare and visualize the levels of genetic redundancy within and across regions of origin, as well as across marker types. A more comprehensive analysis, conducted in collaboration with the cassava collection conserved by the International Institute for Tropical Agriculture (IITA), is currently underway as a separate study. A recent study by Perez-Fon et al. (2023) identified two main gene pools, North & Northwest of the Amazon River basin (ARB) and South & Southeast of ARB, when assessing the genetic diversity of a set of 481 accessions selected as the most heterogeneous and unique cassava landraces. Without conducting additional analysis, our hierarchical clustering analysis using IBS (SNP) and Jaccard (SilicoDART) distances revealed the presence of at least three major groups (Supplementary Figure 4). The major group predominantly comprises accessions from western South America and Central/North America & Caribbean. A second group encompasses a blend of accessions from eastern South America, western South America, Asia, and Africa. Finally, a third group, predominantly comprising accessions from eastern South America, was identified (Table 1; Supplementary Figure 4).

While genomic information proves valuable for assessing genetic redundancy, interpreting molecular data remains complex due to the presence of diverse genetic relationships among potential duplicates (van Treuren and van Hintum, 2003). To complement the MLG analysis results, especially for identified redundant groups across both SNP and SilicoDART approaches, we reviewed and compared additional information from passport data and available historical characterization records. Passport data includes essential details such as genus name, country of origin, acquisition date, unique accession number, among others. Characterization data provides in-depth descriptions of plant germplasm, serving as a tool to confirm their authenticity and identify duplicates in a collection. Cassava experts have agreed upon a set of descriptors for characterization (Bioversity Int and CIAT, 2009) including shape of central leaf, petiole color, color of the first expanded leaf, number of leaf lobes, and color of root pulp.

We selected 20 multi-accession MLGs with varying number of accessions (ranging from 2 to 84) to review the passport and historical characterization data. Discrepancies and agreements were observed among accessions collapsed within multi-accession MLGs (Table 6; Figure 6). These differences can be attributed to various reasons. The five characterization descriptors compared were documented between 10 and 20 years ago, sourced either from the in-field collection maintained until the 1990s or from breeding programs as part of specific projects in the last decade (Mafla et al., 1993). Unfortunately, there is a gap in the information about the source of the five reviewed descriptors from characterization data, although it is known that records were collected before 2002 and root and leaf images were uploaded to the cassava database during 2002 and 2009/2013, respectively. These records may not directly align with the genotyped accessions, implying possible tracking errors during multiple multiplication cycles.

Furthermore, some multi-accession MLGs included accessions with genetic distances higher than the designated thresholds. This is evident from certain accession pairs that exceed the specified genetic threshold for MLG identification (Figure 4B; Supplementary Figure 5), indicating that MLGs with greater dispersion of genetic distances require further inspection. On the other hand, a major obstacle to validating the genetic redundancy of MLGs with the characterization data is its scarcity and incompleteness. Among the five characterization descriptors, missingness ranged from 23 to 80 percent. Hence, our preference is to categorize these groups of accessions, or multi-accession MLGs, as redundant rather than as duplicates, until resources become available to facilitate side-by-side comparison in the field. We consider three possible sources of error that can lead to the MLG clusters not corresponding to the passport and characterization data: (i) potential errors in the traceability of phenotypic data, (ii) potential cumulative errors as a result of decades of *in-vitro* propagation, before implementing the use of barcodes, and (iii) potential errors in the traceability of samples during the DNA extraction and genotyping process.

Considering the high cost of maintaining and distributing cassava collections — estimated at US\$71 USD per accession/year in 2011 (CGIAR Genebanks Consortium, 2011) that would now be the equivalent of US\$97 per accession — identifying duplicates within the collection is crucial for optimizing physical storage space, reducing maintenance costs, enhancing characterization, and ensuring collections' accessibility and usability. Duplicates within a collection may arise by mistake when a variety is introduced in the collection more than once for various reasons, such as different names given to the same varieties by farmers, and/or by inadequate documentation or record-keeping practices resulting in the same variety being catalogued multiple times, among other factors. Furthermore, the process of multiplying accessions over a span of more than 40 years may have caused mixing, leading to duplications under different accession names. This implies that the original diversity preserved in the genebank may have been compromised, to some extent, from the lack of methods ensuring traceability some decades ago.

The identification of distinct accessions is also relevant for managing the collections, for planning for new initiatives within genebanks, and for facilitating access and characterization. Cryopreservation of vegetatively propagated germplasm becomes a viable option for base collections as new techniques are developed (Jenderek and Reed, 2017). Current efforts are underway to establish a cryo-collection at CIAT's genebank. The 1,440 distinct accessions identified through SNP and SilicoDART markers are primary candidates to initiate this process. Similarly, some of the multi-accession MLGs that collapsed accessions falling below the used genetic thresholds are most likely duplicates. On the other hand, there is a need to further revise cases of multi-accession MLGs that group accessions with pairs of genetic distances higher than the set threshold. By conducting comprehensive genetic analyses, curators can make informed decisions regarding the conservation, utilization, and breeding of plant varieties. This information helps in identifying unique traits, understanding evolutionary relationships, and ensuring the preservation of valuable genetic material for future agricultural needs. Consequently, this study offers valuable insights for genebank curators, enabling them to (i) identify a genetically distinct subset of accessions

for targeted cryopreservation efforts, and (ii) discern genetic redundancy and potential accession duplicates. This information not only facilitates follow-up studies but also opens the door for potential removal of duplication from the collection, thereby reducing conservation costs and enhancing accessibility to cassava diversity.

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://dataverse.harvard.edu/privateurl.xhtml?token=cd06716b-4004-4918-9cb3-4030c39f97ce>, Alliance Bioversity Int. and CIAT Dataverse repository and https://gigwa.cgiar.org/FutureSeeds/?module=Cassava_study, Genotype Investigator for Genome-Wide Analyses (Gigwa).

Author contributions

MC-Y: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. JO: Methodology, Writing – review & editing. EA: Methodology, Writing – review & editing. MV-T: Methodology, Writing – review & editing. MC: Visualization, Writing – review & editing. NM: Writing – review & editing. PW: Conceptualization, Funding acquisition, Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for this research, authorship, and/or publication of this article. This work was supported by The Crop Trust, the Federal Ministry for Economic Cooperation and Development (BMZ), the Genebank Platform, and the CGIAR Genebank Initiative. This support enabled the conduct of the research, the provision of researcher positions, and covered publication fees.

References

- Albuquerque, H. Y. G., De Oliveira, E. J., De Brito, A. C., Andrade, L. R. B., De Carmo, C. D., Do Morgante, C. V., et al. (2019). Identification of duplicates in cassava germplasm banks based on single-nucleotide polymorphisms (SNPs). *Scientia Agricola* 76 (4), 328–336. doi: 10.1590/1678-992x-2017-0389
- Allem, A. C. (1994). The origin of *Manihot esculenta* Crantz (Euphorbiaceae). *Genet. Resour. Crop Evol.* 41 (3), 133–150. doi: 10.1007/BF00051630
- Amiteye, S. (2021). Basic concepts and methodologies of DNA marker systems in plant molecular breeding. *Heliyon* 7 (10), e08093. doi: 10.1016/j.heliyon.2021.e08093
- Baptista, P., Carillo, P., Pathirana, R., and Carimi, F. (2022). Management and utilization of plant genetic resources for a sustainable agriculture. *Plants* 11 (15), 2038. doi: 10.3390/PLANTS11152038
- Bellotti, A. C., and Arias, B. (2001). Host plant resistance to whiteflies with emphasis on cassava as a case study. *Crop Prot.* 20 (9), 813–823. doi: 10.1016/S0261-2194(01)00113-2
- Bioversity International & International Center of Tropical Agriculture (2009). Key access and utilization descriptors for cassava genetic resources. Available at: <https://cgspace.cgiar.org/handle/10568/73327>.
- Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34 (5), 562–570. doi: 10.1038/nbt.3535
- Carvalho, L. J. C. B., Anderson, J. V., Chen, S., Mba, C., Doğramaci, M., Carvalho, L. J. C. B., et al. (2017). Domestication syndrome in cassava (*Manihot esculenta* crantz): assessing morphological traits and differentially expressed genes associated with genetic diversity of storage root. *Cassava*. 12 (316). doi: 10.5772/INTECHOPEN.71348
- Chavarriaga-Aguirre, P., Maya, M. M., Tohme, J., Duque, M. C., Iglesias, C., Bonierbale, M. W., et al. (1999). Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. *Mol. Breed.* 5 (3), 263–273. doi: 10.1023/A:1009627231450/METRICS
- Cohen, J. I., Williams, J. T., Plucknett, D. L., and Shands, H. (1991). Ex situ conservation of plant genetic resources: Global development and environmental concerns. *Science* 253 (5022), 866–872. doi: 10.1126/SCIENCE.253.5022.866
- Dean, R. E., Dahlberg, J. A., Hopkins, M. S., Mitchell, S. E., and Kresovich, S. (1999). Genetic redundancy and diversity among “Orange” Accessions in the U.S. National

Acknowledgments

We extend our gratitude to colleagues from the Alliance of Bioversity International and CIAT, Gustavo Cardona, Ana Maria Leiva, and the dedicated staff of the cassava *in-vitro* conservation team for their outstanding collaboration in providing plant material and preparing samples, to Miguel Acosta for assisting the revision of historical characterization data sources, Mathieu Rouard for supporting the upload of data into Gigwa, and Olga Spellman and Glenn Hyman, Alliance of Bioversity International and CIAT Science Writing Service, for English and copy editing of this manuscript. Our special thanks go to Jorge Franco for his insightful conversations and valuable support in the field of biostatistics.

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.2023.1338377/full#supplementary-material>

- sorghum collection as assessed with simple sequence repeat (SSR) markers. *Crop Sci.* 39 (4), 1215–1221. doi: 10.2135/CROPSCI1999.0011183X003900040043X
- Dellaporta, S. L., Wood, J., and Hicks, J. B. (1983). A plant DNA miniprep: Version II. *Plant Mol. Biol. Rep.* 1 (4), 19–21. doi: 10.1007/BF02712670
- Duputié, A., Salick, J., and McKey, D. (2011). Evolutionary biogeography of Manihot (Euphorbiaceae), a rapidly radiating Neotropical genus restricted to dry environments. *J. Biogeogr.* 38 (6), 1033–1043. doi: 10.1111/j.1365-2699.2011.02474.x
- Engels, J. M. M. (2003). Plant genetic resources management and conservation strategies: problems and progress. *Acta Hort.* 623, 179–191. doi: 10.17660/ActaHortic.2003.623.19
- FAO (2010). *The Second Report on the State of the World's Plant* (Rome, Italy: FAO Commission on Genetic Resources for Food and Agriculture assessments) No. 2, 399 p.
- FAO (2014). *Genebank Standards for Plant Genetic Resources for Food and Agriculture* (Rome, Italy: Plant Production and Protection Division), 181 p.
- FAOSTAT (2021). Available at: <https://www.fao.org/faostat/en>.
- Ferguson, M. E., Shah, T., Kulakow, P., and Ceballos, H. (2019). A global overview of cassava genetic diversity. *PLoS One* 14 (11), 1–16. doi: 10.1371/JOURNAL.PONE.0224763
- Fu, Y. B. (2023). Assessing genetic distinctness and redundancy of plant germplasm conserved ex situ based on published genomic SNP data. *Plants* 12 (7), 1476. doi: 10.3390/PLANTS12071476Genesys
- Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). dart: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* 18 (3), 691–699. doi: 10.1111/1755-0998.12745
- Hamblin, M. T., and Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (Manihot esculenta). *Crop Sci.* 54 (6), 2603–2608. doi: 10.2135/cropsci2014.02.0160
- Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* 51 (9), 1349–1355. doi: 10.1038/S41588-019-0487-7
- Hershey, C. H. (2008). A global conservation strategy for cassava (Manihot esculenta) and wild Manihot species. Available at: <https://cgspace.cgiar.org/handle/10568/83105>.
- Hershey, C. H., Iglesias, C. A., Iwanaga, M., and Tohme, J. M. (1994). Definition of a core collection for cassava. *International Crop Network Series* 10, 145–156. Available at: <https://cgspace.cgiar.org/handle/10568/55840>.
- Howler, R., Litaladio, N., and Thomas, G. (2013). Save and grow: cassava. A guide to sustainable production intensification. *FAO publications catalogue* 10 (142).
- Jenderek, M. M., and Reed, B. M. (2017). Cryopreserved storage of clonal germplasm in the USDA National Plant Germplasm System. *In Vitro Cell. Dev. Biol. - Plant* 53 (4), 299–308. doi: 10.1007/s11627-017-9828-3
- Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2 (1), 1–14. doi: 10.7717/peerj.281
- Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol. Biol. (Clifton N.J.)* 888, 67–89. doi: 10.1007/978-1-61779-870-2_5
- Kisha, T. J., and Cramer, C. S. (2011). Determining redundancy of short-day onion accessions in a germplasm collection using microsatellite and targeted region amplified polymorphic markers. *J. Am. Soc. Hortic. Sci.* 136 (2), 129–134. doi: 10.21273/JASHS.136.2.129
- Lefèvre, F., and Charrier, A. (1992). Isozyme diversity within African Manihot germplasm. *Euphytica* 66 (1–2), 73–80. doi: 10.1007/BF00023510
- Leticia, L., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49 (W1), W293–W296. doi: 10.1093/nar/gkab301
- Mafla, G., Roca, W., Reyes, R., Roa E, J. C., Muñoz M, L., Baca, A., et al. (1993). *In vitro* management of cassava germplasm at CIAT. *Proceedings. Centro Internacional de Agricultura Tropical (CIAT), Cali, CO.* p. 168–174. (Working document no. 123). Available at: <https://cgspace.cgiar.org/handle/10568/55702>.
- Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* 51 (7), 1076–1081. doi: 10.1038/s41588-019-0443-6
- Mba, C., and Ogonnaya, F. C. (2022). Utilizing plant genetic resources to develop climate resilient crops. *Agric. Biotechnol. Biodivers. Biores. Conserv. Util.*, 1st Ed. CRC Press, 373–404. doi: 10.1201/9781003178880-22
- McKinney, G. J., Waples, R. K., Seeb, L. W., and Seeb, J. E. (2017). Paralogues are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol. Ecol. Resour.* 17 (4), 656–669. doi: 10.1111/1755-0998.12613
- Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C., and Georges, A. (2022). dartR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods Ecol. Evol.* 13 (10), 2150–2158. doi: 10.1111/2041-210X.13918
- Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., et al. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51 (2), 319–326. doi: 10.1038/s41588-018-0266-x
- Montagnac, J. A., Davis, C. R., and Tanumihardjo, S. A. (2009). Nutritional value of cassava for use as a staple food and recent advances for improvement. *Compr. Rev. Food Sci. Food Saf.* 8 (3), 181–194. doi: 10.1111/J.1541-4337.2009.00077.X
- Motilal, L. A., Zhang, D., Mischke, S., Meinhardt, L. W., and Umaharan, P. (2013). Microsatellite-aided detection of genetic redundancy improves management of the International Cocoa Genebank, Trinidad. *Tree Genet. Genomes* 9 (6), 1395–1411. doi: 10.1007/s11295-013-0645-5
- Murtagg, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif.* 31 (3), 274–295. doi: 10.1007/S00357-014-9161-Z
- Nadeem, M. A., Habyarimana, E., Çiftçi, V., Nawaz, M. A., Karaköy, T., Comertpay, G., et al. (2018). Characterization of genetic diversity in Turkish common bean gene pool using phenotypic and whole-genome DArTseq-generated silicoDArT marker information. *PLoS One* 13 (10), e0205363. doi: 10.1371/journal.pone.0205363
- Noli, E., Teriaca, M. S., and Conti, S. (2013). Criteria for the definition of similarity thresholds for identifying essentially derived varieties. *Plant Breed.* 132 (6), 525–531. doi: 10.1111/pbr.12109
- Ntui, V. O., Tripathi, J. N., Kariuki, S. M., and Tripathi, L. (2023). Cassava molecular genetics and genomics for enhanced resistance to diseases and pests. *Mol. Plant Pathol.* doi: 10.1111/MPP.13402
- Olsen, K. M. (2004). SNPs, SSRs and inferences on cassava's origin. *Plant Mol. Biol.* 56 (4), 517–526. doi: 10.1007/s11103-004-5043-9
- Olsen, K. M., and Schaal, B. A. (1999). Evidence on the origin of cassava: phylogeography of Manihot esculenta. *Proc. Natl. Acad. Sci. United States America* 96 (10), 5586–5591. doi: 10.1073/pnas.96.10.5586
- Orek, K., Kyallo, M., Oluwaseyi, S., and Yao, N. (2023). Genotyping by Sequencing Reveals Genetic Relatedness and Duplicates amongst Local Cassava (Manihot esculenta Crantz) Landraces and Improved Genotypes in Kenya. *Biotechnol. J. Int.* 27 (5), 29–46. doi: 10.9734/bji/2023/v27i5694
- Perez-Fon, L., Ovalle, T. M., Drapal, M., Ospina, M. A., Gkanogiannis, A., Bohorquez-Chau, A., et al. (2023). Integrated genetic and metabolic characterization of Latin American cassava (Manihot esculenta) germplasm. *Plant Physiol.* 192 (4), 2672–2686. doi: 10.1093/plphys/kiad269
- 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 (3), 559–575. doi: 10.1086/519795
- Qi, W., Lim, Y. W., Patrignani, A., Schläpfer, P., Bratus-Neuenschwander, A., Gräter, S., et al. (2022). The haplotype-resolved chromosome pairs of a heterozygous diploid african cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *GigaScience* 11. doi: 10.1093/GIGASCIENCE/GIAC028
- R Core Team (2022). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: <https://www.R-project.org/>.
- Rabbi, I. Y., Kulakow, P. A., Manu-Aduening, J. A., Dankyi, A. A., Asibuo, J. Y., Parkes, E. Y., et al. (2015). Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (Manihot esculenta Crantz). *BMC Genet.* 16 (1), 1–11. doi: 10.1186/s12863-015-0273-1
- Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 1. doi: 10.1038/s41467-020-18404-w
- Sheat, S., Fuerholzner, B., Stein, B., and Winter, S. (2019). Resistance against cassava brown streak viruses from africa in cassava germplasm from South America. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00567
- Soro, M., Pita, J. S., Somé, K., Otron, D. H., Yéou, E., Mutuku, J. M., et al. (2023). Genomic analysis and identification of potential duplicate accessions in Burkina Faso cassava germplasm based on single nucleotide polymorphism. *Front. Sustain. Food Syst.* 7, 1202015. doi: 10.3389/fsufs.2023.1202015
- Tanaka, N., Shenton, M., Kawahara, Y., Kumagai, M., Sakai, H., Kanamori, H., et al. (2021). Investigation of the genetic diversity of a rice core collection of Japanese landraces using whole-genome sequencing. *Plant Cell Physiol.* 61 (12), 2087–2096. doi: 10.1093/PCP/PCAA125
- van Treuren, R., and van Hintum, T. J. L. (2003). Marker-assisted reduction of redundancy in germplasm collections: genetic and economic aspects. *Acta Horticulturae* 623 (623). doi: 10.17660/ActaHortic.2003.623.15
- Virk, P. S., Newbury, H. J., Jackson, M. T., and Ford-Lloyd, B. V. (1995). The identification of duplicate accessions within a rice germplasm collection using RAPD analysis. *Theor. Appl. Genet.* 90 (7–8), 1049–1055. doi: 10.1007/BF00222920
- Westengen, O. T., Skarbo, K., Mulesa, T. H., and Berg, T. (2018). Access to genes: linkages between genebanks and farmers' seed systems. *Food Secur.* 10 (1), 9–25. doi: 10.1007/S12571-017-0751-6
- Zainuddin, I. M., Fathoni, A., Sudarmonowati, E., Beeching, J. R., Grissem, W., and Vanderschuren, H. (2018). Cassava post-harvest physiological deterioration: From triggers to symptoms. *Postharvest Biol. Technol.* 142, 115–123. doi: 10.1016/J.POSTHARVBIO.2017.09.00



OPEN ACCESS

EDITED BY

Axel Diederichsen,
Agriculture and Agri-Food Canada (AAFC),
Canada

REVIEWED BY

Photini V. Mylona,
Hellenic Agricultural Organisation (HAO),
Greece
Jens Weibull,
Swedish Board of Agriculture, Sweden

*CORRESPONDENCE

Nigel Maxted

✉ n.maxted@bham.ac.uk

RECEIVED 11 November 2023

ACCEPTED 24 January 2024

PUBLISHED 22 February 2024

CITATION

Almeida MJ, Barata AM, De Haan S, Joshi BK,
Brehm JM, Yazbek M and Maxted N (2024)
Towards a practical threat assessment
methodology for crop landraces.
Front. Plant Sci. 15:1336876.
doi: 10.3389/fpls.2024.1336876

COPYRIGHT

© 2024 Almeida, Barata, De Haan, Joshi,
Brehm, Yazbek and Maxted. 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.

Towards a practical threat assessment methodology for crop landraces

Maria João Almeida¹, Ana Maria Barata², Stef De Haan³,
Bal Krishna Joshi⁴, Joana Magos Brehm^{1,2}, Mariana Yazbek⁵
and Nigel Maxted^{1*}

¹School of Biosciences, University of Birmingham, Birmingham, United Kingdom, ²Banco Português de Germoplasma Vegetal, Instituto Nacional de Investigação Agrária e Veterinária, Braga, Portugal, ³Andean Food Systems, International Potato Center, Lima, Peru, ⁴National Gene Bank, National Agricultural Research Centre, Kathmandu, Nepal, ⁵Genebank, International Center for Agricultural Research in the Dry Areas, Terbol, Lebanon

Crop landraces (LR), the traditional varieties of crops that have been maintained for millennia by repeated cycles of planting, harvesting, and selection, are genetically diverse compared to more modern varieties and provide one of the key components for crop improvement due to the ease of trait transfer within the crop species. However, LR diversity is increasingly threatened with genetic erosion and extinction by replacement with improved cultivars, lack of incentives for farmers to maintain traditional agricultural systems, and rising threats from climate change. Their active conservation is necessary to maintain this critical resource. However, as there are hundreds of thousands of LR and millions of LR populations for crops globally, active conservation is complex and resource-intensive. To assist in implementation, it is useful to be able to prioritise LR for conservation action and an obvious means of prioritisation is based on relative threat assessment. There have been several attempts to propose LR threat assessment methods, but none thus far has been widely accepted or applied. The aim of this paper is to present a novel, practical, standardised, and objective methodology for LR threat assessment derived from the widely applied IUCN Red Listing for wild species, involving the collation of time series information for LR population range, LR population trend, market, and farmer characteristics and LR context information. The collated information is compared to a set of threat criteria and an appropriate threat category is assigned to the LR when a threshold level is reached. The proposed methodology can be applied at national, regional, or global levels and any crop group.

KEYWORDS

conservation, crop landraces, extinction, genetic erosion, methodology, plant genetic resources, threat assessment

1 Introduction

Globally, 135 million people in 2019 from 55 countries were reported to be facing phase 3 Crisis level food insecurity or worse, which is a 60% increase compared to 2015 when the figure was 80 million, while in total about 850 million people in the world were undernourished in 2021 (FAO et al., 2021). The human population is today 8.09 billion (22nd May 2023) and is predicted to rise to 9.7 billion by 2050, with 86% in developing countries (United Nations, 2021). It is predicted that global food production will need to grow by 60% globally, and 100% in developing countries compared to the 2005/2007 production levels to meet this growing demand (FAO, 2011). At the same time, crop production may decrease by 2% per decade if crop varieties are not adapted to the changing environment (IPCC, 2014). Although there are political reasons for food shortages, there are issues of food wastage and post-harvest losses to consider, plant breeders are increasingly requiring novel genetic diversity to increase production (Litraco and Violle, 2015). This diversity is often found in the traditionally grown, genetically diverse crop landraces (LR), which have not been bred for trait uniformity like modern cultivars.

Camacho-Villa et al. (2005) defined an LR as “a dynamic population(s) of a cultivated plant that has historical origin, distinct identity and lacks formal crop improvement, as well as often being genetically diverse, locally adapted and associated with traditional farming systems”. The importance of the utilisation of LR is well recognised, as they often contain unique trait diversity due to their adaption to the location where they developed, and trait introgression is relatively easy compared to crop wild relatives as there is no crossing barrier and they do not, through linkage drag, bring deleterious alleles that need to be excluded (Ellstrand, 2003). This adaptive trait diversity can sustain yield for LR in marginal environments and mitigate diseases or pest attacks, as well as drought, frost, and salinity tolerance, and even yield enhancement in improved varieties (Harlan, 1975; Frankel et al., 1995; Veteläinen et al., 2009). Importantly, LRs are often maintained by smallholders and indigenous farmers because of the multiple cultural, provisioning, and regulating ecosystem services they provide (de Haan, 2021). These may include diverse benefits such as cultural and local identities, superior organoleptic properties, and relative yield stability in marginal and/or variable environments, among other factors (Perales et al., 2005; Fliedel et al., 2013; Ortman et al., 2023).

Regardless of their obvious economic value, it is well established that LR are increasingly, globally threatened (Vavilov, 1957; Bennett, 1971; 1973; Harlan, 1972; 1975; Frankel, 1970; Frankel, 1972; Frankel, 1973; Hawkes, 1983) and it has been argued that they are the most severely threatened element of all biodiversity (Maxted, 2006). The justification for this proposition being: (i) there are very few inventories of extant LR in each country, each region, or globally (Maxted and Scholten, 2007; FAO, 2011; Jarvis et al., 2011; de Boef et al., 2013; Almeida et al., 2023); (ii) some government agencies and seed companies are actively promoting the replacement of genetically diverse LR by modern genetically uniform cultivars (Frankel and Hawkes, 1975; Harlan, 1975; Negri, 2005); (iii) in most countries no agency is direct responsibility for their conservation (Raggi et al., 2022); (iv) LR sales have been and

are impacted by seed legislation that requires all crop seed to be registered before it can be sold and to comply involves an additional cost to individual growers so inadvertently restricts seed sale and LR production (Maxted et al., 2013); (v) the internationalisation of food systems and pressure of evolving markets predicates varietal standards and uniformity (Negri, 2003; Joshi et al., 2004; Maxted et al., 2013); (vi) LR maintainers are often subsistence farmer growing LR for family or local consumption, but their prime motivation is commercial gain or food production not LR conservation for its own sake (Veteläinen et al., 2009); (vii) LR maintainers are almost always elderly and their number is dwindling each year (average age in the UK was 65 (Scholten et al., 2008); (viii) there is ineffective transmission of LR knowledge (cultivation and marketing) from maintainer generation to generation (Negri, 2003; Camacho-Villa et al., 2005); (ix) the traditional LR maintenance from generation to generation is breaking down with the children of maintainers failing to take over LR maintenance or farming altogether (Negri, 2003); (x) LR maintainers, like other rural populations globally are increasingly migrating from rural areas to cities and LR are often lost (Negri, 2005); and finally, (xi) there is the predicted detrimental impact of climate change on LR diversity (Jarvis et al., 2010). Each of these factors is threatening current LR diversity, both in terms of genetic (Hammer et al., 1996) and cultural/heritage (Negri, 2005) diversity loss and so inevitably likely to negatively impact future food security.

Effective conservation requires planning, which often includes conservation target prioritisation as conservation resources are always too limited to conserve all potential targets simultaneously (Kell et al., 2017). One commonly applied means of prioritisation is relative threat assessment, assessing the relative risk of extinction among competing conservation targets (Maxted et al., 2013). For wild species, the International Union for Conservation of Nature (IUCN) Categories and Criteria are universally recognised and used for threat assessment (IUCN, 2012). However, applying or adapting the IUCN Categories and Criteria for use in LR threat assessment is problematic because (a) it is not species or taxon threat assessment but genetic diversity within species or taxa that are being assessed for LR, (b) LR are crops that have been domesticated and therefore have intrinsically less genetic diversity than wild species, (c) LR populations are always managed by humans and local human management practices and global policies will impact LR maintenance and these factors must also be considered in LR assessment, and (d) government and industrial policies may encourage the promotion of high yielding cultigens and hybrid varieties that replace LR cultivation so actively eradicating LR diversity, such systematic eradication of wild species does not occur. Therefore, it is not feasible to use the standard IUCN Red List approach to LR threat assessment.

However, there is still the requirement for an effective means of LR threat assessment to focus conservation targeting and proposals have been made for LR threat assessment techniques (e.g., Joshi et al., 2004; Porfiri et al., 2009; Padulosi and Dulloo, 2012; de Haan et al., 2016), although no standardised LR threat assessment methodology is currently widely accepted or easily applied. Joshi et al. (2004) proposed categorising LR based on population,

ecological, and social criteria (adapted from Brush, 2000), along with use and modernisation criteria, and there are obvious parallels to the IUCN categories. Hammer and Khoshbakht (2005) developed a list of threatened crop species not LR by correlating the IUCN Red List of Threatened Plants (Walter and Gillett, 1998) results with the list in the 3rd edition of Mansfeld's Encyclopaedia of Agricultural and Horticultural Crops (Hanelt and IPK, 2001). To rationally apply regional funds for sustaining landrace cultivation, Porfiri et al. (2009) assessed LR threat level using five criteria: (i) presence of the product on the market, (ii) presence in the catalogues of seed companies/nurseries, (iii) number of cultivating farmers, (iv) areas under cultivation (as a percentage of the total regional area for the species), (v) new dedicated area trend (presence of new areas reserved to LR cultivation). Antofie et al. (2010) extended the work of Hammer (1991), and Hammer and Khoshbakht (2005) and suggested adapting the Red Listing approach for LR. The authors produced a data sheet for each LR including crop and LR vernacular and scientific names; seed origin; cultivation and location details; conservation status; photographs; authors and references. The data sheet presented information that would help identify LR Red Lists but did not actually assess individual LR threat. Voegel (2012) advocated using diverse crop information (e.g., historical material; statistical registers; lists/inventories of cultivars; scientific literature) to formulate a Red List system, based on the continuity of cultivation and use of a crop and cultivars over time in a certain location. Further in the same year, Padulosi and Dulloo (2012) proposed creating a Red List of cultivated plant species/varieties based on five steps: i) General assessment and inventory of LR; ii) Red List and vulnerable variety list establishment; iii) First validation of Red Lists; iv) Second validation of Red Lists; and v) Documentation and monitoring. More recently de Haan et al. (2016); de Haan et al. (2019), stress the importance of time series data in LR population monitoring, they suggest using (i) hotspot identification, (ii) total diversity, (iii) relative diversity, (iv) spatial diversity, and (v) collective knowledge as indicators of threat. Despite the individual merits of each of these approaches and their evolving refinement over time, most do not fully address the requirement to assess LR infra-specific level of threat, nor have they been widely applied by the global agrobiodiversity community. Also, the lack of information about LR (e.g., LR checklists or baseline assessment; LR statistical registers) in most countries would hinder their practical application. Indeed, having robust spatially implicit baseline data is a prerequisite for any threat assessment or LR monitoring. Then why rely on such data when the costs associated with genomic analysis are becoming less expensive? The reason why biodiversity and LR threat assessment is not done routinely using genomics is the sheer number of taxa or landraces that exist. FAO (2010) estimates there are about 7,000 crops cultivated routinely globally but there is no estimate we are aware of for the number of existent LR, but for rice alone, there are estimated to be approximately 120,000 LRs (Das et al., 2013), though this is probably a high number for a major crop. Even so, an estimate of a total number of over 400M LRs could exist, and routine threat assessment of this large of a cohort using molecular techniques is unrealistic.

As outlined, there have been several diverse attempts to propose a method to threat assess LR material, which in itself demonstrates the urgent requirement for such a method to aid LR conservation planning and maintenance. However, none has been widely applied in practice. Therefore, here we bring together some of the previous LR threat assessment authors and together propose a novel standardised and quantitative method that can be applied to objectively assess LR threat risk at any geographic level or crop.

2 Landrace threat assessment methodology

2.1 Pre-threat assessment

For LR threat assessment, the unit to be assessed is a LR, but here are preliminary issues that need to be resolved prior to making the actual assessment. These issues are often associated with gathering the necessary information that the assessment is based upon. Depending on the LR to be assessed, much information may already exist, and the process is primarily collation, but for other LRs it may involve generating additional information, commonly time series data related to LR population range, population trend, market and farmer characteristics, and cultivation context. It is also the case that assessment for either Red Listing or LR threat assessment is iterative, meaning the assessment is necessarily repeated because the assessment information for a LR changes over time – therefore there is a need to continue to gather assessment information and periodically repeat the assessment.

The process of gathering assessment information and periodically repeating the threat assessment would normally be discussed by a range of potential stakeholders from the LR maintainer/researcher community (= assessment team) with a particular interest in the LR to be assessed. The issues they might discuss and agree on are likely to include:

- a. *LR definition*: The assessment team will need to discuss and agree on what constitutes a LR. LRs are difficult to define precisely (Harlan, 1975; Brush, 2000; Negri, 2003; Camacho-Villa et al., 2005; and Negri et al., 2009). Zeven (1998) believed they were impossible to define, while agreeing they existed, and their conservation was a priority. However, a pragmatic working definition was proposed by Maxted et al. (2020) that a LR is a dynamic population of a cultivated plant species that has a: distinct diagnostic identity (defined in terms of pheno- and genotypic expression), historical origin, not been formally bred recently (with at least 10 generations post initial varietal release), and is also commonly intrinsically genetically diverse, locally adapted to its geographic location, associated with traditional cultivation systems, and with local cultural associations.
- b. *Nomenclatural/phenotypic/genomic distinction*: Practically, further clarification is required between genomic,

phenotypic, or nomenclatural distinction: is the LR to be assessed defined on its nomenclatural, phenotypic (morphological), or genomic identity? As an individual LR is not as easily identified as biologically distinct species using phenotypic distinction, genomic techniques would be required to decisively identify the populations that represent a specific LR. However, in practice, this would be excessively expensive to enact for the hundreds of thousands of LR and millions of LR populations that exist and might result in the identification of individual genotypes rather than genetically diverse recognised LRs. Therefore, practically LRs are almost always phenotypically (morphologically) and/or nomenclaturally defined. A group of LR populations share distinct, easily observed, and correlated morphological characteristics and/or are known by a single name. Most often a local community will recognise a distinct LR by its morphological characteristics and then use a local name to distinguish that LR. In which case, we assume the populations that have the same name and share morphological characteristics have a unique genetic identity, which is different from other LRs. It is noted that issues related to how landraces are practically recognised and studied are far from novel, some of the pioneers of genetic resources proposed elaborate scientific methods to use classical taxonomical approaches to describe and define basic units of genetic diversity. For example, the ‘ecogeographical classifications’ suggested by Vavilov (1926); Vavilov (1931) and elaborated by Sinskaya (1969) and Mansfeld (1951).

c. *Choice of assessment unit*: The choice of which LR to be assessed is often expedient; if conservation funding becomes available in a particular region, or an array of LR have breeder required trait (s), or a research project generates sufficient LR population descriptive and management data to facilitate threat assessment, then the LR is assessed and those most threatened can then be prioritised and actively conserved. When choosing which LR to threat assess, it could also be argued that care needs to be taken to avoid bias because (i) LR that are assessed as LC or NT will be preferentially assessed because by definition they are more abundant and more likely to be known to farmers/experts, as is evidenced by IUCN Red Listing (Hayward et al., 2015), (ii) LR that are assessed as VH or HI may also be preferentially assessed because they are known by farmers/experts as rare or threatened and assessors wish their preconception confirmed.

d. *Geographic scope (gene flow)*: It is preferable to assess each LR threat status throughout its range to supply the most comprehensive view of its threat status and avoid the need to replicate threat assessment at separate times by different authors in segments of its range. However, this is not always possible, the assessor may not have knowledge of the full geo-political range of the LR, or they may be professionally limited to working on national LR only so LRs found across national borders would be excluded, or a LR may be found on either side of a barrier to gene flow (e.g., mountains, sea) or germplasm exchange (e.g., different ethnic groups, nationality, or even gender). The critical issue is whether gene flow can or is thought to occur among LR populations – if there is gene flow the LR populations can be assessed as one LR but if there is no gene flow the LR populations should be assessed separately. As such, a LR may be assessed at a multi-national, national, national regional, or more restricted level, but in each case the most appropriate geographic scope for the assessment, or rather associated level of gene flow, needs to be agreed pragmatically by the assessment team based on the information available, particularly incorporating knowledge gained from discussion with those cultivating the LR.

2.2 Proposed landrace threat assessment methodology

The LR threat assessment method proposed is in part derived from the IUCN Red Listing method (IUCN, 2001) which is very widely used to assess biodiversity threats and has proven a globally invaluable tool for biodiversity conservation planning, but which is, as argued above, unsuitable for LR threat assessment. Like the IUCN Red List threat assessment so is the LR threat assessment method, but they should not be confused. The generalised principles of both involve five basic steps, but the approach taken is different in its application (Figure 1).

Step 1 – the assessment is focused on a single LR composed of one to many representative populations, a particular LR is selected on the basis of available assessment data and the wish to use the assessment in conservation planning.

Step 2 – involves the collation of LR representative population descriptive and management data.

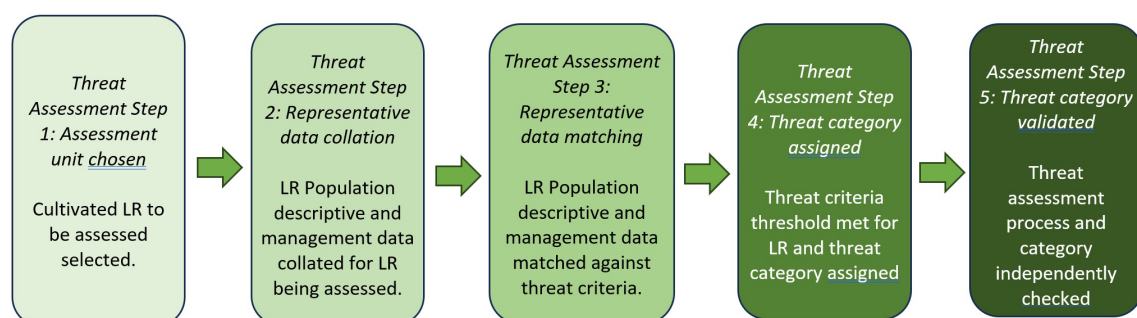


FIGURE 1
Generalised LR threat assessment procedure.

Step 3 – involves the matching of this LR representative population descriptive and management data against the LR threat criteria based on population and range sizes and changes over time, the market and farmer characteristics, and current conservation status. For example, when scoring subcriteria A1.1 LR Geographic Range, the extent of occurrence or area within which the LR population(s) are cultivated is 10km², then a score of 3 would be recorded. This process would be repeated for each subcriteria that data were available and therefore could be scored. The LR threat scores for all the subcriteria scored are summed and the threat percentage is calculated. For example, if scoring a LR 18 out of the 24 subcriteria can be scored, this gives a maximum potential score of 90 (18 subcriteria multiplied by 5, the maximum score for each). Then the actual score for the 18 subcriteria that could be scored is calculated as a percentage of the maximum score possible; in this example 75 out of 90, which is a threat assessment score of 83%.

Step 4 – the percentage threat score for the criteria that could be assessed is assessed against the threat category threshold and the categories to be assigned for the LR to be assessed is given. If in the example, the threat assessment score is 83% then the LR would be threat-assessed as *Very High (VH)* as the percentage Threat Assessment Score was over 80% for the criteria that could be scored and the LR is facing an extremely high risk of cultivation extinction.

Step 5 – involves validation, where the threat data, the justification for the threat assessment proposed and the LR threat category proposed summarised in the Assessment Report are checked by a Reviewer in a similar manner to the academic paper standard peer review process. If necessary the reviewer can request changes or approve the LR threat assessment.

To acknowledge the link between the Red Listing and LR threat assessment, but also to help avoid confusion between the two approaches, the LR threat categories used are distinct where they are not synonymous with those threat categories used in IUCN Red Listing. Such that the threatened categories for LR assessment are

Extinct (EX), Extinct On-farm (EO), Very High (VH), High (HI), Moderate (MO), Low (LO), Very Low (VL), Near Threatened (NR), Least Concern (LC) as well as Data Deficient (DD) and Not Evaluated (NE), as opposed to the IUCN Red List categories (IUCN, 2001, IUCN, 2012) Extinct (EX), Extinct in the Wild (EW), Critically Endangered (CE), Endangered (EN), Vulnerable (VU), Near Threatened (NR), Least Concern (LC), Data Deficient (DD), and Not Evaluated (NE). Both methods use the same terms for the categories Extinct (EX), Near Threatened (NR), Least Concern (LC), Data Deficient (DD), and Not Evaluated (NE), and therefore the definition is identical for both IUCN Red Listing and LR threat assessment as defined here. The definition of the LR unique threat categories is provided in section 2.4 below. See Figure 2 for a schematic representation of the LR threat assessment process.

2.3 Proposed LR threat criteria

LR threat assessment is based on a review of LR descriptive and management information for single or multiple LR populations representative of the LR being assessed. This information is based on available published and grey literature, personal observation of the LR, or focus group meetings with the local communities maintaining the LR. The assessment is based on matching the threat criteria against the characteristics of the LR populations; the criteria are partitioned to indicate a relative threat to LR sustainability and the greater the perceived risk the more likely genetic erosion or extinction.

The threat assessment criteria proposed are split into 4 main criteria, from A to D (A – LR Population Range; B – LR Population Trend; C – Market and Farmer Characteristics; D – LR Cultivation Context), and 24 subcriteria each partitioned to differentiate relative threat. Each subcriteria is divided into relative threat assessment ranges from most (score = 5) to least threatening (score = 1). For subcriteria that cannot be assessed, no score is recorded and they are not included in the threat summary calculation. For an assessment, the scores for each individual subcriteria (5 = most threatened to 1 = least threatened) that can be scored are summed and then converted to an assessment

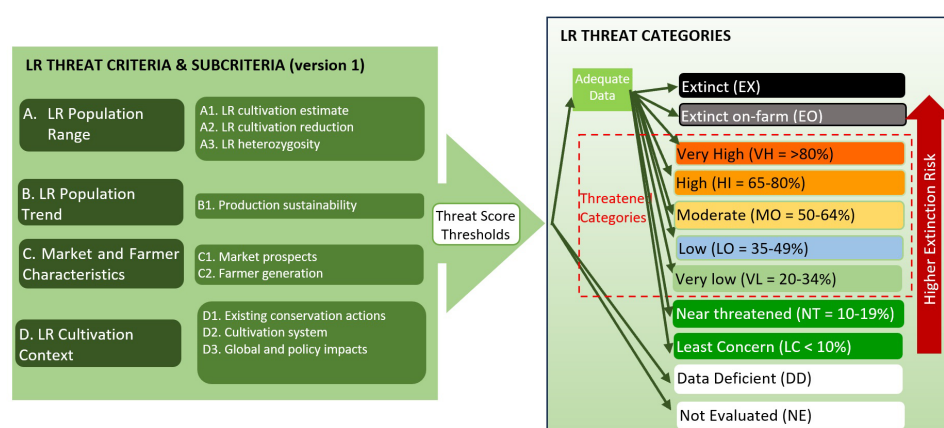


FIGURE 2
Summary of LR threat assessment criteria and categories.

percentage and this is matched to the LR threat categories, and a category assigned. It may not be possible to score all 24 subcriteria for every LR being assessed, but to ensure that the assessment maintains objectivity it is proposed at least two-thirds of subcriteria are scorable (that is, 16 out of 24 are scorable), if less than two-thirds of subcriteria can be scored then the LR is assessed as Data Deficient.

For several criteria threat is being assessed over time, but what is a scientifically justified time interval to provide meaningful threat assessment in the case of LR? IUCN (2012) in a similar situation uses the last 10 years or three generations, “because measuring changes over shorter time periods is difficult and does not reflect timescales for human interventions” (Mace et al., 2008). However, crops and their LR populations are genetically dynamic, their genetic diversity will change each year, perhaps even more so than wild taxa because they are subject to natural evolutionary pressures, as well as maintainer selection each generation. Even though dynamic change will occur, it will be within limits or the LR itself would lose its distinguishing features. It is also possible that in marginal environments the relative abundance of a particular LR can change substantially year on year, so a 10-year periodicity seems a justifiable timeframe for annual crops. However, it is recognised that this periodicity may need to be amended following further practical application of the LR threat assessment methodology proposed. A further consideration may be the increasing use of citizen science which is likely to facilitate more intense and frequent measurements, if deemed beneficial.

Although practically it is the number of generations that is important not the actual number of years as LR genetic diversity loss can only occur when there is generational change, not within a particular single generation’s lifetime. Further, Jain (1961) investigating the loss of genetic diversity during regeneration found that after 19 generations of bulk composite crossing in annual self-pollinating cereals 50-70% of variation for height and heading was lost but that after 10 generations significant loss of genetic diversity could be detected. Therefore, here the time interval for assessment proposed is over 10 generations, 10 years for an annual crop but longer for a perennial. Ten years may also be thought of as a LR maintainer’s detailed knowledge retention time, about the time a LR maintainer can accurately remember details of the LR they maintain or have knowledge of. It is also recognised that the number of generations may need to be changed when dealing with non-seed-based crops such as those clonally propagated.

The 24 subcriteria are described below and summarised in Table 1:

A: LR Population Range

A1: LR cultivation estimate

A1.1: Geographic range – LR population health is estimated as the geographic spread of a LR estimated using its cultivated extent of occurrence (EOO) (see IUCN, 2012): the smaller the geographic range the greater the LR extinction risk.

A1.2: Geographic concentration – estimated as the geographic concentration of the LR using its cultivated area of occurrence (AOO) (see IUCN, 2012): a relatively smaller area of cultivation indicates the relative risk of extinction. A1.1/A1.2 can be assessed using GeoCAT (Bachman et al., 2011) or participatory mapping (Plasencia et al., 2018).

A1.3: LR maintainer number – estimated as the number of LR maintainers today repeatedly planting, cultivating, and seed saving: the lower the number of maintainers each going through the cultivation cycle the greater the LR extinction risk.

A2: LR cultivation reduction

A2.1: Geographic range reduction – estimated as the change in the geographic spread of a LR estimated using its cultivated extent of occurrence (EOO), where the relative decrease in cultivated EOO indicates the relative risk of extinction: the larger the decrease the greater the risk. This is assessed as the average range change over 10 generations. Ten generations is sufficiently long to avoid annual sowing variation being recorded, while permitting distinction of significant long-term changes.

A2.2: Geographic concentration reduction – estimated as the change in the geographic concentration of a LR estimated using its cultivated area of occurrence (AOO), where the relative decrease in cultivated AOO indicates the relative risk of extinction: the larger the decrease the greater the risk. This assessed over 10 generations, so 10 years for an annual crop but longer for a perennial.

A2.3: Geographic constancy – LR population health is estimated by consistency in cultivation levels (roughly similar areas planted or numbers of plants sown and harvested), in terms of range and concentration assessed over 10 generations, whether maintainers cultivate roughly the same geographic range and concentration for LR generation to generation over the latest 10 generation period. Greater instability of cultivation indicates the rise and fall of LR population levels over time which increases the relative risk of extinction. LR population rise, and fall, is estimated by percentage of population change magnitude (increase or decrease) from generation to generation. Therefore, this is assessed as the average generational change in the LR range and average generational change in LR concentration over 10 generations divided by two.

A2.4: Maintainer number reduction – estimated by the relative number of maintainers cultivating LR over 10 generations: reduction in the number of maintainers between the number in year one compared to year ten would be an indication of increased relative risk of extinction.

A3: LR heterozygosity

A3.1: LR phenotypic diversity – estimated as the amount of phenotypic diversity observed in the LR populations: the greater the diversity the less likely the LR is to be threatened by natural or anthropogenic changes. Phenotypic diversity should be assessed using the standard phenotypic descriptor lists, Bioversity International lists numerous crop-based descriptor lists (<https://alliancebioversityciat.org/publications-data>), as well as the generalised FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 (Alercia et al., 2015). Here, phenotypic diversity is calculated as the percentage of phenotypic descriptors with at least two or more descriptor states recorded for the LR. Ideally, it is recommended to undertake on-farm characterisation trials with all LR from the region over two cropping seasons, with a minimum of one cropping season. As a minimum the assessment team could interview the maintainers and receive guidance on the relative number of descriptors showing phenotypic variation.

A3.2: LR exchange – estimated as the percentage of maintainers that exchange LR material after harvest with other locally-based

TABLE 1 Version 1 of criteria, subcriteria groups and subcriteria, and indicators of relative threat.

Criteria	Subcriteria	Threat assessment scores					Data sources
		5	4	3	2	1	
A. LR Population Range	A1: LR cultivation estimate						
	A1.1: Geographic range	<1 km ²	1-5 km ²	6-20 km ²	21-40 km ²	≥40 km ²	Obs.
	A1.2: Geographic concentration	<0.5 km ²	0.5-1 km ²	2-3 km ²	4-10km ²	≥10 km ²	Obs.
	A1.3: LR maintainer number	1	2-5	6-15	16-25	≥26	Obs.
	A2: LR cultivation reduction						
	A2.1: Geographic range reduction	≥90%	70-89%	50-69%	30-49%	<30%	Obs.
	A2.2: Geographic concentration reduction	≥90%	70-89%	50-69%	30-49%	<30%	Obs.
	A2.3: Geographic constancy	≥90%	70-89%	50-69%	30-49%	<30%	Obs.
	A2.4: Maintainer number reduction	≥90%	70-89%	50-69%	30-49%	<30%	Obs.
	A3: LR heterozygosity						
	A3.1: LR phenotypic diversity	<30%	30-49%	50-69%	70-89%	≥90%	Farmer Sur.
	A3.2: LR exchange	<30%	30-49%	50-69%	70-89%	≥90%	Farmer Sur.
B. LR Population Trend	B1: Production sustainability						
	B1.1: Ease of multiplication	<20%	21-40%	41-60%	61-80%	>80%	Farmer Sur.
	B1.2: Maintainer continuation	<30%	30-49%	50-69%	70-89%	≥90%	Farmer Sur.
	B1.3: LR known loss	>4	3	2	1	0	Farmer Sur.
	B1.4: Cultivation of modern cultivars	90%	70%	50%	30%	10%	Farmer Sur.
C. Market Farmer Characteristics	C1: Market prospects						
	C1.1: LR support applied	No support	–	LR generic	–	LR specific	Farmer Sur.
	C1.2: Market range	Local	–	Regional	–	National	Farmer Sur.
	C.1.3 Food system embeddedness	Weak (few households)	–	Intermediate (mid nos. households)	–	Strong (most households)	Farmer Sur.
	C2: Farmer generation						
	C2.1: Maintainer age	≥70	56-69	41-55	26-40	≤25	Farmer Sur.
D. LR Context	D1: Existing conservation actions						
	D1.1: Conserved in situ	No routine maintenance	1-9 pops. on-farm	≥10 pops. on-farm	1-9 pop. conserved	≥10 pops. conserved	Obs.
	D1.2: Conserved in situ backup	< 5% pops. duplication	5-30% duplicated	31-70% pops. duplicated	71-95% pops duplicated	>95% pops. duplication	Obs.
	D1.3: Conserved ex-situ	No conservation	1-9 pops. conserved	≥10 pops. conserved	1-9 pops. conserved in last 10 yrs.	≥10 pops. conserved in last 10 yrs.	Obs.
	D2: Cultivation system						
	D2.1: Type of cultivation system	<30%	30-49%	50-69%	70-89%	≥90%	Farmer Sur.

(Continued)

TABLE 1 Continued

Criteria	Subcriteria	Threat assessment scores					Data sources
		5	4	3	2	1	
	D2.2: Herbicide and fertilizer usage	≥90%	70-89%	50-69%	30-49%	≤10%	Farmer Sur.
	D3: Global and policy impacts						
	D3.1: Distorting incentives	Direct distorting incentives	–	Indirect distorting incentives	–	No distorting incentives	Farmer Sur.
	D3.2: Global stochastic impact	≥90%	70-89%	50-69%	30-49%	≤10%	Farmer Sur.

In terms of data sources Obs., Assessment team observation and Farmer Int., Farmer survey.

maintainers. LR material exchange promotes continued heterozygotic diversity and resilience to natural or anthropogenic changes, so reducing extinction risk.

B: LR Population Trend

B1: Production sustainability

B1.1: Ease of multiplication – estimated as the percentage of farmers that report that LR seed/material is abundant and/or potentially easily propagated: relative ease of potential multiplication is an indication of reduced extinction risk.

B1.2: Maintainer continuation – estimated as the percentage of LR maintainers that report that within their families or the local community, there is interest in maintaining the LR post current maintainer retirement: the stronger the indication that the next generation of maintainers will continue LR maintenance the smaller the extinction risk.

B1.3: LR known loss – estimated as the number of all LR from the same local area known to be no longer cultivated by local maintainers over the last 10 years: the greater the number of LR lost the greater the likelihood that further LR will cease to be cultivated. As above, 10 years may be used as this may be thought of as the LR maintainer's detailed knowledge retention time, about the time a LR maintainer can accurately remember details of the LR maintained.

B1.4: Cultivation of modern cultivars – estimated as the proportion of arable land of the same crop being covered with modern cultivars as the LR being assessed: the greater the proportion of cultivars grown the more likelihood that further LR will cease to be cultivated as maintainers potentially switch to cultivar production.

C. Market & Farmer characteristics

C1: Market prospects

C1.1: LR support applied – identified as any external support (financial or other), primarily from governmental sources, provided to the maintainer or seller that encourages cultivation or marketing of the specific LR being assessed: the presence of LR maintenance incentives indicates reduced threat. Such incentives may be specific, such as particular support for individual LR as recognition under Commission Directive 2008/62 EC, as 'conservation varieties' or designation using a quality label, or a regional uniqueness scheme, like the European PDO (Protected Designation of Origin) or PGI (Protected Geographical Indication), which covers agricultural

products and foodstuffs. Incentives may also be generic, support for any LR such as Payment for Environmental Services (PES) under the UK Agricultural Bill (UK Parliament, 2020) or the voluntary benefit sharing scheme applied for potato LR in Peru called AGUAPAN, where the private sector directly make payments to LR diversity guardians (see: www.aguapan.org). There are also countries where no specific or generic support for LR maintenance or marketing is provided and here LR are more likely to be threatened with cultivation cessation and extinction.

C1.2: Market range – estimated as the breadth of sales and marketing of LR or LR-derived products in the national, sub-national regional, or local markets: the broader the geographic range of the market for the LR or LR-derived products the less likely the maintainer will cease cultivation. It should be noted that in purely subsistence-based farming systems, there will be relatively low engagement with markets so maintainers will not receive market-based security and are more susceptible to stopping growing LR.

C1.3 Food system embeddedness - estimated as the likelihood of LR use in the regional food system or cuisine: the more LR are embedded in the local cuisine the less likely they are to be threatened. Many LR in purely subsistence-based farming systems may not engage with markets but are conserved at the household level because of their superior quality or organoleptic traits.

C2: Farmer generation

C2.1: Maintainer age – estimated as the average age of the maintainers that are cultivating and marketing/consuming the LR: the older the maintainer cultivating the LR the more threatened the LR will be as all maintainers must eventually retire.

D. LR Context

D1: Existing conservation actions

D1.1: Conserved in situ – identified by the relative *in situ* on-farm conservation effort: with the most conservation secure LR having more populations actively conserved *in situ* on-farm and the most threatened being those populations of the LR where there is no active on-farm maintenance. Brown and Briggs (1991) suggested that five populations would effectively capture 90-95% common alleles, but this is a minimum number so using 10 populations would aid security of maintenance. Also, here we distinguish between active and passive on-farm conservation, where active

on-farm conservation means the maintainer is provided with some form of support to retain existing LR diversity, while passive conservation is where the LR maintainer themselves alone wishes to maintain the LR. Therefore, relatively active on-farm conservation is more secure than passive on-farm maintenance, with a representation of genetic diversity in multiple populations being preferable to a few or single on-farm population, and no regular on-farm maintenance most threatened.

D1.2: Conserved *ex-situ* backup – identified by the proportion of *in situ* populations of the LR sampled and backed up in an *ex-situ* collection: the greater the backup the less likely the LR is to be threatened. It is widely recognised that to be effectively conserved, *in situ* or on-farm populations should be backed-up *ex-situ*. This has two advantages, it means that if the *in situ* on-farm populations are lost they might be reintroduced and restored from the *ex-situ* backup, and the *ex-situ* backup sample might be used to meet any user requirement. As such, it is likely that *ex-situ* backups provide improved chances of survival, as backed-up and used populations are perceived as having higher value and so less threatened.

D1.3: Conserved *ex-situ* – identified by the number and timing of *ex-situ* sampling: with the most conservation secure having higher numbers of LR population and more recent samples conserved as *ex-situ* accessions. To ensure that the genetic diversity in the on-farm populations is relatively well represented in the samples held *ex-situ*, the samples recognised should have been collected and entered the *ex-situ* facility within the past 10 years.

D2: Cultivation system

D2.1: Type of cultivation system – estimated as the percentage of maintainers with sustainable or traditional farming systems, rather than more commercial or industrial farming systems in the area where the LR is maintained: the greater the number of LR populations maintained within more sustainable or traditional farming systems, the less likely the LR is to be threatened.

D2.2: Chemical herbicide and fertiliser usage – estimated as the percentage of maintainers that routinely use chemical herbicides, fungicides, and fertiliser to stimulate production and yield: the greater the proportion of maintainers with LR populations maintained by using more sustainable or traditional farming systems the less likely the LR is to be threatened.

D3: Global impacts

D3.1: Distorting incentives – Distorting or perverse incentives are benefits provided to LR maintainers by those wishing LR growers to switch to potentially more productive crop varieties. These incentives may be supplied by governments or companies that have a vested interest in promoting cultigen or hybrid production. Distorting incentives may be direct or indirect, meaning they are focused either directly on LR or on the farming system and have an indirect impact on the LR. The more direct the distorting incentives the more likely LR maintainers will switch production and the LR will be eroded or lost.

D3.2: Global stochastic impact – estimated as the percentage of maintainers reporting their LR maintenance is being impacted by global deleterious factors such as environmental change, floods, heat, droughts, and wildfires, although these events may be beyond the control of the local community, they can seriously threaten LR maintenance (Jarvis et al., 2010).

2.4 Proposed LR threat categories

The LR Threat Categories¹ used to describe relative LR threat are as follows:

Extinct (EX) – A LR is extinct when there is no reasonable doubt that the last population of the LR has been lost on-farm and there are no samples held using *ex-situ* techniques. A taxon is presumed Extinct when exhaustive surveys in known and/or expected regions of cultivation throughout its historic range and *ex-situ* collection surveys have failed to record any cultivated or conserved populations of the LR.

Extinct on-farm (EO) – A LR is Extinct On-farm when it is known only to survive in active *ex-situ* conservation, primarily as a seed sample in a genebank, but also possibly as a living plant in a field genebank or seed or tissue culture held in *in vitro* culture or frozen at -196°C in cryopreservation; when exhaustive surveys of previously known areas of cultivation have found no known cultivation either on-farm or in a home garden throughout its historic range it is Extinct On-farm.

Very High (VH) – A LR has a Very High risk of extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score over 80% for the criteria that can be scored, and it is therefore considered to be facing an extremely high risk of cultivation extinction.

High (HI) – A LR has a High risk of extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of 65-80% for the criteria that can be scored, and it is therefore considered to be facing a high risk of extinction from cultivation.

Moderate (MO) – A LR has a MOderate risk of extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of 50-64% for the criteria that can be scored, and it is therefore considered to be facing a moderate risk of extinction from cultivation.

Low (LO) – A LR has a LOw risk of extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of 35-49% for the criteria that can be scored, and it is therefore considered to be facing a low risk of extinction from cultivation.

Very low (VL) – A LR has a Very Low risk of extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of 20-34% for the criteria that can be scored, and it is therefore

¹ It is acknowledged that the description of the LR threat categories is derived from the IUCN Red List categories (IUCN, 2001).

considered to be facing a very low risk of extinction from cultivation.

Near threatened (NT) – A LR is Near Threatened by extinction when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of 10–19% for the criteria that can be scored, and it is therefore considered to be facing an extremely low risk of extinction from cultivation but is sufficiently close to qualifying for or is likely to qualify for a threatened category in the near future, so the LR should be monitored and reassessed regularly.

Least Concern (LC) – A LR is Least Concern when the best available evidence indicates, following LR criterion scoring, that it has a percentage Threat Assessment Score of <10% for the criteria that can be scored, and it is therefore considered to be facing negligible risk of extinction from cultivation. Its cultivation is widespread and locally abundant.

Data Deficient (DD) – A LR is Data Deficient when there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on the available distribution and/or management data. To effectively estimate threat at least two-thirds of subcriteria must be scorable or ≥16 out of 24 are scorable, if less it is assessed as Data Deficient. Listing an LR in this category indicates that more information is required to make an assessment.

Not Evaluated (NE) – A LR is Not Evaluated when it has not yet been evaluated against the criteria.

2.5 Proposed threat subcriteria data collation

A key component of the LR assessment is collating the data for the assessment subcriteria and, in practice, using a standard questionnaire when interviewing LR maintainers was helpful. The questionnaire was developed from those used by Kell et al. (2009), Fonseca (2004), and the *Banco Português de Germoplasma Vegetal* (BPGV). The data recorded related to the LR maintainer (e.g., farmer's age, gender); socio-economic conditions; cultivated crops; cultural practices; qualities of LR; and seed characteristics were collected using the questionnaire (see Table 2). However, there is also a range of other tools, including quantitative instruments, that can aid the assessment of subcriteria.

3 Discussion

LR diversity is increasingly recognised as a critical resource for contemporary crop improvement (Vavilov, 1957; Frankel, 1970; Frankel, 1972; Harlan, 1972; Bennett, 1973; Hawkes, 1983; Veteläinen et al., 2009; FAO, 2011; Jarvis et al., 2011). Anecdotal evidence and the few LR cultivation reviews undertaken (Veteläinen et al., 2009; Raggi et al., 2022) indicate, despite these resources being

a crucial basis for future food security, LR genetic diversity is highly threatened, subject to genetic erosion and extinction, and LR genetic diversity is inadequately conserved therefore unavailable to farmers and breeders for use. However, this general reality differs between LR in crop gene pools and/or geographies (Khoury et al., 2021), and to date there are very few efforts involving systematic PGR *in situ*, even less for LR populations on-farm, monitoring. In this context, it is unnecessarily difficult to plan and conserve LR diversity.

A pivotal factor is the lack of an objective and repeatable method for LR threat assessment is significantly impeding effective conservation planning and implementation, and unavailable LR resources cannot be used (Veteläinen et al., 2009). The intrinsic characteristics of LR, notably the range of diversity/numbers of extant LR, non-standardised nomenclature, lack of comprehensive national LR inventories and the fact that LR populations are maintained by primarily farmers and cultivation is subject to prevailing food systems and market forces, and not conservationists with a single focus on conserving the resource, each makes them a challenging subset of biodiversity to threat assess. The fact that LR conservation focuses on an entirely human-managed resource, not a wild species governed by ecological laws and existing regulatory frameworks, as well as the need to focus conservation at the genetic and not species level, means the straight adaptation of the IUCN Red List method is inappropriate for LR threat assessment and this derived method is urgently needed.

What is presented is a standardised and repeatable method for LR threat assessment derived using the principles that underlie IUCN Red Listing. Initial unpublished case studies testing demonstrates the methodology indicates it is relatively simple to apply, is applicable for multiple crops at multi-national, national, or local levels and would therefore meet the confirmed requirement for an aid to crop and LR conservation planning. However, undoubtedly, the LR threat assessment method proposed requires ground truthing and refinement through actual application on diverse crops in diverse global localities to enhance its value. The current authors are undertaking this task at present. Therefore, it is stressed that what is presented here is version 1 of a LR threat assessment methodology. Just like the IUCN Red List methodology itself it is likely the LR methodology will pass through several revisions following initial practical applications.

In terms of method revision, it is likely that the percentage scores necessary for triggering the seven subcriteria scorable categories and the appropriate time interval for assessment of several of the subcriteria (over 10 generations is proposed here) may need to be revised following practical implementation. Similarly, some subcriteria, such as A2.3 (Geographic constancy), A3.1 (LR phenotypic diversity), and B1.1 (Ease of multiplication) may prove difficult to score practically, if the LR maintainer cannot supply the information needed and those that regularly remain unscorable should be possibly dropped. It is also hoped that practical LR assessments will identify potential additional subcriteria that could be reviewed and possibly added to the methodology. It should also be noted that threat category identification is not the last stage in the process of IUCN Red Listing, once the appropriate category has been proposed the draft

<p>LR maintainer information</p> <p>Name: _____ Date: ____/____/____</p> <p>Altitude: _____ m. Lat.: _____ N/S Long.: _____ W/E</p> <p>ID: _____ Maintainer gender: M/F Contact later: Y/ N</p> <p>Address: _____</p> <p>Locality: _____ Parish: _____</p> <p>Municipality: _____ Country: _____</p> <p>Tel.: _____ Mobile: _____</p> <p>Email: _____</p> <p>Name of grower association? _____</p> <p>How many LR generations have you or your extended family been maintaining the assessed LR? _____</p> <p>Do you work in other activity besides agriculture? Y/ N</p> <hr/> <p>A. LR Population Range</p> <p>Cultivation EOO of assessed LR 10 generations ago: <1 km² 1–5 km² 6–15 km² 16–25 km² ≥26 km²</p> <p>Cultivation EOO of assessed LR today: <1 km² 1–5 km² 6–15 km² 16–25 km² ≥26 km²</p> <p>Cultivation AOO of assessed LR 10 generations ago: <0,5 km² 0,5–1 km² 2–3 km² 4–10 km² ≥10 km²</p> <p>Cultivation AOO of assessed LR today: <0,5 km² 0,5–1 km² 2–3 km² 4–10 km² ≥10 km²</p> <p>Number assessed LR maintainers 10 gen. ago: _____</p> <p>Number assessed LR maintainers today: _____</p> <p>Plant phenotypic variability observation(s): _____</p> <p>_____</p> <p>Assessed LR exchanged – local community: _____ Y/ N</p> <p> – outside local community: Y/N</p>	<p>B. LR Population Trend</p> <p>LR easily multiplied? _____ Y/ N</p> <p>Are family members willing to maintain LR? _____ Y/ N</p> <p>Number of other LR known to be lost in the same area as the LR being assessed: _____</p> <p>Number modern cultivars of same crop cultivated: _____</p> <p>Nos. people in the locality maintaining assessed LR? _____</p> <hr/> <p>C. Market and farmer characteristics</p> <p>Assessed LR support applied: PDO / TSG / other(s) _____</p> <p>_____</p> <p>Assessed LR market range: national / regional / local.</p> <p>Maintainer age: _____</p> <hr/> <p>D. LR context</p> <p>Conserved <i>in situ</i> assessed LR population maintained: 1–4 pop. on-farm ≥5 pop. on-farm 1–4 pop. conserved ≥5 pop. conserved</p> <p>Conserved <i>ex situ</i>: No conservation 1–4 pop. conserved ≥5 conserved.</p> <p>Type of cultivation: Traditional / Commercial, industrial</p> <table border="0" style="width: 100%;"> <tr> <td style="width: 70%;">Use of chemical herbicides:</td> <td style="text-align: right;">Y/ N</td> </tr> <tr> <td style="padding-left: 150px;">fungicides:</td> <td style="text-align: right;">Y/ N</td> </tr> <tr> <td style="padding-left: 150px;">fertilizers:</td> <td style="text-align: right;">Y/ N</td> </tr> </table> <p>Incentives/benefits to use modern cultivars: _____ Y/ N</p> <p>Stochastic impacts: floods / droughts / wildfires / other(s)</p> <hr/> <p>Observation</p> <p>Cultural practices: Irrigation / rotation / organic fertilizers / inorganic fertilizers / animal traction / mechanization / LR exchange</p>	Use of chemical herbicides:	Y/ N	fungicides:	Y/ N	fertilizers:	Y/ N
Use of chemical herbicides:	Y/ N						
fungicides:	Y/ N						
fertilizers:	Y/ N						

4 Conclusion

The proposed LR threat assessment method presents a first-of-a-kind standardised protocol that can be used globally: in different countries, regions, and with different crops. It would be helpful for the LR threat assessment method to be further evaluated in other

global regions and on a full range of crops to see if it is as universal as it currently appears. Nonetheless, the growing LR community interest in developing such a robust threat assessment methodology supports the general need to activate a network for systematic LR monitoring for key crop gene pools globally, to aid their systematic conservation, extend farmer/breeder LR usage and help provide global food and nutritional security.

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

MA: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. AB: Investigation, Methodology, Writing – review & editing. SD: Investigation, Methodology, Writing – review & editing. BJ: Methodology, Writing – review & editing. JB: Investigation, Methodology, Writing – review & editing. MY: Investigation, Methodology, Writing – review & editing. NM: Conceptualization, Investigation, Methodology, 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. Partial

research funding was provided by the Horizon 2020 Framework Programme of the European Union through the Networking, Partnerships, and Tools to Enhance *in situ* Conservation of European Plant Genetic Resources (Farmer's Pride) project 774271. Costs for open-access publishing were funded by the University of Birmingham.

Acknowledgments

We would like to thank Ehsan Dulloo and Valeria Negri for fruitful discussions of the topic, the staff of the IUCN Red List unit in Cambridge, UK for their support, and reviewers for their careful reading and helpful comments on 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.

References

- Alercia, A., Diulgheroff, S., and Mackay, M. (2015) *FAO/Bioversity Multi-Crop Passport Descriptors (MCPD V.2.1)*. Rome: Bioversity International. pp. 1–11.
- Almeida, M. J., Pinheiro de Carvalho, M.Â.A., Barata, A. M., Magos Brehm, J., and Maxted, N. (2023). Crop landraces inventory for Portugal. *Genet. Resour. Crop Ev.* 70, 1151–1161. doi: 10.1007/s10722-022-01492-6
- Antofie, M.-M., Sand, M., Ciotea, G., and Iagrăru, P. (2010). Data sheet model for developing a Red List regarding crop landraces in Romania. *Ann. Food Sc. Tech.* 11, 45–49.
- Bachman, S., Moat, J., Hill, A., de la Torre, J., and Scott, B. (2011). Supporting Red List threat assessments with GeoCAT: Geospatial Conservation Assessment Tool. *ZooKeys* 150, 117–126. doi: 10.3897/zookeys.150.2109
- Bennett, E. (1971). "The origin and importance of agroecotypes in south-east Asia," in *Plant life in south-east Asia*. Eds. P. H. Davis, P. C. Harper and I. C. Hedge (Botanical Society of Edinburgh, Edinburgh).
- Bennett, E. (1973). *Survey of crop genetic resources in their centres of diversity: first report*. Ed. O. H. Frankel (Rome: FAO/IBP).
- Brown, A. H. D., and Briggs, J. D. (1991). "Sampling strategies for genetic variation in ex situ collections of endangered plant species," in *Genetics and conservation of rare plants*. Eds. D. A. Falk and K. E. Holsinger (Oxford university Press, New York), 99–119.
- Brush, S. (2000). "The issues of *in situ* conservation of crop genetic resources," in *Genes in the field: on-farm conservation of crop diversity*. Ed. S. B. Brush (International Development Research Centre and International Plant Genetic Resources Institute, Boca Raton, Lewis Publishers).
- Camacho-Villa, T., Maxted, N., Scholten, M., and Ford-Lloyd, B. (2005). Defining and identifying crop landraces. *Plant Genet. Resour.-C.* 3, 373–384. doi: 10.1079/PGR200591.
- Das, B., Sengupta, S., Parida, S. K., Roy, B. P., Ghosh, M., Prasad, M., et al. (2013). Genetic diversity and population structure of rice landraces from Eastern and North Eastern States of India. *BMC Genet.* 14, 71. doi: 10.1186/1471-2156-14-71
- de Boef, W. S., Subedi, A., Peroni, N., Thijssen, M., and O'Keeffe, E. (2013). *Community biodiversity management: promoting resilience and the conservation of plant genetic resources* (London: Earthscan / Routledge). doi: 10.4324/9780203130599
- de Haan, S. (2021). "Community-based conservation of crop genetic resources," in *Plant Genetic Resources: a review of current research and future needs*. Ed. E. Dulloo (Burleigh Dodds Science Publishing, Cambridge).
- de Haan, S., Burgos, G., Liria, R., Rodriguez, F., Creed-Kanashiro, H. M., and Bonierbale, M. (2019). The nutritional contribution of potato varietal diversity in Andean food systems: a case study. *Am. J. Potato Res.* 2, 151–163. doi: 10.1007/s12230-018-09707-2
- de Haan, S., Polreich, S., Rodriguez, F., Juarez, H., Ccanto, R., Alvarez, C., et al. (2016). "A long-term systematic monitoring framework for on-farm conserved potato landrace diversity. pp. 289–296," in *Enhancing Crop Genepool Use: capturing wild relative and landrace diversity for crop improvement*. Eds. N. Maxted, E. Dulloo and B. V. Ford-Lloyd (CABI International, Wallingford).
- Ellstrand, N. C. (2003). *Dangerous liaisons? when cultivated plants mate with their wild relatives* (Baltimore, Maryland: John Hopkins University Press).
- FAO. (2010) *Second report on the State of the World's Plant Genetic Resources for Food and Agriculture*, (Rome, Italy: Food and Agriculture Organization of the United Nations). Available online at: <https://www.fao.org/agriculture/seed/sow2/en/>.
- FAO. (2011) *Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture* (Rome, Italy: Food and Agriculture Organization of the United Nations). Available online at: <http://www.fao.org/docrep/015/i2624e/i2624e00.htm> (Accessed 13.09.2023).

- FAO, IFAD, UNICEF, WFP and WHO (2021) *The state of food security and nutrition in the world 2017: Building resilience for peace and food security* (Rome: FAO. Food and Agriculture Organization of the United Nations, Rome, Italy). Available online at: <https://www.fao.org/policy-support/tools-and-publications/resources-details/en/c/1107528/> (Accessed 13.09.2023).
- Fliedel, G., Koreissi, Y., Boré, F., Dramé, D., Brouwer, I., and Ribeyre, F. (2013). Sensory diversity of fonio landraces from West Africa. *Afr. J. Biotechnol.* 12, 1836–1844. doi: 10.5897/AJB.
- Fonseca, J. (2004). *Colher para Semear - Manual prático para a colheita e conservação de sementes* (Lisbon, Portugal: Colher para Semear).
- Frankel, O. H. (1970). Genetic danger in the green revolution. *World Ag* 19, 9–14.
- Frankel, O. H. (1972). Genetic conservation - A parable of the scientist's social responsibility. *Search* 3, 193–201.
- Frankel, O. H. (1973). *Survey of crop genetic resources in their centres of diversity: first report*. Ed. O. H. Frankel (Rome: FAO/IBP).
- Frankel, O., Brown, A., and Burdon, J. (1995). *The conservation of plant biodiversity* (Cambridge: Cambridge University Press).
- Frankel, O., and Hawkes, J. (1975). "Genetic Resources – the past ten years and the next," in *Crop Genetic Resources for today and tomorrow*. Eds. O. H. Frankel and J. G. Hawkes (Cambridge University Press, Cambridge).
- Hammer, K. (1991). *Checklist and germplasm collecting* Vol. 85 (FAO/IBPGR. PGR Newsletter, Rome Italy), 15–17.
- Hammer, K., and Khoshbakht, K. (2005). Towards a "red list" for crop plant species. *Genet. Resour. Crop Ev.* 52, 249–265. doi: 10.1007/s10722-004-7550-6.
- Hammer, K., Knapf, H., Xhuvli, L., and Perrino, P. (1996). Estimating genetic erosion in landraces – Two case studies. *Genet. Resour. Crop Ev.* 43, 329–336. doi: 10.1007/BF00132952.
- Hanelt, P. (2001). *Mansfeld's Encyclopaedia of Agricultural and Horticultural Crops* Vol. 6 (Berlin: Springer). doi: 10.1007/978-3-540-30442-5.
- Harlan, J. R. (1972). Genetics of disaster. *J. Env. Qual.* 1, 212–215. doi: 10.2134/jeq1972.00472425000100030002x.
- Harlan, J. (1975). Our vanishing genetic resources. *Science* 188, 617–621. doi: 10.1126/science.188.4188.617.
- Hawkes, J. G. (1983). *The diversity of crop plants* (Cambridge, Mass. USA: Harvard University Press). doi: 10.4159/harvard.9780674183551.
- Hayward, M., Child, M., Kerley, G., Lindsey, P., Somers, M., and Burns, B. (2015). Ambiguity in guideline definitions introduces assessor bias and influences consistency in IUCN Red List status assessments. *Front. Ecol. Evol.* 3. doi: 10.3389/fevo.2015.00087
- IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Eds. R. K. Pachauri and L. A. Meyer (Geneva, Switzerland: International Panel for Climate Change). Core Writing Team 151pp.
- IUCN (2001). *IUCN Red List Categories and Criteria: Version 3.1* (IUCN, Gland, Switzerland and Cambridge, UK: IUCN Species Survival Commission).
- IUCN (2012). *IUCN Red List Categories and Criteria: Version 3.1. 2nd ed.* (IUCN, Gland, Switzerland and Cambridge, UK: IUCN Species Survival Commission).
- Jain, S. K. (1961). Studies on the breeding of self-pollinated cereals. The composite cross bulk population method. *Euphytica* 10, 315–324. doi: 10.1007/BF00039102.
- Jarvis, D. I., Hodgkin, T., Sthapit, B., Fadda, C., and Lopez-Noriega, I. (2011). An heuristic framework for identifying multiple ways of supporting the conservation and use of traditional crop varieties within the agricultural production system. *Crit. Rev. Plant Sci.* 30, 125–176. doi: 10.1080/07352689.2011.554358.
- Jarvis, A., Upadhyaya, H., Gowda, C. L. L., Aggarwal, P. K., Fujisaka, S., and Anderson, B. (2010). *Climate change and its effect on conservation and use of plant genetic resources for food and agriculture and associated biodiversity for food security* (Rome, FAO: ICRISAT/FAO. Thematic background study for the Second Report on The State of the World's Plant Genetic Resources for Food and Agriculture).
- Joshi, B., Upadhyay, M. P., Gauchan, D., Sthapit, B. R., and Joshi, K. D. (2004). Red listing of agricultural crop species, varieties, and landraces. *Nep. Agr. Res. J.* 5, 73–80.
- Kell, S. P., Ford-Lloyd, B. V., Magos Brehm, J., Iriondo, J. M., and Maxted, N. (2017). Broadening the base, narrowing the task: prioritizing crop wild relative taxa for conservation action. *Crop Sci.* 57, 1042–1058. doi: 10.2135/cropsci2016.10.0873
- Kell, S. P., Maxted, N., Allender, C., Astley, D., and Ford-Lloyd, B. (2009). *Vegetable landrace inventory of England and Wales. Unpublished report* (Birmingham, UK: The University of Birmingham), 117. pp.
- Khoury, C. K., Brush, S. B., Costich, D. E., Curry, H. A., De Haan, S., Engels, J. M. M., et al. (2021). Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytol.* 233, 84–118. doi: 10.1111/nph.17733
- Litrico, I., and Violle, C. (2015). Diversity in Plant Breeding: A new conceptual framework. *Trends Plant Sci.* 20, 604–613. doi: 10.1016/j.tplants.2015.07.007.
- Mace, G. M., Collar, N. J., Gaston, K. J., Hilton-Taylor, C., Akçakaya, H. R., Leader-Williams, N., et al. (2008). Quantification of extinction risk: IUCN's system for classifying threatened species. *Conserv. Biol.* 22, 1424–1442. doi: 10.1111/j.1523-1739.2008.01044.x.
- Mansfeld, R. (1951). Das morphologische System des Saatweizens, *Triticum aestivum* L. s.l. - *Züchter*. 21, 41–60.
- Maxted, N. (2006). UK land-races – a hidden resource? *Plant Talk* 44, 8–10.
- Maxted, N., Hunter, D., and Ortiz Rios, R. O. (2020). *Plant genetic conservation* (Cambridge: Cambridge University Press), 560. pp.
- Maxted, N., Magos Brehm, J., and Kell, S. (2013). *Resource book for preparation of national conservation plans for crop wild relatives and landraces* (Birmingham, UK: University of Birmingham). Available online at: http://www.fao.org/fileadmin/templates/agphome/documents/PGR/PubPGR/ResourceBook/TEXT_ALL_2511.pdf (Accessed 13.09.2023).
- Maxted, N., and Scholten, M. A. (2007). "Methodologies for the creation of National / European inventories," in *Report of a Task Force on On-farm Conservation and Management, Second Meeting, 19–20 June 2006, Steglitz, Germany*. Eds. A. Del Greco, V. Negri and N. Maxted (Biodiversity International, Rome, Italy), 11–19.
- Negri, V. (2003). Landraces in central Italy: where and why they are conserved and perspectives for their on-farm conservation. *Genet. Resour. Crop Ev.* 50, 871–885. doi: 10.1023/A:1025933613279.
- Negri, V. (2005). Agro-biodiversity conservation in Europe: ethical issues. *J. Agr. Environ. Ethic.* 18, 3–25. doi: 10.1007/s10806-004-3084-3.
- Negri, V., Maxted, N., and Veteläinen, M. (2009). "European landrace conservation: an introduction," in *European Landraces: On-farm conservation, Management and Use. Biodiversity Technical Bulletin*, vol. 15. Eds. M. Veteläinen, V. Negri and N. Maxted (Biodiversity International, Rome, Italy), 1–22.
- Ortman, T., Sandström, E., Bengtsson, J., Watson, C. A., and Bergkvist, G. (2023). Farmers' motivations for landrace cereal cultivation in Sweden. *Biol. Agric. Hortic.* 39 (4), 247–268. doi: 10.1080/01448765.2023.2207081
- Padulosi, S., and Dulloo, M. E. (2012). "Towards a viable system for monitoring agrobiodiversity on-farm: A proposed new approach for Red Listing of cultivated plant species," in *On farm conservation of neglected and underutilized species: status, trends, and novel approaches to cope with climate change. Proceedings of an International Conference, Frankfurt, 14–16 June 2011*. Eds. S. Padulosi, N. Bergamini and T. Lawrence (Biodiversity International, Rome, Italy), 171–187.
- Perales, H. R., Benz, B. F., and Brush, S. B. (2005). Maize diversity and ethnolinguistic diversity in Chiapas, Mexico. *P. Natl. Acad. Sci. U.S.A.* 102, 949–954. doi: 10.1073/pnas.0408701102.
- Plasencia, F., Juárez, H., Polreich, S., and De Haan, S. (2018). Assessment of the spatial distribution of potato biodiversity in the districts of Challabamba in Cusco and Quilcas in Junin through the use of participatory mapping. *Rev. del Instituto Investigaciones la Facultad Geología Minas Metalurgia y Cienc. Geográfica* 21, 17–24.
- Porfiri, O., Costanza, M., and Negri, V. (2009). "Landrace inventories in Italy and the Lazio region case study," in *European landraces on-farm conservation, management, and use. Biodiversity Technical Bulletin No. 15*. Eds. M. Veteläinen, V. Negri and N. Maxted (Biodiversity International, Rome, Italy).
- Raggi, L., Ciro Pacifico, L., Caproni, L., Álvarez-Muñoz, C., Annamaa, K., Maria Barata, A., et al. (2022). Analysis of landrace cultivation in Europe: a means to support in situ conservation of crop diversity. *Biol. Conserv.* 267, 109460. doi: 10.1016/j.biocon.2022.109460.
- Scholten, M., Maxted, N., Ford-Lloyd, B. V., and Green, N. (2008). *Hebridean and Shetland oat (Avena strigosa Schreb.) and Shetland cabbage (Brassica oleracea L.) landraces: occurrence and conservation issues. PGR Newsletter*. FAO/IBPGR, Rome Italy 154:1–5
- Sinskaya, E. N. (1969). Historical geography of cultivated floras (at the dawn of agriculture). *Leningrad, (USSR: Kolos)*.
- UK Parliament (2020) *The UK Agriculture Act 2020. London, House of Commons Library*. Available online at: <https://commonslibrary.parliament.uk/research-briefings/cbp-8702/> (Accessed 13.09.2023).
- United Nations (2021). *World Population Prospects: The 2020 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP.241* (New York, United Nations: Department of Economic and Social Affairs, Population Division).
- Vavilov, N. I. (1926). Studies on the origin of cultivated plants. *Tr. po prikl. bot. Gen. i sel.* 16, 3–248.
- Vavilov, N. I. (1931). The Linnean species as concept. *Tr. po prikl. bot. Gen. i sel.* 26, 109–134.
- Vavilov, N. I. (1957). *World resources of cereals, leguminous seed crops, and flax, and their utilization in plant breeding. Agro-ecological survey of the principal field crops* (Leningrad: USSR Academy of Science Press).
- Veteläinen, M., Negri, V., and Maxted, N. (2009). "A European strategic approach to conserving crop landraces," in *European landraces on-farm conservation, management, and use. Biodiversity Technical Bulletin No. 15*. Eds. M. Veteläinen, V. Negri and N. Maxted (Biodiversity International, Rome).
- Voegel, R. (2012). "Red List for crops – a tool for monitoring genetic erosion, supporting re-introduction into cultivation and guiding conservation efforts," in *On farm conservation of neglected and underutilized species: status, trends, and novel approaches to cope with climate change. Proceedings of an International Conference, Frankfurt, 14–16 June 2011*. Eds. S. Padulosi, N. Bergamini and T. Lawrence (Biodiversity International, Rome).
- K. Walter and H. Gillett (Eds.) (1998). *1997 IUCN Red List of Threatened Plants compile by the World Conservation Monitoring Centre*. Gland (Switzerland and Cambridge, U.K.: IUCN - The World Conservation Union), 862.
- Zeven, A. C. (1998). Landraces: A review of definitions and classifications. *Euphytica* 104, 127–139. doi: 10.1023/A:1018683119237.



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Benoit Bizimungu,
Agriculture and Agri-Food Canada (AAFC),
Canada
Photini V. Mylona,
Hellenic Agricultural Organisation (HAO),
Greece

*CORRESPONDENCE

Alfonso H. del Rio
✉ adelrioc@wisc.edu

RECEIVED 20 December 2023

ACCEPTED 12 February 2024

PUBLISHED 05 March 2024

CITATION

Arcos-Pineda JH, del Rio AH, Bamberg JB,
Vega-Semorile SE, Palta JP, Salas A,
Gomez R, Roca W and Ellis D (2024)
An international breeding project using
a wild potato relative *Solanum
commersonii* resulted in two new
frost-tolerant native potato cultivars
for the Andes and the Altiplano.
Front. Plant Sci. 15:1358565.
doi: 10.3389/fpls.2024.1358565

COPYRIGHT

© 2024 Arcos-Pineda, del Rio, Bamberg,
Vega-Semorile, Palta, Salas, Gomez, Roca and
Ellis. 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.

An international breeding project using a wild potato relative *Solanum commersonii* resulted in two new frost-tolerant native potato cultivars for the Andes and the Altiplano

Jesus H. Arcos-Pineda¹, Alfonso H. del Rio^{2,3*}, John B. Bamberg²,
Sandra E. Vega-Semorile³, Jiwan P. Palta³, Alberto Salas⁴,
Rene Gomez⁴, William Roca⁴ and David Ellis⁴

¹Instituto Nacional de Innovación Agraria (INIA), Estación Experimental Agrícola (EEA) Illpa-CE Salcedo, Puno, Peru, ²U.S. Department of Agriculture (USDA)/Agricultural Research Service, Potato Genebank, Sturgeon Bay, WI, United States, ³Department of Plant and Agroecosystem Sciences, University of Wisconsin—Madison, Madison, WI, United States, ⁴International Potato Center (CIP), Lima, Peru

This breeding project, initiated at the United States Potato Genebank (USPG) in collaboration with Peruvian partners Instituto Nacional de Innovación Agraria (INIA), International Potato Center, Peru (CIP), and local farmers, sought to enhance cold hardiness and frost tolerance in native potato cultivars in Peru. The Andes and Altiplano are often affected by frost, which causes significant reduction in yield; creating varieties with superior resilience is a critical undertaking. The goal was to transfer outstanding non-acclimated cold tolerance and acclimation capacity found in wild potato species *Solanum commersonii* (*cmm*). Breeding families segregating for cold hardiness were created using (a) a somatic hybrid *cmm* + haploid *Solanum tuberosum* (*tbr*) (cv. Superior, US variety from Wisconsin) as male and (b) seven cultivars native to Peru of the species *S. tuberosum* sbsp. *andigenum* (*adg*) as females. All plant materials were part of the USPG germplasm collection. Sexual seeds of each family were sent to Peru for evaluations under the natural conditions of the Andean highlands and Altiplano. The plants were assessed for their response to frost, and genotypes showing exceptional tolerance were selected. Plants were also evaluated for good tuber traits and yield. Initial planting involving ~2,500 seedlings in five locations resulted in selecting 58 genotypes with exceptional frost tolerance, good recovery capacity after frost, and good tuber traits. Over the years, evaluations continued and were expanded to replicated field trials in the harsher conditions of the Altiplano (Puno). All trials confirmed consistency of frost tolerance over time and location, tuber quality, and yield. After 8 years, two advanced clones were considered for cultivar release because of their exceptional frost tolerance and superior field productivity that outyielded many of the established cultivars in the region. In November 2018, a new native cultivar named *Wiñay*, a Quechua word meaning “to grow” was released in Peru. In 2022, a second cultivar followed with the name *Llapanchispaq*

(meaning “for all of us”). This project evidenced that a multinational and all-encompassing approach to deploy valuable genetic diversity can work and deliver effective results. This is even more significant when outcomes can promote food security and sustainability in very vulnerable regions of the world.

KEYWORDS

andean potato cultivars, climate change, benefit sharing, frost tolerance, genetic diversity, potato wild relatives, *Solanum commersonii*, U.S. Potato Genebank

Introduction

In the Andes and the Altiplano of Peru and Bolivia, frost is responsible for serious damage to agriculture affecting many smallholders in a situation of high vulnerability, i.e., subsistence potato farmers in conditions of poverty or extreme poverty (Condori et al., 2014). Approximately 74% of the agricultural communities there are exposed to frost (Hijmans, 1999). Though the damages can be irreversible, potato farmers try to counteract these impacts through the ancient tradition of planting mixtures of many native cultivars expecting that some can survive and produce. For example, cultivars of cultivated potato species *Solanum juzepczukii* and *Solanum curtilobum* have excellent cold hardiness and are part of the farmers’ cultivar pool. However, their tubers are bitter due to high glycoalkaloid content and cannot be eaten fresh unless the ancient freeze-drying processes of *chuño* and *tunta* are used to remove alkaloids (Brush et al., 1981). It is estimated that approximately 25% of all potatoes in the Altiplano are bitter; hence, adding frost-tolerant non-bitter cultivars could be a very significant contribution. In that way, farmers can consume the tubers and/or place them in regional markets immediately with no need for additional processing to eliminate bitterness.

Screening and identifying crop germplasm expressing genetic traits with enhanced resilience to abiotic stresses is especially relevant today. The impact of climate change in agriculture is serious and responsible for food shortages and economic distress (Morton, 2007). The possibility of incorporating valuable genetic diversity to create new varieties with enhanced adaptation to new climates must be considered a strategy to prioritize. Potato germplasm, which includes the potato crop and its wild relatives, is an outstanding source of genetic resistance to different pests and diseases and to extreme tolerance to abiotic stresses (Li and Palta, 1978; Palta and Li 1979; Hanneman, 1989; Tiwari et al., 2022). In particular, the wild relative species *Solanum commersonii* (*cmm*), endemic to Uruguay, Argentina, and Brazil, possesses outstanding freezing tolerance and ability to cold acclimate—two important

genetic traits that, if transferred to cultivated potatoes, can make a large impact in resilience to low temperatures and frost (Li and Palta, 1978; Palta, 1991; Stone et al., 1993; Vega and Bamberg, 1995; Aversano et al., 2015; Cho et al., 2016). Every year, the potato experiences major yield losses worldwide due to frost (Li and Palta, 1978; Palta and Li 1979). The good news is that major increases in frost tolerance are not needed to have significant impacts. Hijmans et al. (2003a, b) estimated that if cold tolerance only increases by 1° C or 2°C, potato yields can be improved by 26% and 40%, respectively.

Because of its outstanding cold hardiness, germplasm enhancement and breeding programs investigated the possibility of using *cmm* to enhance freezing tolerance. However, it was found that some wild potato species, including *cmm*, are not possible for use in crossbreeding because of their reproductive incompatibilities. Johnston et al. (1980) reported that inter-specific crosses between *cmm* and cultivated potatoes failed because of post-zygotic barriers. This resulted in investigating other options to overcome cross incompatibility. Hence, it was discovered that techniques of somatic fusion of leaf protoplasts were effective to overcome incompatibility and combine the genes of otherwise incompatible species. The fact that potatoes can be clonally propagated was important because all useful gene combinations established in the somatic hybrids were maintainable. Mattheij and Puite (1992) and Millam et al. (1997) showed that methods of protoplast fusion and somatic hybridization were effective in generating viable somatic hybrids between *cmm* and other incompatible species. Moreover, Cardi et al. (1993) generated *cmm* + *tbr* hybrids with improved levels of frost hardiness and acclimation capacity proving that introgression of *cmm* was effectively attained and expressed. Later, field evaluations of selfed and backcrossed progenies derived from somatic hybrids *cmm* + *tbr* were conducted in the US, which resulted in finding lines with excellent tuber quality and production (Millam et al., 1995; Chen et al., 1999a, b, c). This unlocked the idea of testing them in Peru with the expectation that their remarkable cold hardiness and productivity could be replicated in the Peruvian highlands. However, they gave poor yields as they were not adapted to the shorter daylength of the Andes (because of the *tbr* background). That setback led to a new plan of developing breeding lines for cold hardiness in Peru, but this time utilizing the well-adapted native cultivars of *adg* as the parental lines.

Abbreviations: *acl*, *Solanum acaule*; *adg*, *Solanum tuberosum* sbsp. *andigenum*; CIP, International Potato Center, Peru; *cmm*, *Solanum commersonii*; INIA, Instituto Nacional de Innovación Agraria, Peru; *tbr*, *Solanum tuberosum*; USPG, The United States Potato Genebank.

TABLE 1 List of germplasm used as parents to generate frost-tolerant families at the USPG.

Family Code	Female parent	Local name of the native cultivar (if reported in GRIN database)
	USPG PI numbers*	
H1	281078	
H2	281065	
H3	292086	<i>Ccompis</i>
H4	281070	
H5	308886 x 308885	<i>Color uncuna x Chaquillo</i>
H6	246555	<i>Suytta</i>
H7	281063	

Some of the *adg* accessions have a local cultivar name in Peru listed in the database and all originated in Cusco, Peru. The same male parent (somatic hybrid *cmm* + *tbr*; PI number GS393) was used to generate each family.

*Additional information is available through the Germplasm Resources Information Network (GRIN) <https://npgsweb.ars-grin.gov/gringlobal/search>.

Here, we present the results of an international breeding project initiated at the USPG that examined the prospects of using the potato wild relative *cmm* to transfer enhanced cold hardiness, cold acclimation capacity, and frost tolerance into potato cultivars of *S. tuberosum* subsp. *andigenum*, the potato cultivated in the Andean and the Altiplano regions. From a broader perspective, this project was a model of international cooperation to promote participatory work and share benefits in the utilization of genetic resources.

Materials and methods

Development of frost-tolerant breeding families

The parental materials used to generate the breeding families were (1) a somatic hybrid between *S. commersonii* + a haploid of *S. tuberosum* cv. Superior (USW-13122) developed in the Wisconsin breeding program (Chen et al., 1999a; Kim-Lee et al., 2005) and (2) seven native cultivars of the species *S. tuberosum* ssp. *andigenum* (*adg*) as female parents (Table 1). The somatic hybrid *cmm* + *tbr* is a genetic stock maintained in the USPG collection as accession number GS393. The accessions used as female *adg* parents were not the original cultivar clones, but the sexual progeny (seed lots) derived from them and maintained at the gene bank as botanical seeds. All plant materials were part of the USPG germplasm collection in Sturgeon Bay, Wisconsin, USA.

The F1 progenies generated were then subsequently bulk intermated yielding F2 populations. To estimate potential cold hardiness, assorted genotypes from the different families were assessed for relative freezing tolerance (RFT) in non-acclimated and acclimated stages. RFT was estimated in the laboratory by measurements of ion leakage of excised terminal leaflets subjected to ice nucleation and simulated freeze–thaw stress (Li and Palta, 1978).

RFT values in non-acclimated genotypes ranged from -2.4°C to 4.0°C , while in acclimated stage, they ranged from -3.0°C to -5°C . Seeds of these seven breeding families expected to segregate for freezing tolerance were sent to Peru in 2009 to be evaluated for their frost response under field conditions in the Andes and the Altiplano.

Selection of genotypes expressing frost tolerance

Field experiments

The initial field trial in 2009–2010 included approximately 2,500 randomly selected seedlings, which were a proportional share of several genotypes from each of the seven families. Plants were initially grown in greenhouses at the International Potato Center, Peru (CIP) in Lima, Peru. When the plants were approximately 15 cm, they were transported by car to the different fields in the highlands.

The distances to the fields from the city of Lima were between 300 and 1,000 km. Fields were in farmers' communities in the highlands, in the towns of *San Jose de Aymara*, *Huancavelica* (-12.243 , -75.051 ; 3,906 m); *Aramachay*, *Junin* (-11.916 ; -71.413 ; 3,687 m); *CCorao*, *Cusco* (-13.478 ; -71.923 ; 3,596 m); and *Kayra*, *Cusco* (-13.541 ; -71.888 ; 3,426 m).

All fields were rainfed except for *Kayra*, which is an experimental field in the University San Antonio de Abad in Cusco. In addition, the fields were in areas where freezing temperatures and unexpected frosts are common during the growing season. As this project promoted participatory work, the activities of land preparation, planting, hilling, selection, and evaluations were done with the help of local farmers at each community.

Assessment of plant damage to determine levels of frost tolerance

One week after the frost event, the fields affected were visited to assess the levels of damage in the plants and to estimate the survival rates. This 1-week waiting made more evident the effects of frost on the plants and facilitated the assessments and ratings. Symptoms of freezing damage in the plants ranged from minor ones, like darkening of the leaf margins, to more serious ones, like wilted and/or dead leaf and stem tissue. The magnitude of leaf and whole-plant injury was determined by visual estimations and reported as percentage of plant damage.

After 2 to 3 weeks, the affected fields were visited for a second time to assess the levels of plant recovery. At the end of the season, the genotypes rated with good levels of frost tolerance and ability to recover (and desirable tuber features) were selected for the next round of field trials. For field trials, the genotypes were coded with the acronym HSP, which stands for H=frost (as for the Spanish word *heladas*), S=row number (Spanish *surco*), and P=plant number in the row (Spanish *planta*). The number after H identified the family where the genotype originated (see Table 1).

Seedling selection and clonal generations

Seedling selection aimed to identify genotypes with superior tolerance to frost. Because of the *adg* background, the breeding lines were anticipated to be adapted and respond well to the traditional practices in the Andes. Adaptation implied plants with short-day tuberization and ability to grow in soils with low fertility and low pH. After seedling selection, clonal generations were generated for the following season of 2010–2011 using the tubers harvested as seeds. This sought to confirm if selections were consistent in exceptional frost responses and the expression of desirable tuber traits. Trials of clonal generation selections involved larger number of plants in the field, so the results were more representative as sampling variation was included in the analyses. Hence, the experimental field trials consisted of tuber seeds in single rows with up to five plants per clonal line. The same methods to assess plant injury, indicated above, were used to determine levels of frost damage in the plants. Yield was measured as tuber number and weight. Farmers from the local rural communities nearby participated in the evaluation and selection process.

Advanced selections and evaluations for cultivar release

Field trials for advanced selections were carried out under the harsher conditions found in experimental fields in Puno, in the Altiplano region, where crops are constantly negatively affected by environmental stresses. Gilles and Valdivia (2009) pointed out that because of frost and other abiotic stresses in the Altiplano, the potato never reaches its full yield potential. The experimental fields were in *Salcedo* (−15.868, −69.995, 3,838 m) and *Illpa* (−15.832, −70.019, 3,815 m). Both are part of the System of Agricultural Stations for Peru's National Program for Agriculture (INIA).

The advanced selections were planted together with other common native cultivars of the region as comparisons. All were evaluated in randomized complete block designs with up to five replicates per location. Plots consisted of single rows of up to 10 plants spaced at ~0.60 m between plants. At harvest, tubers of each plot were counted and weighed to determine the averaged tuber yield. Throughout the different years of field trials, the popular cultivars of the Altiplano region used were *Imilla Blanca*, *Imilla Negra*, *Ccompis*, *Mariva*, *Pucamama*, *Lekecho*, and *Oke* (Cahuana-Quispe and Arcos-Pineda, 2002).

In the final step, the advanced clones identified as potential new cultivars were validated under the INIA guidelines for cultivar release stated under Peru's National Policy number 047-2000-INIA (El Peruano, 2000). The guidelines indicate the procedures needed to test the potential cultivars for parameters a new plant variety must meet, namely, distinguishability, homogeneity, stability, validation, and homologation. The guidelines can be found in <https://www.fao.org/faolex/results/details/en/c/LEX-FAOC020078>.

The validation approach expected multilocation field trials of the potential cultivars, which must be grown with local cultivars as controls. These trials must be done in at least 10 different farmers'

communities in the region, and then, surveys must be conducted in a subset of the population at each community to confirm the support on accepting these clones as new cultivars for the region.

Statistical analysis

The results of the field evaluations for frost and yield followed an analysis of variance using JMP Pro 15 statistical software to get additional insights on the levels of differential responses among genotypes. A Tukey–Kramer LSD test was used to determine the significance of the differences between mean values at a probability level $p \leq 0.05$ (JMP® Pro, Version 15.0.0, 2019).

Results

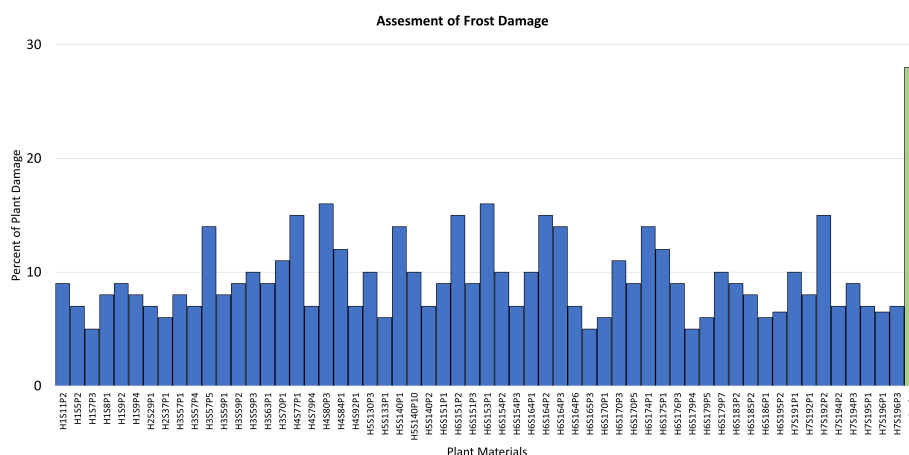
Detection of genotypes expressing good levels of frost tolerance

After a frost, fields were assessed to determine damage to seedlings. The extent of damage observed in the fields ranged broadly from the total loss of the plant to a minimum plant damage of ~5%. In general, frost episodes in that season were in the range of −2°C to −5°C. Only the genotypes with plant damage under 20% were considered for selection, and that produced a total of 58 genotypes. All of them showed plant damage levels under 20% and expressed recovery ability after a frost rated as either good or very good. Assessments for plant damage after a frost for those 68 selections are presented in Figure 1. To get an idea of the levels of cold hardiness, these genotypes were compared to a native potato cultivar of the region named *Locka*. This is a cultivar of the species *S. jusepczukii* characterized by bitter tubers but cold hardy and adapted to the Altiplano (Cahuana-Quispe and Arcos-Pineda, 2002). Though evaluations at harvest were limited to only one plant, the 58 selections were also chosen for their attractive tuber characteristics.

Assessment of frost tolerance and tuber yield in the first clonal generation

In the season of 2011–2012, field trials were focused on verifying the consistency of frost tolerance responses for the 58 genotypes selected. In addition, these trials helped to better characterize tuber traits and estimate yield potential. Tubers harvested in the previous season were used as seeds that allowed more plants per genotype (up to 10 plants) to be assessed. Trials were done at the experimental station of *Kayra* in Cusco and the cold hardy cultivar *Locka* was used again as a comparative control.

The data collected at harvest consisted of tuber number and tuber weight per plant, which gave insights on the levels of yield variability among genotypes and allowed identification of genotypes with outstanding productivity. The levels of variation for yield among genotypes were very significant ($p < 0.001$). The average tuber number per plant fluctuated between 1 and 54 (an average



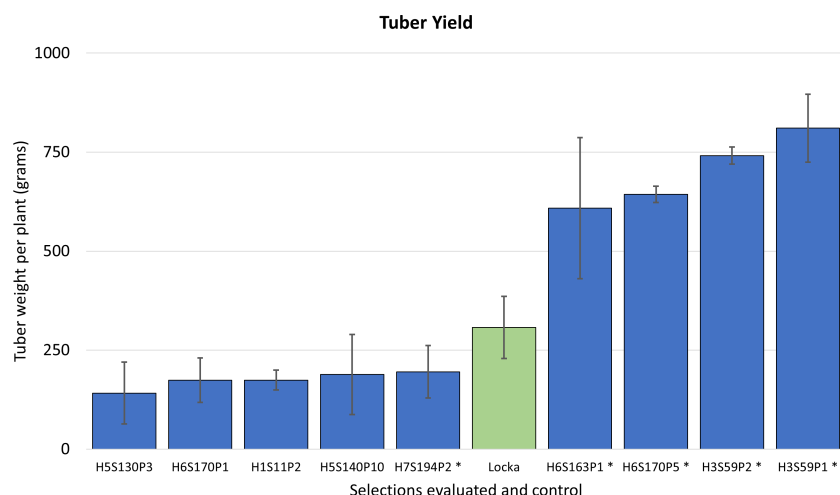


FIGURE 3

Yield (determined by average tuber weight per plant in grams) for the nine advanced selections evaluated in Salcedo and Illpa, Puno. Error bars represent standard deviations of the mean. Clones in asterisks are the five selections advancing to additional field trials and selection.

At harvest, it was found that some of these clones exhibited tuber yields that were significantly superior to the native cultivar used as the control ($p < 0.001$) (Figure 3). This resulted in narrowing down selections to five clones, coded as H3S59P1, H6S163P1, H7S194P2, H6S170P5, and H3S59P2, that were rated as exceptional and chosen to be evaluated as potential cultivars (Figure 3). It is important to note that selection H7S194P2 was not particularly exceptional for frost tolerance and/or yield. However, farmers in the Andes/Altiplano do not always consider high yields as the main priority in selecting for a cultivar. Our experience of many years working with local farmers made us aware that varieties with attractive tuber features, including taste, are sometimes preferred over high yielders. In this case, the farmers

considered that the clone H7S194P2 satisfied those attributes and advanced for additional testing.

Comparison of selections to regional cultivars

Evaluations conducted in season 2014–2015, in replicated trials, included the five advanced clones selected in the previous season plus additional native potato cultivars common to the region. The cultivars included were *Lekecho*, *Mariva*, *Imilla Negra*, *Oke*, *Pucamama*, and *Imilla Blanca*. All are very common in stores and local farmers' markets in the Altiplano and Southern Peru and

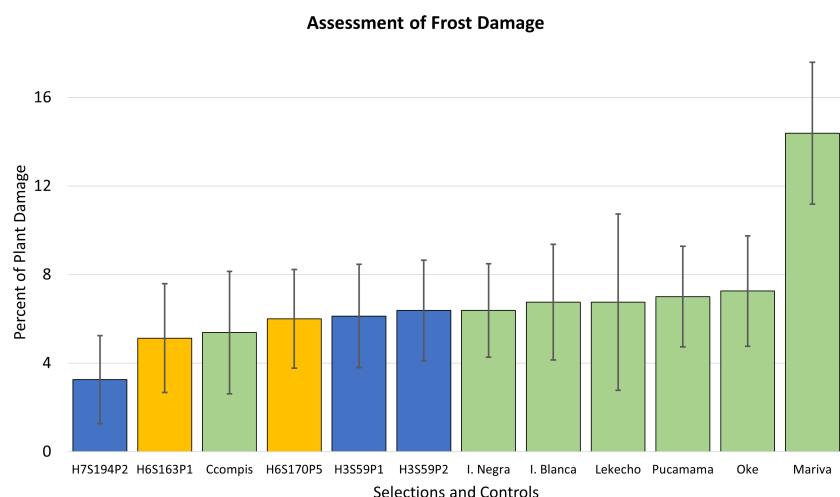


FIGURE 4

Frost damage scores are shown in percent of plant damage for five advanced frost-tolerant selections, which are compared to seven native cultivars from the region. Clones in yellow bars are the advanced selections that later resulted in cultivar releases. Error bars represent standard deviations of the mean.

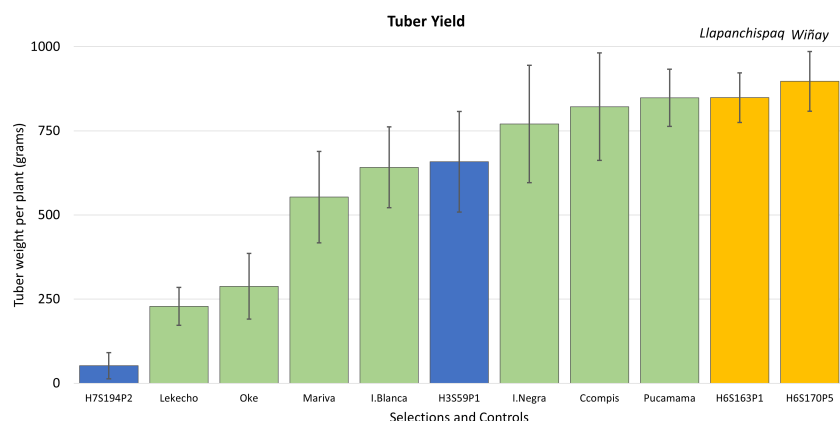


FIGURE 5

Yield (determined by tuber weight in grams per plant) from harvest data in four advanced selections and seven native cultivars evaluated in the fields of Salcedo and Illpa, Puno. Error bars are the standard deviations of the mean value. Yellow bars are the advanced clones later released as new cultivars; on top of the bar, their given released cultivar names are indicated.

have good levels of cold hardiness. The cultivar *Imilla Negra* is one of the most popular commercial cultivars in that region.

The results of evaluations in the fields of Puno confirmed that the five advanced selections responded well to the climatic conditions of the region and exhibited excellent levels of frost tolerance. The scores of plant damage were in the range of 3% to 6%, and the traditional cultivars also exhibited good tolerance in the range of 7%–8% of plant damage except the cultivar *Mariva* with 14% (Figure 4).

At harvest, all advanced clones (except for H7S194P2) showed good tuber yields that were comparable to or better than those in the traditional cultivars. During the season, however, one of the top performers, H6S170P5, showed high levels of foliar spot disease and significant susceptibility to drought; for those reasons, this clone was excluded for further testing.

Final years of selection and field trials for cultivar release

In the next two seasons, which included the years 2015 to 2017, evaluations came down to four clones: H3S59P1, H6S163P1, H6S170P5, and H7S194P2. This 2-year assessment corroborated that tuber yields were excellent for these selections as they were comparable or better than some of the common cultivars (Figure 5). The average tuber yield per plant in all the selections combined was 601 g, and two clones, H6S170P5 and H6S163P1, exhibited the highest yields in the group with 897 and 849 g of tuber weight per plant, respectively. Therefore, it was decided to consider these two clones (H6S170P5 and H6S163P1) for potential release as new cultivars.



FIGURE 6

Pictures of the new potato cultivars with enhanced cold hardiness released in Peru: *Wiñay* H6S170P5 (on the left) and *Llapanchispaq* H6S163P1 (on the right).

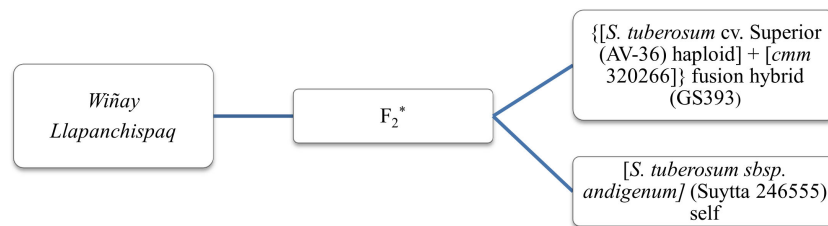


FIGURE 7

Pedigree of *Wiñay* (H6S170P5) and *Llapanchispaq* (H6S163P1), two new native potato varieties for Peru bred from a cross between female Peruvian native potato cultivar "Suytta" and male "cmm + tbr" fusion hybrid GS393. *The F₁ progenies from crosses were subsequently bulk intermated to generate the F₂ populations that were sent to Peru.

Release of new potato cultivars for Peru

The field trials and surveys at the farmers' communities to comply with Peru's regulations for cultivar release, as indicated in INIA's policies and guidelines, confirmed that clones H6S170P5 and H6S163P1 were acceptable in quality and productivity. Though glycoalkaloid levels in tubers were not measured, taste panels done among the farmers at the communities confirmed no perception of bitterness in either selection. As noted in the breeding scheme, these selections derived from non-bitter cultivars and hybrid lines (which were selected for reduced quantities of alkaloids). In addition, most of the wild *cmm* genome in the *cmm-tbr* hybrid was eliminated through backcrosses.

The selections were then collectively accepted by the rural communities and, hence, approved as new cultivars for Peru. In November of 2018, clone H6S170P5 completed the required paperwork and was released as INIA 330 with the cultivar name of *Wiñay* (a Quechua word meaning *to grow*). The first two letters of the cultivar name *Wi-* were also chosen to remember Wisconsin as the origin of the breeding line and the international collaboration. In November 2022, clone H6S163P1 also completed the requirements and was released as INIA 334 with the name of *Llapanchispaq* (a Quechua word meaning *for all of us*) (Figure 6). The pedigree of these two new cultivars is shown in Figure 7.

Discussion

An inspection of the potato seedlings after they were transplanted in the fields showed that the plants exhibited symptoms of environmental stress (as plants moved from steady conditions in the greenhouses to the natural fields in the highlands). Stresses included reduced water availability combined with lower humidity levels in the air and the colder temperatures common to high elevations of Andean environments. In fact, regardless of the time of the year, high elevation sites are characterized by unexpected cold spells at night and early morning. The farmers also indicated that the plants in two locations, *San Jose de Aymara* and *Aramachay*, endured periods of drought early in the season, which caused additional losses. The effects of drought can be very detrimental; more than 80% of the agriculture in the highlands of

Peru is rainfed, and irrigation is not an alternative (Cruz et al., 2017). On the other hand, some fields experienced waterlogging stress after heavy rainfalls, especially in Cusco. All these situations illustrate that strong selective pressure, not just by frost, takes place in the Andes, which provides a glimpse of how challenging and unpredictable agriculture in the region can be.

During the growing seasons, frost episodes occurred at each location at different times (the potato growing season in the highlands of Central and Southern Peru runs from planting in October–November to harvest in May–June). The farmers living in the communities near the fields indicated that each frost they experienced was different in terms of its duration, temperature, and severity. For instance, a frost reported in *Ccorao* (one of the most frost-prone localities identified in the Cusco region) generated very low temperatures with the lowest point in the vicinity of -8°C . That variability and unpredictability of the conditions of each frost event underlines the fact that advances in research and breeding for frost tolerance are very challenging. Each frost occurrence has unique characteristics, so developing cold hardiness that can be expressed in a broader range of conditions could be the most advantageous.

Plants from the distinct breeding families showed different degrees of adaptation and fitness in the fields (in terms of the ability of the plant to grow and develop). As anticipated, the plants resembled the female parent, that is, they presented the morphological plant characteristics of native Peruvian potatoes of the cultivated species *adg*. This was sought in this project and was an important aspect in developing a breeding program there—it was critical to have breeding materials with the cultivar characteristics preferred and wanted by the local farmers. During the selection process, farmers were involved and shared their views, which were fully respected and used to make decisions on what genotypes to select. They mainly stressed the importance of selecting the attractive tuber characteristics favored in the region in terms of tuber shapes and colors. They also indicated their preference for plants with robust foliage.

All assessments done from seedlings to clonal generations were helpful to identify and confirm individuals with exceptional frost tolerance. Additional field trials of the advanced selections confirmed the consistency of frost hardiness at different years and locations. In the later trials, including more plants per each genotype not only helped to estimate yield potential but also to confirm expression of the tuber features, that is, shape, size, quality,



FIGURE 8

Some examples of tuber diversity for shape and color in the breeding lines assessed for this project. The clonal genotypes presented in the pictures are (1) H1S11P2, (2) H7S194P2, (3) H3S59P1, (4) H6S170P1, (5) H3S59P2, (6) H5S130P3, (7) H6S170P5, (8) H5S140P10, (9) H3S63P1, (10) H6S151P2, (11) H6S163P1, and (12) shows a picture of a large display of tubers that were used for selecting tuber features at the *Kayra* Experimental Station in Cusco.

and color (Figure 8). Tuber quality implied tubers with nice, healthy skin, resistant to bruising, and without significant external and internal defects like scab, browning, or hollow heart. In addition, the plants in the field were visually checked for signs of susceptibility to pests/diseases during the season, in particular, late blight (*Phytophthora infestans*) and the Andean potato weevil (*Premnotrypes* spp.), two of the most widespread biotic constraints affecting potato in the Andes. Altogether, we concluded that even though all the initial selections showed good levels of frost tolerance, not all of them were considered for more testing. Most of them (~84%) were found to not have enough yield and/or tuber quality to be included in further assessments. However, the few selections that reached the final stages of this breeding project were exceptional and outperformed many of the traditional cultivars included as controls. This finding is even more remarkable if we consider that the plants were exposed to the colder and tougher environmental conditions of the Altiplano, in the fields of *Salcedo* and *Illpa*. Under those circumstances, plant damage after frost was under 10%, which was an indication that genes for enhanced cold hardiness and acclimation capacity from *cmm* were likely transferred and expressed in the selections.

Another interesting result was that the two clones selected for cultivar release originated from the same breeding family coded as H6-. In this family, the female *adg* parent was a native cultivar

named *Suytta* (which was donated to the USPG from Peru in 1958). The amount of data in the passport record of this collection at the USPG is limited. It reports that this accession was collected in the town of Pisac (near Cusco City), (probably) acquired at a local farmers' market. About its characteristics, the records indicate that



FIGURE 9

Cooperator Ladislao Palomino from the INIA-Cusco showing additional advanced native potato clones generated from the USPG breeding lines sent to Peru.

the tubers are round showing some blue or red and white colors and that plants do not produce many tubers, and yield is poor. The descriptions certainly conflict with what was observed in the field and agronomic evaluations, but we have to keep in mind that these new cultivars are also expressing traits of *tbr* and *cmm*.

In a recent development, plants of *Wiñay* and *Llapanchispaq* growing in Puno in 2022 responded exceptionally well to a prolonged drought affecting the Altiplano region (J. Arcos, personal communication). Drought is also a major abiotic constraint common in the region and responsible for crop losses every year. Finding that the cultivars also have drought tolerance is excellent news and opens opportunities for enhancing food security and sustainability for the farmers in the Altiplano. Also, while this project was in the process of selecting *Wiñay* and *Llapanchispaq*, other additional promising clones shared as breeding stocks with other potato breeding and development efforts in Peru were evaluated (L. Palomino, personal communication). As a result, the potato breeding program at the INIA Cusco has reported several advanced clones that originated from the Wisconsin breeding lines, and that have excellent frost tolerance with potential to become new cultivars (Figure 9).

Conclusions

The development and release of new cultivars *Wiñay* (H6S170P5) and *Llapanchispaq* (H6S163P1) are good examples of implementing a breeding project after effective assessment, use, and deployment of helpful diversity expressed in a wild potato relative. More significantly, this project properly aligns with global efforts aimed at solutions to counter climate change. These cultivars offer farmers opportunities for crop sustainability in regions where frost is serious and extensive. From a broader perspective, this derived from many years of scientific and technical advancement done at the USPG and UW. All led to the discovery of extreme cold hardiness in *cmm*, and later, to figure out that protoplast fusion and somatic hybridization can be used to overcome reproductive barriers in this species, and therefore, access its useful diversity for breeding.

In summary, this effectively integrated past and present accomplishments in research and in multi-institutional efforts in germplasm exchange, conservation, screening, and breeding. The results—two new palatable native cultivars for the Andes and the Altiplano with enhanced frost tolerance that out-yielded current popular varieties. This appears to be the first time that cold hardiness from *cmm* has been incorporated into released native Andean cultivars and that breeding lines created in the USA ended up as new potato cultivars for Peru. This approach proved that global cooperation was possible and successful in achieving the desired important goal of benefit sharing (i.e., *promoting the use of genetic resources for fair and equitable sharing of benefits from their use*) (SCB, 2011; Brink and van Hintum, 2020).

Data availability statement

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

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article. If the image/s was/were reproduced from an existing publication.

Author contributions

JA: Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. AR: Conceptualization, Funding acquisition, Investigation, Supervision, Writing – original draft, Writing – review & editing. JB: Conceptualization, Funding acquisition, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. SV-S: Investigation, Methodology, Validation, Writing – review & editing. JP: Formal Analysis, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. AS: Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. RG: Conceptualization, Investigation, Supervision, Validation, Writing – review & editing. WR: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. DE: Funding acquisition, Investigation, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project was financially supported by the U.S. Potato Genebank, the Plant Stress Physiology Program at the University of Wisconsin-Madison, the College of Agriculture and Life Sciences (CALS) Global Program at the UW-Madison, and Peru's Instituto Nacional de Innovación Agraria (INIA).

Acknowledgments

The authors thank all the farmers from the rural communities in Peru who were very important over the years and in all aspects of this breeding project. We would like to thank Jennifer Kushner and the University of Wisconsin—Madison CALS Global for supporting part of this project. We thank Dr. Celfia Obregon from CITE-Papa in Peru for the administrative and logistic support. We also thank all the help from the US Potato Genebank and staff for their cooperation in the initial steps of developing the breeding families.

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.

References

- Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., et al. (2015). The *Solanum commersonii* Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. *Plant Cell* 27, 954–968. doi: 10.1105/tpc.114.135954
- Brink, M., and van Hintum, T. (2020). Genebank operation in the arena of access and benefit-sharing policies. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01712
- Brush, S. B., Carney, H. J., and Huamán, Z. (1981). Dynamics of Andean potato agriculture. *Econ. Bot.* 35, 70–88. doi: 10.1007/BF02859217
- Cahuana-Quispe, R., and Arcos-Pineda, J. (2002). *Variedades nativas y mejoradas de papa en Puno* (Puno, Perú: INIA), 116.
- Cardi, T., D'Ambrosio, E., Consoli, D., Puite, K. J., and Ramulu, K. S. (1993). Production of somatic hybrids between frost-tolerant *Solanum commersonii* and *S. tuberosum*: characterization of hybrid plants. *Theor. Appl. Genet.* 87, 193–200. doi: 10.1007/BF00223764
- CENEPRED (Peru's National Center for Disaster Risk Estimation, Prevention and Reduction) (2022). *Escenario de riesgo por bajas temperaturas del departamento de Puno* (Lima, Peru: CENEPRED Report 14399).
- Chen, Y. K. H., Bamberg, J. B., and Palta, J. P. (1999c). Expression of freezing tolerance in the interspecific F1 and somatic hybrids of potatoes. *Theor. Appl. Genet.* 98, 995–1004. doi: 10.1007/s001220051160
- Chen, Y. K. H., Palta, J. P., and Bamberg, J. B. (1999b). Freezing tolerance and tuber production in selfed and backcross progenies derived from somatic hybrids between *Solanum tuberosum* L. and *S. commersonii* Dun. *Theor. Appl. Genet.* 99, 100–107. doi: 10.1007/s001220051213
- Chen, Y. K. H., Palta, J. P., Bamberg, J. B., Kim, H., Haberlach, G. T., and Helgeson, J. P. (1999a). Expression of nonacclimated freezing tolerance and cold acclimation capacity in somatic hybrids between hardy wild *Solanum* species and cultivated potatoes. *Euphytica* 107, 1–8. doi: 10.1023/A:1003583706491
- Cho, K. S., Cheon, K. S., Hong, S. Y., Cho, J. H., Mekapogu, M., Yu, Y., et al. (2016). Complete chloroplast genome sequences of *Solanum commersonii* and its application to chloroplast genotype in somatic hybrids with *Solanum tuberosum*. *Plant Cell Rep.* 35, 2113–2123. doi: 10.1007/s00299-016-2022-y
- Condori, B., Hijmans, R. J., Ledent, J. F., and Quiroz, R. (2014). Managing potato biodiversity to cope with frost risk in the high andes: A modeling perspective. *PLoS One* 9, e81510. doi: 10.1371/journal.pone.0081510
- Cruz, P., Winkel, T., Ledru, M. P., Bernard, C., Egan, N., Swingedouw, D., et al. (2017). Rain-fed agriculture thrived despite climate degradation in the pre-Hispanic arid Andes. *Sci. Adv.* 3, e1701740. doi: 10.1126/sciadv.1701740
- El Peruano (2000). Issue Number 7247: 185955-185959.
- Gilles, J. L., and Valdivia, C. (2009). Local forecast communication in the Altiplano. *Bull. Am. Meteorol. Soc.* 90, 85–91. doi: 10.1175/2008BAMS2183.1
- Hanneman, R. E. (1989). The potato germplasm resource. *Am. Potato J.* 66, 655–667. doi: 10.1007/BF02853985
- Hijmans, R. J. (1999). "Estimating frost risk in potato production on the Altiplano using interpolated climate data," in *International Potato Center, Impact on a Changing World, Program Report 1997–1998* (International Potato Center, Lima), 373–380.
- Hijmans, R. J., Condori, B., Carrillo, R., and Kropff, M. J. (2003b). A quantitative and constraint-specific method to assess the potential impact of new agricultural technology: the case of frost resistant potato for the Altiplano (Peru and Bolivia). *Agric. Syst.* 76, 895–911. doi: 10.1016/S0308-521X(02)00081-1
- Hijmans, R. J., Jacobs, M., Bamberg, J. B., and Spooner, D. M. (2003a). Frost tolerance in wild potatoes: Assessing the predictivity of taxonomic, geographic, and ecological factors. *Euphytica* 130, 47–59. doi: 10.1023/A:1022344327669
- JMP®, Version 15.0.0. SAS Institute Inc., Cary, NC, (2019).
- Johnston, S. A., den Nijs, T. P. M., Peloquin, S. J., and Hanneman, R. E. Jr. (1980). The significance of genic balance to endosperm development in interspecific crosses. *Theoret. Appl. Genet.* 57, 5–9. doi: 10.1007/BF00276002
- Kim-Lee, H. Y., Moon, J. S., Hong, Y. J., Kim, M. S., and Cho, H. M. (2005). Bacterial wilt resistance in the progenies of the fusion hybrids between haploid of potato and *Am. J. Potato Res.* 82, 129–137. doi: 10.1007/BF02853650
- Li, P. H., and Palta, J. P. (1978). "Frost hardening and freezing stress in tuber bearing potato species," in *Plant Frost Hardiness and Freezing Stress: Mechanism and Crop Implications*. Eds. P. H. Li and A. Sakai (Academic Press, New York), 49–71, ISBN: .
- Mattheij, W. M., and Puite, K. J. (1992). Tetraploid potato hybrids through protoplast fusions and analysis of their performance in the field. *Theor. Appl. Genet.* 83, 807–812. doi: 10.1007/BF00226701
- Millam, S., Davie, P. A., Harding, K., and Dale, M. F. (1997). Non-transgenic applications of plant tissue culture in potato. *Annu. Rev. Scott. Crop Res. Inst.* 7, 50–52.
- Millam, S., Payne, L. A., and Mackay, G. R. (1995). The integration of protoplast fusion-derived material into a potato breeding programme — a review of progress and problems. *Euphytica* 85, 451–455. doi: 10.1007/BF00023979
- Morton, J. F. (2007). The impact of climate change on smallholder and subsistence agriculture. *Proc. Natl. Acad. Sci.* 104, 19680–19685. doi: 10.1073/pnas.0701855104
- Palta, J. P. (1991). Mechanisms for obtaining freezing stress resistance in herbaceous plants, p. 219–250. In: H. T. Stalker and J. P. Murphy (eds.). *Plant breeding in the 1990s*. Wallingford, UK: CAB Intl. Press.
- Palta, J. P., and Li, P. H. (1979). Frost hardiness in relation to leaf anatomy and natural distribution of *Solanum* species. *Crop Sci.* 19, 665–671. doi: 10.2135/cropsci1979.0011183X001900050031x
- SCB (Secretariat of the Convention on Biodiversity) (2011). *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity: Text and Annex* (Montreal, Canada: Secretariat of the Convention on Biodiversity). 15pp. doi: 10.25607/OBP-789
- Stone, J. M., Palta, J. P., Bamberg, J. B., Weiss, L. S., and Harbage, J. F. (1993). Inheritance of freezing resistance in tuber-bearing *Solanum* species: Evidence for independent genetic control of nonacclimated freezing tolerance and cold acclimation capacity. *Proc. Natl. Acad. Sci.* 90, 7869–7873. doi: 10.1073/pnas.90.16.7869
- Tiwari, J. K., Buckseth, T., Zinta, R., Bhatia, N., Dalamu, D., Naik, S., et al. (2022). Germplasm, breeding, and genomics in potato improvement of biotic and abiotic stresses tolerance. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.805671
- Vega, S. E., and Bamberg, J. B. (1995). Screening the U.S. potato collection for frost hardiness. *Am. Potato J.* 72, 13–21. doi: 10.1007/BF02874375



OPEN ACCESS

EDITED BY

Tao Zhou,
Xi'an Jiaotong University, China

REVIEWED BY

Yu Feng,
Chinese Academy of Sciences (CAS), China
Neng Wei,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE

Min Chen

✉ chenmin@cnbg.net

RECEIVED 19 January 2024

ACCEPTED 22 February 2024

PUBLISHED 06 March 2024

CITATION

Lu R, Hu K, Sun X and Chen M (2024) Low-coverage whole genome sequencing of diverse *Dioscorea bulbifera* accessions for plastome resource development, polymorphic nuclear SSR identification, and phylogenetic analyses.
Front. Plant Sci. 15:1373297.
doi: 10.3389/fpls.2024.1373297

COPYRIGHT

© 2024 Lu, Hu, Sun 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.

Low-coverage whole genome sequencing of diverse *Dioscorea bulbifera* accessions for plastome resource development, polymorphic nuclear SSR identification, and phylogenetic analyses

Ruisen Lu^{1,2}, Ke Hu^{1,2}, Xiaoqin Sun^{1,2,3} and Min Chen^{1,2*}

¹Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China, ²Jiangsu Key Laboratory for the Research and Utilization of Plant Resources, Nanjing, China, ³Jiangsu Provincial Science and Technology Resources Coordination Platform (Agricultural Germplasm Resources) Germplasm Resources Nursery of Medicinal Plants, Nanjing, China

Dioscorea bulbifera (Dioscoreaceae), a versatile herbaceous climber native to Africa and Asia, holds significant nutritional and medicinal value. Despite extensive characterization and genetic variability analyses of African accessions, studies on the genetic variation of this species in China are limited. To address this gap, we conducted low-coverage whole genome sequencing on *D. bulbifera* accessions from diverse regions across mainland China and Taiwan island. Our initial investigation encompassed comprehensive comparative plastome analyses of these *D. bulbifera* accessions, and developing plastome resources (including plastome-derived repetitive sequences, SSRs, and divergent hotspots). We also explored polymorphic nuclear SSRs and elucidated the intraspecific phylogeny of these accessions. Comparative plastome analyses revealed that *D. bulbifera* plastomes exhibited a conserved quadripartite structure with minimal size variation mainly attributed to intergenic spacer regions, reinforcing prior observations of a high degree of conservation within a species. We identified 46 to 52 dispersed repeats and 151 to 163 plastome-derived SSRs, as well as highlighted eight key divergent hotspots in these *D. bulbifera* accessions. Furthermore, we developed 2731 high-quality candidate polymorphic nuclear SSRs for *D. bulbifera*. Intraspecific phylogenetic analysis revealed three distinct clades, where accessions from Southeast China formed a sister group to those from South China and Taiwan island, and collectively, these two clades formed a sister group to the remaining accessions, indicating potential regional genetic divergence. These findings not only contributed to the understanding of the genetic variation of *D. bulbifera*, but also offered valuable resources for future research, breeding efforts, and utilization of this economically important plant species.

KEYWORDS

Dioscorea bulbifera, plastome resources, polymorphic nSSRs, intraspecific phylogeny, molecular breeding

1 Introduction

Dioscorea bulbifera L., commonly referred to as the air potato, air yam, bitter yam, cheeky yam, potato yam, is a dioecious herbaceous climber belonging to the yam family, Dioscoreaceae (Coursey, 1967; Kundu et al., 2021). This species is native to Africa and Asia, but has widely naturalized and is cultivated across various regions, including Central and South America, Nepal, China, the Americas, the West Indies, Pacific Islands, Southeast Asia, and even parts of Australia (Coursey, 1967; Guan et al., 2017; Kundu et al., 2021; Kuncari, 2022). This species is characterized by a twining stem with a sleek surface and alternately arranged vibrant green leaves in a broadly cordate shape (Burkill, 1960; Ding and Gilbert, 2000). The emergence of purplish-brown bulbils (aerial tubers) from leaf axils is particularly noteworthy, as these serve as the primary organ for asexual propagation of the species (Terauchi et al., 1991; Kundu et al., 2021). Additionally, the plant generates underground tubers that bear a resemblance to petite potatoes (Terauchi et al., 1991; Kundu et al., 2021).

Dioscorea bulbifera is a rich source of primary metabolites, encompassing carbohydrates, starch, sugars, proteins, lipids, vitamins, minerals, and fibers (Abara et al., 2011). Its tubers offer versatile culinary possibilities, being adaptable to roasting and cooking as a vegetable, providing sustenance for tribal communities during food crises (Dutta, 2015; Ojinnaka et al., 2017). Significantly, the presence of essential amino acids, such as threonine and phenylalanine, coupled with significant mineral content, notably iron, enhances its nutritional importance (Ezeocha et al., 2014; Otegbayo et al., 2018). Beyond its nutritional richness, *D. bulbifera* holds a profound place in traditional medicine serving as a purgative, anthelmintic, diuretic, rejuvenating tonic, and exhibiting aphrodisiac qualities (Kumar et al., 2017). In traditional Chinese medicine, *D. bulbifera* is employed to address conditions such as cough, pharyngitis, skin infections, piles, hemoptysis, and goiter (Guan et al., 2017). Recent studies have highlighted the potency of *D. bulbifera* against cancer, demonstrating its efficacy in inhibiting tumor growth in various cells, including colon and liver cancer (Guan et al., 2017).

The vast nutritional and medicinal benefits of *Dioscorea bulbifera* have triggered substantial characterization and genetic variability analyses in diverse regions like Brazil (Silva et al., 2016), Ethiopia (Beyene, 2013; Mulualem and Weldemichel, 2013), Nigeria (Jayeola and Oyebola, 2013), Uganda (Ikiriza et al., 2023), and West Africa (Osugwu and Edem, 2020). These investigations have significantly propelled the advancement of breeding techniques aimed at enhancing its desirable traits for both food and medicinal purposes in these areas (Osugwu and Edem, 2020). Nevertheless, despite the abundance of resources in China (Guan et al., 2017), the characterization and genetic variation analyses of this species lag far behind. It remains underutilized, marginalized, and less cultivated in China. Consequently, there exists an urgent necessity for comprehensive characterization, particularly at the molecular level, and genetic variation analysis of this plant in China.

Nowadays, the application of low-coverage whole genome sequencing has emerged as a cost-effective and efficient strategy for selectively capturing high-copy elements such as the plastome, ribosomal DNA, and SSRs across diverse plant species (Straub et al., 2012;

Dodsworth, 2015; Lu et al., 2022). Among these molecular markers, whole plastome sequences, have demonstrated immense value in plant phylogenetic studies owing to their distinctive traits, including the absence of recombination, small effective population sizes, low nucleotide substitution rates, and typically uniparental inheritance (Birky et al., 1983). Additionally, conducting comparative plastomes analyses can facilitate the identification of regions with sequence variation, thereby aiding in accurate and rapid species discrimination—a critical element for the optimal utilization and conservation of plant species (Lu et al., 2021). More significantly, utilizing assembled nuclear sequences derived from low-coverage whole genome sequencing data has successfully enabled the extensive exploration of polymorphic nuclear SSRs (nSSRs) in non-model plant species (Liu et al., 2018; Liu et al., 2021), which play pivotal roles in population genetic analyses and marker-assisted selection (Kumar et al., 2015).

In this study, we conducted low-coverage whole genome sequencing of *Dioscorea bulbifera* accessions spanning diverse regions across mainland China and Taiwan island. Our objectives were to: (1) explore and compare *D. bulbifera* plastomes to unravel their evolutionary patterns; (2) pinpoint plastome-derived markers including repetitive sequences, plastomic SSRs, and divergent hotspots; (3) develop polymorphic nuclear SSRs using multiple assembled nuclear sequences; and (4) reconstruct the phylogenetic relationships among *D. bulbifera* accessions based on plastome data. These discoveries will not only broaden our understanding of the genetic variations within *D. bulbifera* accessions but also furnish essential genetic resources pivotal for advancing molecular characterization and commercial breeding schemes of this species.

2 Materials and methods

2.1 Plant materials, DNA extraction and genome sequencing

We gathered 10 *Dioscorea bulbifera* accessions from various regions across mainland China, spanning Anhui (AHLA), Fujian (FJZZ), Guangdong (GDSG), Guangxi (GXQZ), Henan (HeNXY), Hunan (HuNXX), Jiangsu (JSYX), Sichuan (SCEM), Zhejiang (ZJLS) provinces, along with Taiwan island (TWXB) (Table 1), for comprehensive analysis. For each accession, the pristine, fresh green leaves were harvested from a wild mature individual, and then preserved by desiccation with silica gel. Voucher specimens were deposited at the Herbarium of Institute of Botany, Jiangsu Province and the Chinese Academy of Sciences (NAS). Genomic DNA was then extracted from ~50 mg of silica-dried leaves using the DNAsecure Plant Kit (Tiangen Biotech, Beijing, China), with the purity and integrity of the isolated genomic DNA evaluated through agarose gel electrophoresis and spectrophotometric analysis.

Approximately 1 µg genomic DNA was broken into small fragments using the Covaris E210 sonicator (Covaris Inc., MA, USA). Fragments were then size-selected selected by Agencourt AMPure XP-Medium kit (Thermo Fisher Scientific, USA) to attain sizes ranging from 200 to 400 bp. Following end repair, 3'adenylation, adaptor ligation, PCR amplification, and purification, the resulting double-stranded PCR products were

TABLE 1 Summary of plastome characteristics of 10 *Dioscorea bulbifera* accessions.

Characteristics	AHLA	FJZZ	GDSG	GXQZ	HeNXY	HuNXX	JSYX	SCEM	TWXB	ZJLS
Locality	Lu'an, Anhui	Zhangzhou, Fujian	Shaoguan, Guangdong	Qinzhou, Guangxi	Xinyang, Henan	Xiangxi, Hunan	Yixing, Jiangsu	Emeishan, Sichuan	Xinbei, Taiwan	Lishui, Zhejiang
Clean Reads	45,979,402	51,380,484	37,387,226	40,006,520	43,568,546	39,480,040	49,054,050	45,186,582	53,141,458	41,480,186
Latitude (N)/ Longitude (E)	31°25'57"/ 116°8'31"	24°33'06"/ 117°20'07"	24°55'33"/ 113°01'22"	21°59'06"/ 108° 42'16"	30°59'45"/ 116° 04'49"	29°07'40"/ 110° 27'48"	31°10'05"/ 119° 40'55"	29°35'48"/ 103°22'13"	24°52'03"/ 121° 24'51"	27°54'50"/ 118° 55'23"
Total plastome length (bp)	153,074	153,002	153,099	153,002	152,970	153,093	153,074	153,093	153,002	153,074
LSC	83,225	83,152	83,249	83,152	83,120	83,240	83,225	83,240	83,152	83,225
SSC	18,851	18,852	18,852	18,852	18,852	18,855	18,851	18,855	18,852	18,851
IR	25,499	25,499	25,499	25,499	25,499	25,499	25,499	25,499	25,499	25,499
Total GC content (%)	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0	37.0
LSC	34.8	34.8	34.8	34.8	34.8	34.8	34.8	34.8	34.8	34.8
SSC	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8
IR	43.0	43.0	43.0	43.0	43.0	43.0	43.0	43.0	43.0	43.0
Total number of genes	113	113	113	113	113	113	113	113	113	113
PCGs	79	79	79	79	79	79	79	79	79	79
tRNA genes	30	30	30	30	30	30	30	30	30	30
rRNA genes	4	4	4	4	4	4	4	4	4	4
Duplicated genes	19	19	19	19	19	19	19	19	19	19

transformed into single-stranded circular DNA (ssCir DNA), through heat denaturation and circularization using a splint oligo sequence. The ssCir DNA was formatted as the final library, and sequenced on DNBSEQ-T7 sequencing platform to generate 150-bp paired-end reads.

2.2 Plastome assembly and annotation

After preprocessing raw data with Trimmomatic v.0.36 (Bolger et al., 2014) to remove adaptor sequences, contamination, and low-quality reads, each accession yielded a total of 39,480,040–53,141,458 clean reads utilized for subsequent plastome assembly. *De novo* assembly of the plastome was executed using GetOrganelle v.1.7.6 (Jin et al., 2020), with recommended parameters: -R 15 -k 21,45,65,85,105 -F embplant_pt. Plastome sequences of all 10 accessions were assembled in complete circular sequences. Initial annotation was conducted with the MAFFT v.7 plugin (Katoh and Standley, 2013) integrated within Geneious Prime® 2022.0.1 (https://www.geneious.com), by aligning them with the previously published ones of *Dioscorea bulbifera* (MG805604) and *D. nipponica* (OQ525997) as references. Reference annotations were then transferred to these newly assembled plastomes, followed by meticulous manual verification to ensure precision of exon/intron boundaries and accurate start/stop codon placement. To present comprehensive insights of *D. bulbifera* plastomes, high-resolution

circular plastome maps were generated using the web-based tool OrganellarGenomeDRAW (OGDRAW) v.1.3.1 (Greiner et al., 2019).

2.3 Whole plastome sequence comparison

The mVISTA program (http://genome.lbl.gov/vista/mvista/submit.shtml) was employed to assess the structural resemblance of complete plastome sequences among *Dioscorea bulbifera* accessions, with the annotation from the AHLA plastome sequence serving as the reference. Plastome sequences were aligned using the Shuffle-LAGAN mode with default parameters (Brudno et al., 2003), and the resulting alignments were displayed using the VISTA program (Frazer et al., 2004). Additionally, to detect potential expansions or contractions in the inverted repeat (IR) regions within *D. bulbifera* plastomes, a comparison and visualization of the four junctions between the inverted repeat (IR) and the large single copy (LSC)/small single copy (SSC) regions were conducted using IRscope (https://irscope.shinyapps.io/irapp/, Amiryousefi et al, 2008).

2.4 Plastome-derived markers development

Repetitive sequences, comprising forward (direct), reverse, complement, and palindromic repeats within *Dioscorea bulbifera*

plastomes, were identified through the online tool REPuter (Kurtz et al., 2001). The parameters for repetitive sequences identification involved a minimum repeat size of 30 bp, a sequence identity of at least 90%, and a hamming distance of 3. Simple sequence repeats (SSRs) within the 10 *D. bulbifera* plastome sequences were identified using the MISA-web application (Beier et al., 2017). The SSR search criteria specified a minimum of 10 repeat units for mononucleotide SSRs, 5 for dinucleotide SSRs, 4 for trinucleotide SSRs, and 3 for tetra-, penta-, and hexanucleotide SSRs, respectively.

For an in-depth exploration of divergent hotspots within *Dioscorea bulbifera* plastomes, the 10 newly assembled plastome sequences were aligned using MAFFT v.7 (Katoh and Standley, 2013) in Geneious Prime® 2022.0.1. Regions within this alignment, including protein coding areas, intergenic spacers, and introns, displaying a total mutation count above 0 and an aligned length exceeding 200 bp, were systematically extracted from the alignment matrix. Subsequently, the nucleotide diversity (π) of these extracted regions was computed using DnaSP v.6.12.03 (Rozas et al., 2017).

2.5 Polymorphic nuclear SSRs identification

Low-coverage whole genome sequencing data from three geographically distinct *Dioscorea bulbifera* accessions (AHLA, SCEM, and TWXB) were employed to develop polymorphic nuclear SSR markers. Clean reads of these accessions were aligned to the genome sequences of *D. alata* (Bredeson et al., 2022) and *D. zingiberensis* (Li et al., 2022) to exclude mitochondria and chloroplast reads using BWA-MEM v.0.7.17 (Li, 2013). The resulting alignment files were sorted and converted into Binary Alignment/Map (BAM) format using SAMtools v.1.9 (Li et al., 2009). Subsequently, BAM files were transformed into FastQ files with SAMtools bam2fq. The obtained reads were then *de novo* assembled into contigs using CLC Genomics Workbench v.23.0.4 (CLC bio, Aarhus, Denmark) with default settings. Utilizing these assembled nuclear sequences, CandiSSR (Xia et al., 2016) was employed to identify polymorphic nuclear SSRs (nSSRs) within *D. bulbifera*, using the default parameters except for specifying a flanking sequence length of 200 bp.

2.6 Phylogenetic analyses within *Dioscorea bulbifera*

Phylogenetic analyses were performed using two datasets: complete plastome sequences and 79 protein coding genes shared across all 10 *Dioscorea bulbifera* accessions (Table 1), employing *D. nipponica* (OQ525997), *D. elephantipes* (EF380353) and *D. alata* (OP787123) as outgroups, based on Noda et al. (2020). For the complete plastome dataset, both maximum likelihood (ML) and Bayesian inference (BI) analyses were conducted employing two partitioning scenarios: (1) unpartitioned, and (2) partitioned by each gene and intergenic region (265 partitions). In contrast, the protein coding gene dataset underwent four partitioning scenarios: (1) unpartitioned, (2) partitioned by codon position (three partitions), (3) partitioned by each gene (79 partitions), and (4) partitioned by PartitionFinder v.2.1.1 (Lanfear et al., 2017) (19

partitions). Alignments of both complete plastome sequences and protein coding sequences were executed using the MAFFT v.7 plugin (Katoh and Standley, 2013) within Geneious Prime® 2022.0.1. Except for the partition scenario determined through PartitionFinder, the optimal substitution model was obtained using PartitionFinder, while the remaining partition schemes selected GTR + I + G as the optimal substitution based on the Akaike Information Criterion (AIC) computed by jModelTest v.2.1.4 (Darriba et al., 2012). Maximum Likelihood (ML) analyses were carried out using RAxML v.8.2.12 (Stamatakis, 2014) via the CIPRES Science Gateway v.3.3 (<http://www.phylo.org/portal2/>), utilizing 1000 bootstrap replications. Bayesian Inference (BI) analyses were conducted using MrBayes v.3.2.7 (Ronquist et al., 2012), utilizing Markov Chain Monte Carlo (MCMC) runs for 1×10^6 generations with a sampling frequency of 1000 trees. The initial 1000 trees were discarded as 'burn-in', and the remaining trees were employed to generate a majority-rule consensus tree and estimate posterior probabilities (PPs).

3 Results

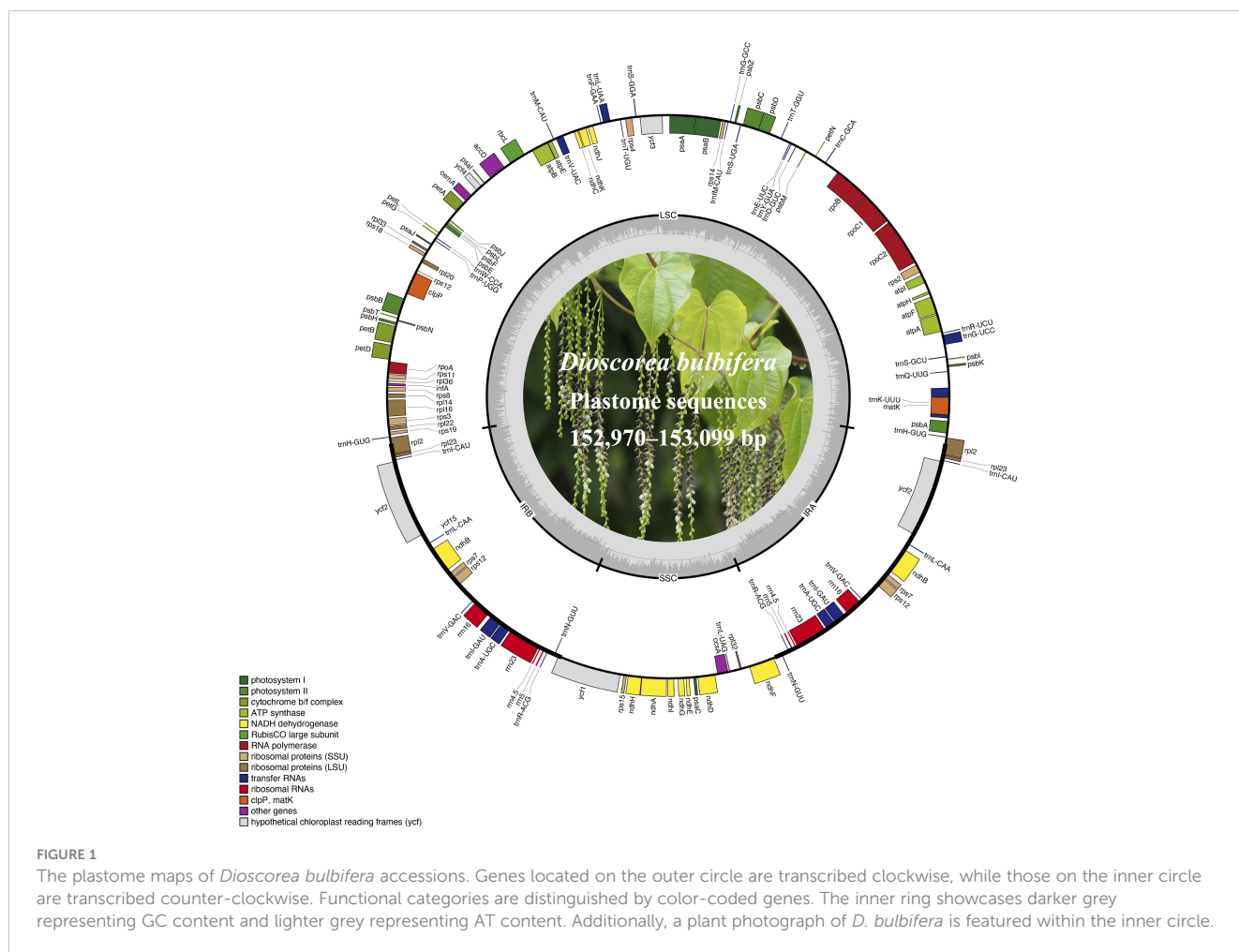
3.1 Plastome characteristics of *Dioscorea bulbifera*

The plastome sizes exhibited narrow variation among these 10 *Dioscorea bulbifera* accessions: GDSG displayed a length of 153,099 bp, HuNXX and SCEM had a size of 153,093 bp, AHLA, JSYX, and ZJLS shared a size of 153,074 bp, FJZZ, GXQZ, and TWXB showcased a size of 153,002 bp, while HeNXY exhibited a size of 152,970 bp (Figure 1; Table 1). Each of these *D. bulbifera* plastomes maintained the typical circular quadripartite structure, comprising a pair of inverted repeat (IR) regions (25,499 bp) separated by a large single copy (LSC) region ranging between 83,120–83,249 bp, and a small single copy (SSC) region varying from 18,851–18,855 bp. The GC content across the entire plastome sequence (37.0%), as well as in the LSC (34.8%), SSC (30.8%), and IR (43.0%) regions, remained consistent among all 10 *D. bulbifera* accessions.

All 10 *Dioscorea bulbifera* plastomes shared an exact set of 113 unique genes, encompassing 79 protein-coding genes (PCGs), 30 transfer RNA (tRNA) genes, and four ribosomal RNA (rRNA) genes (Figure 1; Table 1). Out of these, 19 genes (seven PCGs, eight tRNA genes, and all four rRNA) were duplicated within the inverted repeats (IRs), resulting in a cumulative count of 132 genes (Table 1; Supplementary Table S1). Among these unique genes, eight PCGs and six tRNAs contained a single intron, while three PCGs harbored two introns each (see details in Supplementary Table S1). An intact gene encoding initiation factor IF1 (*infA*) was identified, while the *rps16* gene was independently lost in *D. bulbifera* plastomes (Figure 1).

3.2 Plastome comparison within *Dioscorea bulbifera*

Using the accession AHLA as the reference, mVISTA results indicated high sequence similarity among these *Dioscorea bulbifera*



plastomes, especially in the IR regions (Figure 2). Moreover, the coding regions demonstrated notably higher similarity levels compared to non-coding regions, including introns and intergenic spacers. Notably, the intergenic spacers, specifically *trnK-trnQ* and *psbM-trnD*, displayed the lowest sequence similarity (Figure 2). Within the *trnK-trnQ* region, AHLA, JSYX, HeNXY, HuNXX, SCEM, and ZJLS showed a 24 bp gap compared to FJZZ, GDSC, GXQZ, and TWXB. Similarly, in the *psbM-trnD* region, FJZZ, GXQZ, and TWXB presented a more substantial 96 bp gap compared to the other accessions.

Comparative analysis of IR/SC junctions underscored the stability of *Dioscorea bulbifera* plastomes, revealing no expansion or contraction in the IR regions among these accessions (Figure 3). Across these 10 plastomes, the LSC/IRA junction (JLA) was situated in the *psbA-trnH* intergenic spacer region, 87 bp away from the adjacent gene *psbA*. Concurrently, the LSC/IRb junction (JLB) was positioned within the intergenic spacer of *rps19-trnH*, maintaining an 8 bp distance from the *rps19* gene (Figure 3). Notably, the *ycf1* gene traversed the LSC/IRb junction (JLB), maintaining a consistent length of 365 bp within the IRb region and extending to 5199 bp in the SSC region. Meanwhile, the *ndhF* gene was completely located in the SSC region, merely 4 bp away from the SSC/IRA junction (JSA) (Figure 3).

3.3 Plastome-derived markers of *Dioscorea bulbifera*

The types and lengths of dispersed repeats, including forward (direct), reverse, complement, and palindromic repeats, as well as simple sequence repeats (SSRs) were detected and analyzed within these 10 *Dioscorea bulbifera* plastomes. A total of 250 dispersed repeats were detected across all 10 *D. bulbifera* plastomes, comprising three repeat types: forward (120), reverse (10) and palindromic (120) repeats (Figure 4A). Among all 10 plastomes, FJZZ, GDSC, GXQZ and TWXB displayed the highest count of repeats (total: 26, forward: 13, reverse: 1, and palindromic: 12), followed by AHLA, JSYX and ZJLS (total: 25, forward: 12, reverse: 1, and palindromic: 12), and HuNXX and SCEM (total: 24, forward: 11, reverse: 1, and palindromic: 12), while HeNXY exhibited the fewest (total: 23, forward: 10, reverse: 1, and palindromic: 12) (Figure 4A). Across each *D. bulbifera* plastome, a substantial majority of repeats (61.5% in HeNXY to 76.0% in AHLA, JSYX, and ZJLS) ranged in size between 30 and 40 bp (Figure 4B).

The count of plastome-derived SSRs within the *Dioscorea bulbifera* plastomes varied from 151 (AHLA, JSYX, and ZJLS) to 163 (HuNXX and SCEM) (Figure 5; Supplementary Table S2).

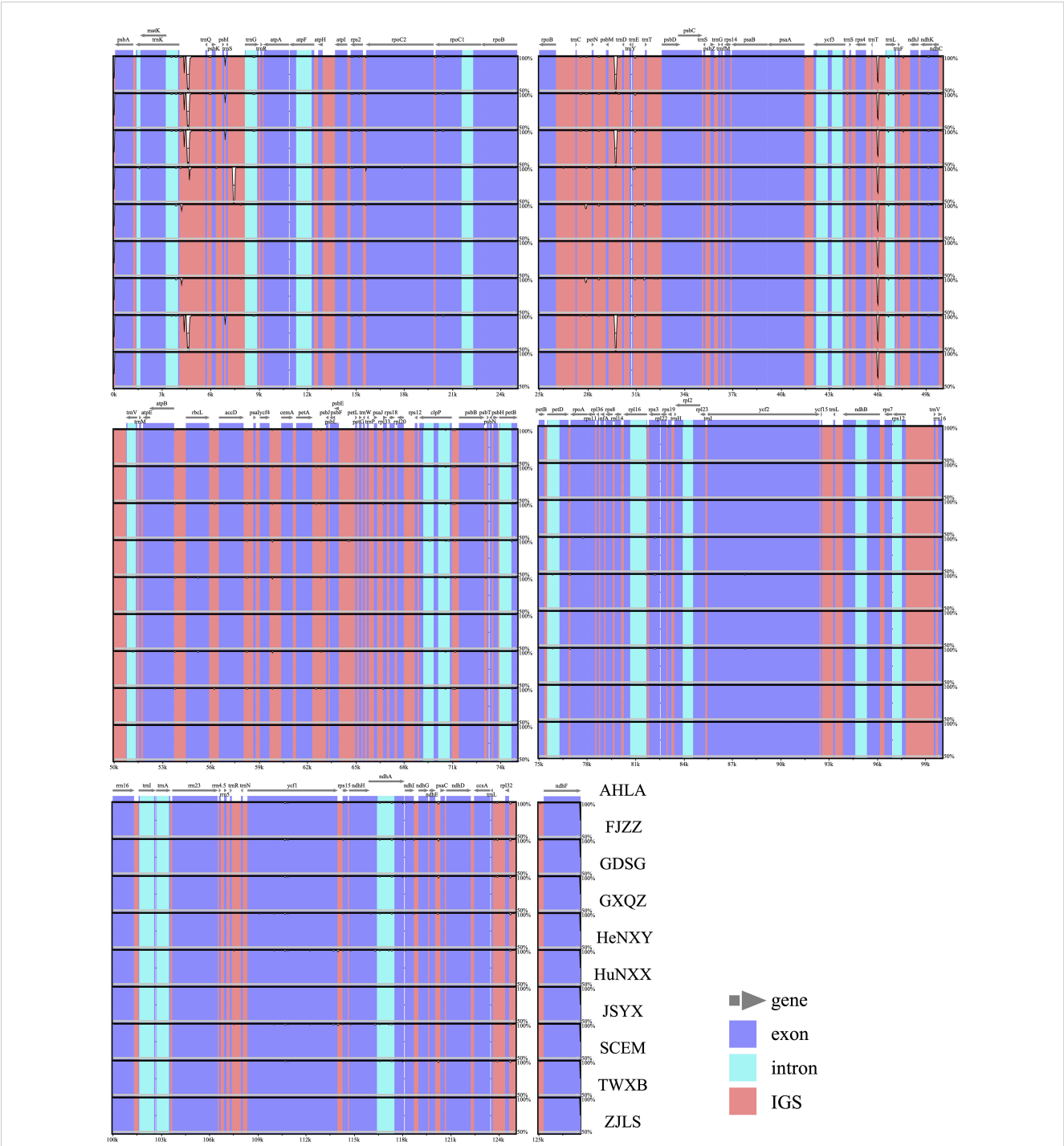
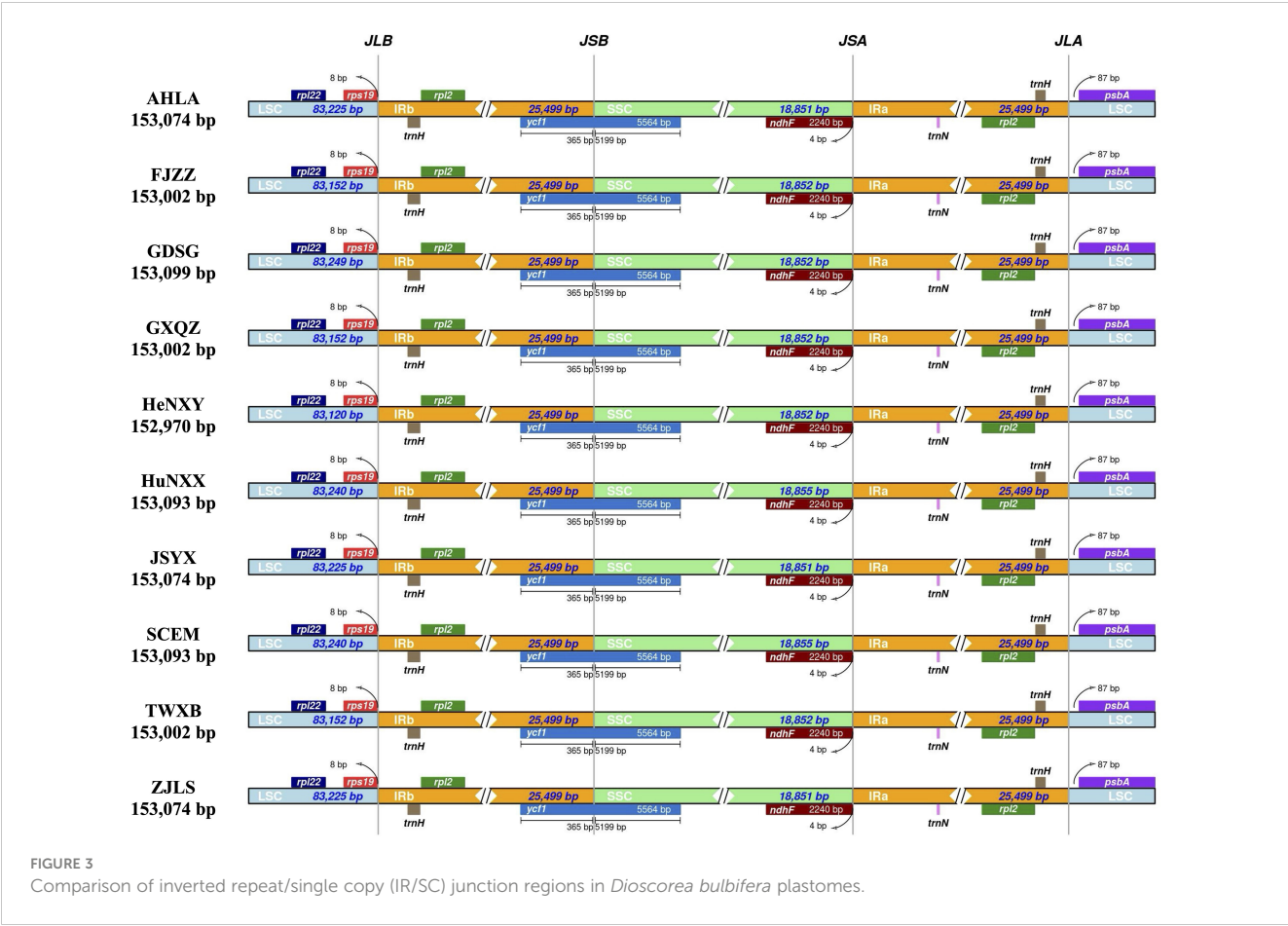


FIGURE 2 Sequence similarity plots among *Dioscorea bulbifera* plastomes, using the accession AHLA as a reference. Annotated genes are shown along the top. The vertical scale indicates percent identity, ranging from 50% to 100%. Genome regions are color-coded, distinguishing between exons, introns, and intergenic spacers (IGS).

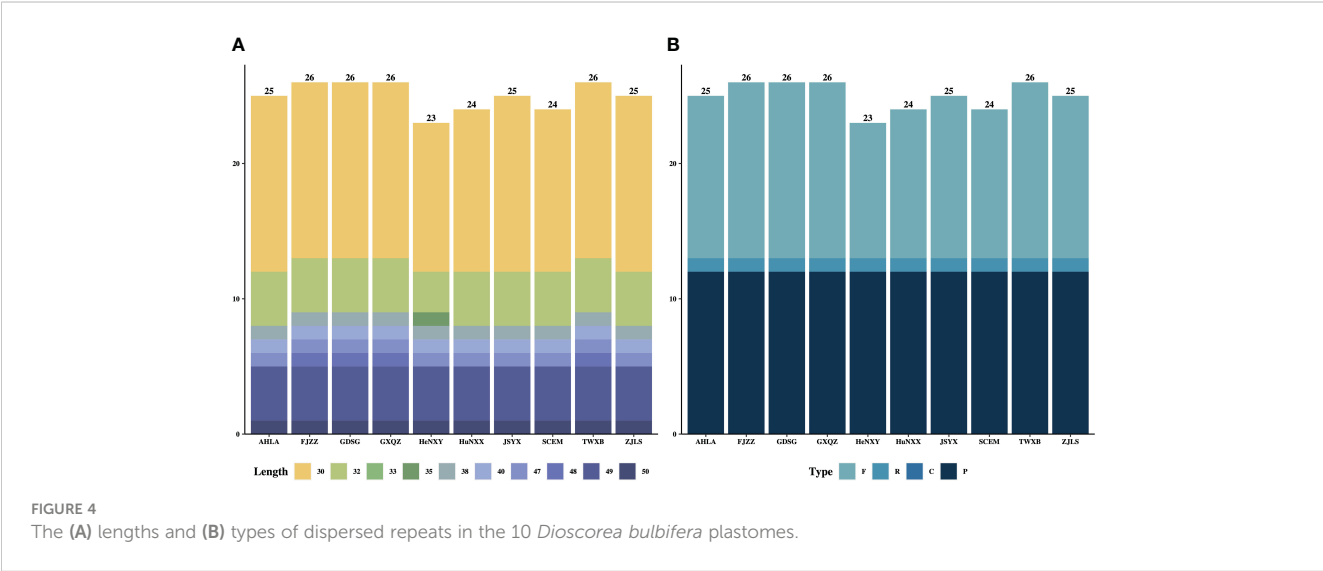
Among these, dinucleotides emerged as the most prevalent SSR type, ranging from 62 in AHLA, JSYX, and ZJLS to 68 in HuNXX and SCEM. Following dinucleotides were mononucleotides, which varied from 46 in AHLA, JSYX, and ZJLS to 52 in HuNXX and SCEM. Trinucleotides and tetranucleotides exhibited a same pattern and were next in frequency, ranging from 17 in FJZZ, GDSC, GXQZ, HeNXY, and TWXB to 19 in HuNXX and SCEM.

Conversely, pentanucleotides were observed in only 4 instances (HeNXY, HuNXX, SCEM) or 5 in the remaining accessions, while hexanucleotides were least common, occurring once in FJZZ, GDSC, GXQZ, HeNXY, HuNXX, SCEM, and TWXB, or twice in the other accessions, (Figure 5; Supplementary Table S2). Among the motifs in the SSRs, A/T and AA/TT were the most frequently occurring motifs, followed by AT/AT, AAA/TTT and AAAA/



TTTT, while the remaining types appeared relatively infrequently (Figure 5; Supplementary Table S2). Moreover, a specific set of at least eight plastome-derived SSRs—(A/T)_{16,17}, (C/G)_{10,11,12}, (AT/AT)_{6,7}, (AT/AT)_{8,9}, (CC/GG)_{5,6}, (CCC/GGG)₄, (ATAT/ATAT)₃, (CGCG/CGCG)₃, (ATATAT/ATATAT)₃ could effectively distinguish these *D. bulbifera* accessions into two to three distinct groups (Supplementary Table S2).

Although a total of 133 regions (62 CDS, 52 IGS, 13 introns and 6 tRNAs) showed an aligned length exceeding 200 bp, only 38 regions (12 CDS, 20 IGS, 3 introns and 3 tRNAs) had a mutation count greater than zero. Consequently, these 38 regions were selected from the alignment of all 10 *Dioscorea bulbifera* plastomes to identify divergent hotspots (Figure 6). These 38 regions displayed π values ranging from 0.000053 (CDS *rpoC2*) to 0.0036 (IGS *ndhE*–





3.4 Polymorphic nuclear SSRs for *Dioscorea bulbifera*

[illegible]

frontiersin.org

above 0.98, indicating their high transferability across *D. bulbifera* accessions. Subsequent filtration, eliminating low-quality PolynSSRs with transferability (similarity) < 95% and a missing rate (MR) \geq 0.5, resulted in a collection of 2433 high-quality candidate PolynSSRs (Supplementary Table S3). Out of these, 2331 high-quality PolynSSRs could be designed for primers, encompassing 95.81% of the total (Supplementary Table S3). Within this set of high-quality candidate PolynSSRs, tetranucleotide repeats comprised the majority at 1041 (42.79%), followed by tri-, penta-, and hexanucleotide repeats, accounting for 32.71%, 12.33%, and 12.17% of the total, respectively (Supplementary Table S3).

3.5 Intraspecific phylogeny of *Dioscorea bulbifera*

Both the ML and BI analyses, employing complete plastome sequences and 79 shared protein coding genes under different partitioning strategies, yielded identical tree topologies. Consequently, only the phylogenetic trees based on complete plastome sequences are presented here (Figure 7). Phylogenetic analyses revealed three main distinct clades within these *Dioscorea bulbifera* accessions. Specifically, AHLA, ZJLS, and JSYX from Southeast China formed a distinctive monophyletic clade (Clade I), that is sister to the Clade II encompassing accessions from South China (FJZZ, GDSG, and GXQZ) and TWXB from Taiwan island. These two clades collectively form a sister group to the remaining accessions (HeNXY, HuNXX, and SCEM) (Clade III) (Figure 7).

4 Discussion

4.1 Plastome characteristics and evolution of *Dioscorea bulbifera*

The comprehensive exploration of 10 *Dioscorea bulbifera* plastomes from diverse geographic regions across mainland China and Taiwan island unveiled intriguing insights into the plastome structure, genetic composition, and variation of this species. Across all 10 accessions, the plastomes maintained a conserved quadripartite structure, with minimal size variation predominantly residing within the single copy regions (Figure 2; Table 1). The consistency in gene content (including unique and duplicated genes), gene order, and GC content of these *D. bulbifera* plastomes reinforced earlier findings that highlighted a high degree of conservation among plastomes within a species in terms of structure, gene composition, and gene order synteny (Muraguri et al., 2020; Lu et al., 2022). The absence of the *rps16* gene in *D. bulbifera* plastomes corroborated our prior observation that this gene may be absent across *Dioscorea* clades, excluding the *Stenophora* clade (= *D. sect. Stenophora*) (Jansen et al., 2007; Hu et al., 2023a). Conversely, the intact presence of the initiation factor IF1 gene (*infA*) in *D. bulbifera* plastomes aligned with its occurrence in other *Dioscorea* species (Lu et al., 2023; Hu et al., 2023a). Notably, among monocots, the depletion of *infA* genes is particularly concentrated (>70%) in Alismatales, Commelinales, Liliales, Pandanales, with minimal loss occurrences of

7.69% observed in Dioscoreales, including 12.50% within Dioscoreaceae (Lu et al., 2021).

Plastome size variations typically arise from two primary factors: i) the dynamic changes in the junctions between the inverted repeat (IR) and single copy (SC) regions (Kim and Lee, 2004; Lu et al., 2016), and ii) the variability of gene spacer regions, and the presence or absence of genes and introns (Jansen et al., 2007; Li et al., 2021). Our investigation of IR/SC junctions across *Dioscorea bulbifera* accessions unveiled a consistent structural configuration without observable expansions or contractions (Figure 3). Considering the conserved nature of gene content and structure, it appeared that differences in plastome sizes among these accessions mainly result from alterations in gene spacer regions. Significantly, the mVISTA analysis underscored that specific intergenic spacers, particularly *trnK-trnQ* and *psbM-trnD*, exhibited low sequence similarity, displaying gaps among the identified clades (Figure 2). This observation, coupled with our previous studies on plastomes of *D. alata* (Lu et al., 2023) and *D. nipponica* (Hu et al., 2023b), suggested that variations in intergenic spacers, particularly *trnK-trnQ*, could be contributing to the diversity in plastome sizes within individual *Dioscorea* species. However, further investigation is warranted to comprehensively understand these variations.

4.2 Molecular markers for *Dioscorea bulbifera*

The recognition of the importance of conserving medicinal plants, enhancing cultivars with desirable traits, and comprehending germplasm diversity has witnessed significant growth in recent years (Baruah et al., 2017; Marakli, 2018). This growing emphasis has led to the utilization of various molecular markers that offer elaborate genomic insights surpassing the capabilities of phenotypic methods (Marakli, 2018). Plastome-derived markers have emerged as valuable assets, enabling the identification of germplasm resources and contributing to their conservation and breeding efforts (Daniell et al., 2016). Despite the high conservation of plastome sequences within *Dioscorea bulbifera*, nucleotide substitutions, SSRs, and indels could serve as valuable markers to elucidate the genetic diversity and guide molecular breeding of this medically important plant (Li et al., 2020). In this study, we successfully identified a remarkable array of plastome-derived SSRs, ranging from 151 (AHLA, JSYX, and ZJLS) to 163 (HuNXX and SCEM) (Figure 5; Supplementary Table S2), and revealed at least eight potentially polymorphic SSRs, highlighting their utility in marker-assisted studies and population genetics.

Previous phylogenetic studies in *Dioscorea* primarily relied on *matK*, *rbcl*, and *trnL-F* genes, which often lacked sufficient phylogenetic resolution within closely related species and within a single species (Hu et al., 2023a). Recent comparative plastome studies have emphasized the concentration of divergent hotspot regions in non-protein-coding areas across *Dioscorea* species. Notable examples included six IGS regions (i.e., *ndhD-ccsA*, *petA-psbI*, *psbZ-trnG*, *rpl32-ndhF*, *trnD-trnY*, and *trnL-rpl32*) and the *rps16* intron sequence emerged as potential molecular markers for species within

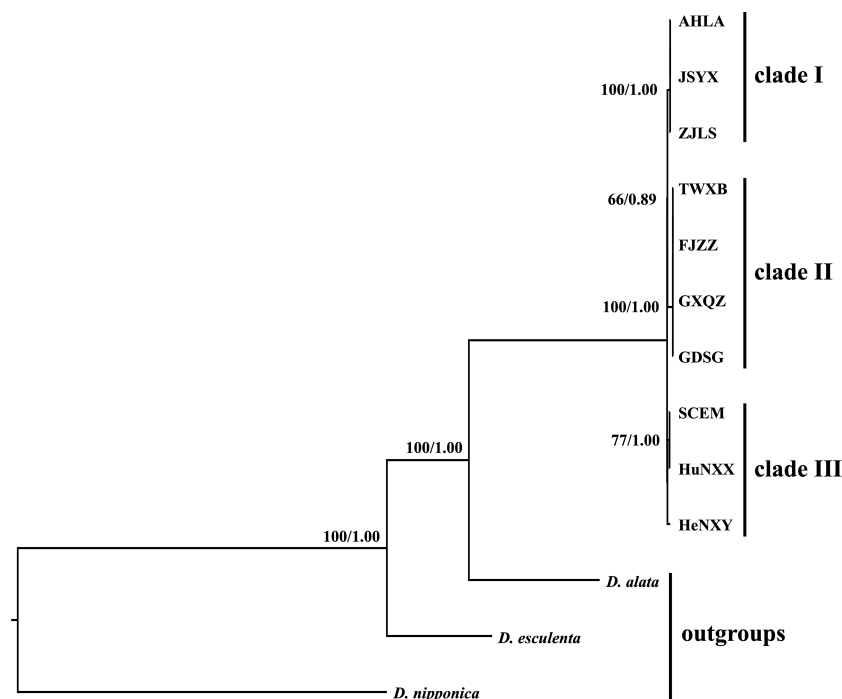


FIGURE 7

Intraspecific phylogenetic relationships among the 10 *Dioscorea bulbifera* accessions originating from different regions across mainland China and Taiwan island, inferred from the methods of maximum likelihood (ML) and Bayesian inference (BI). The ML bootstrap values/BI posterior probabilities are displayed above the lines.

D. sect. Stenophora (Hu et al., 2023a). Similarly, regions such as *ndhD-ccsA*, *trnC-petN*, and *trnL-rpl32* were identified as potential DNA barcodes for species within *D. sect. Enantiophyllum* (Lu et al., 2023). Notably, regions like three intergenic spacers (*rps16-trnQ*, *trnE-trnT*, and *trnL-rpl32*) and two intron regions (intron 1 of *clpP* and intron *trnG*) promised substantial insights into assessing intraspecific genetic variability of *D. nipponica* (Hu et al., 2023b). Given these findings, further exploration of molecular markers, particularly in non-coding regions, becomes imperative in *Dioscorea* species. The comparative analysis across 10 *D. bulbifera* plastomes unveiled four IGS regions (*ndhE-psaC*, *trnF-ndhJ*, *psaJ-rpl33*, and *trnC-petN*), two tRNA regions (*trnL-UAA*, *trnG-UCC*), and two CDS regions (*rps2* and exon 1 of *ndhA*) with notably high values of nucleotide diversity (Figure 6), holding substantial promise for population genetic and intraspecific phylogenetic studies of *D. bulbifera*.

Nuclear SSR markers (nSSRs) have demonstrated their value in diverse applications, including population genetic analyses, cultivar and germplasm identification, and marker-assisted selection, due to their high polymorphism and co-dominant inheritance in a Mendelian fashion (Kaldete et al., 2017). Recent advancements in sequencing technologies and bioinformatic analyses have opened an unprecedented window for identifying high-quality, polymorphic nuclear SSR markers in non-model organisms, offering effective results within optimized cost and time frames (Xia et al., 2016). In this study, a significant discovery of 2433 high-quality candidate PolynSSRs was made (Supplementary Table S3), providing potent tools for conducting population genetic studies of *Dioscorea bulbifera*. In summary, the identified intraspecific plastome-

derived and nuclear markers could offer complementary insights into the genetic structure, differentiation, and gene flow among *D. bulbifera* populations, being crucial for their conservation and efficient management. Furthermore, these markers can be used to develop genetic maps and conduct marker assisted breeding.

4.3 Phylogenetic relationships of *Dioscorea bulbifera*

Nowadays, the utilization of whole plastome sequences has become widespread in elucidating the phylogenetic relationships among plant species (Lu et al., 2016; Hu et al., 2023b). Within the *Dioscorea* genus, the application of whole plastome sequences has significantly clarified previously ambiguous phylogenetic aspects in certain taxa. For instance, Magwé-Tindo et al. (2018) utilized whole plastomes to construct a robust and well-supported phylogenetic tree of West African *Dioscorea* species, revealing six monophyletic groups within them. Additionally, Hu et al. (2023a) conducted phylogenetic analyses for *D. sect. Stenophora* using plastome sequences, suggesting that *D. biformifolia* and *D. banzhuana* represent successive sister species to the remaining *Stenophora* species. Despite these advancements, limited research has explored intraspecific variation and phylogeny of *Dioscorea* species, using whole plastome data. In this study, phylogenetic analyses based on plastome sequences delineated three distinct clades among *D. bulbifera* accessions originating from diverse regions across mainland China and Taiwan island (Figure 7). The identification of three distinct clades implied potential genetic

divergence among populations from different geographic regions. It is noteworthy that accession TWXB unexpectedly clustered with accessions from southern mainland China (FJZZ, GDSG, and GXQZ), displaying no mutations in their plastomes. This finding is surprising given that Taiwan Island became isolated around 10,000 years ago due to rising sea levels (Voris, 2000), leading to disrupted gene flow in many species through the formation of the Taiwan Strait (Lin et al., 2014). One plausible explanation for these results is that glaciations may have caused lowered sea levels, facilitating dispersal between Taiwan and mainland China and thereby obscuring the genetic endemism of Taiwanese accessions (Qu et al., 2015).

Overall, these findings underscored the significance of plastome phylogenomics in resolving intraspecific variation and phylogenetic relationships within *Dioscorea bulbifera*. Moving forward, it is imperative to acquire additional plastomes from *D. bulbifera* accessions in tropical Asia, Northern Australia, America, and sub-Saharan Africa (Kundu et al., 2021). This expansive dataset will provide a comprehensive perspective on the evolutionary relationships and processes of *D. bulbifera*, laying a robust foundation for further exploration of this economically significant species.

5 Conclusions

In conclusion, this study presented a comprehensive analysis of *Dioscorea bulbifera*, a versatile herbaceous climber with substantial nutritional and medicinal importance, through low-coverage whole genome sequencing. The investigation covered diverse accessions from mainland China and Taiwan, shedding light on the genetic variation within this species. Comparative plastome analysis revealed conserved structural features across accessions, with variations mainly attributed to intergenic spacer regions. The identification of plastome-derived markers, including dispersed repeats, SSRs, and divergent hotspots, along with high-quality polymorphic nuclear SSRs, provided valuable tools for population genetic studies and molecular breeding of *D. bulbifera*. The phylogenetic analysis revealed three distinct clades in these *D. bulbifera* accessions, indicating potential genetic divergence among populations from different geographic regions. Overall, this study not only addressed the existing gap in genetic variation studies of *D. bulbifera* in China but also laid the groundwork for further exploration and utilization of this valuable plant species.

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: All newly generated plastome sequences were deposited in GenBank (accession numbers: PP130724–PP130733). The low-coverage whole genome

sequencing data of 10 *Dioscorea bulbifera* accessions generated in this study have been submitted to the NCBI SRA database (<https://www.ncbi.nlm.nih.gov/sra/>), under accession numbers: SRR27556260–SRR27556269.

Author contributions

RL: Software, Writing – original draft. KH: Software, Writing – original draft. XS: Resources, Writing – review & editing. MC: Conceptualization, Funding acquisition, Resources, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (32200194).

Acknowledgments

The authors gratefully acknowledge helpful comments from reviewers on earlier versions of this 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.2024.1373297/full#supplementary-material>

References

- Abara, A., Tawo, E., Obi-Abang, M., and Obochi, G. (2011). Dietary fiber components of four common Nigerian *Dioscorea* species. *Pakistan J. Nutr.* 10, 383–387. doi: 10.3923/pjn.2011.383.387
- Amiryousefi, A., Hyvönen, J., and Pocza, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220
- Baruah, J., Gogoi, B., Das, K., Ahmed, N. M., Sarmah, D. K., Lal, M., et al. (2017). Genetic diversity study amongst *Cymbopogon* species from NE-India using RAPD and ISSR markers. *Ind. Crop Prod.* 95, 235–243. doi: 10.1016/j.indcrop.2016.10.022
- Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198
- Beyene, T. M. (2013). Genetic diversity of aerial yam (*Dioscorea bulbifera* L.) accessions in Ethiopia based on agronomic traits. *Agric. For. Fish* 2, 67–71. doi: 10.11648/j.aff.20130202.12
- Birky, C. W., Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics* 103, 513–527. doi: 10.1093/genetics/103.3.513
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Bredeson, J. V., Lyons, J. B., Oniyinde, I. O., Okereke, N. R., Kolade, O., Nnabue, I., et al. (2022). Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat. Commun.* 13, 2001. doi: 10.1038/s41467-022-29114-w
- Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., et al. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721–731. doi: 10.1101/gr.926603
- Burkhill, I. H. (1960). The organography and evolution of Dioscoreaceae, the family of yams. *Bot. J. Linn. Soc.* 56, 319–412. doi: 10.1111/j.1095-8339.1960.tb02508.x
- Coursey, D. G. (1967). *Yams: an account of the nature, origins, cultivation and utilization of the useful members of the Dioscoreaceae* (London, UK: Longmans, Green and Co), 230.
- Daniell, H., Lin, C. S., Yu, M., and Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 1–29. doi: 10.1186/s13059-016-1004-2
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109
- Ding, Z. Z., and Gilbert, M. G. (2000). “Dioscoreaceae,” in *Flora of China*, vol. 24. Eds. Z. Y. Wu and P. H. Raven (Science Press & St. Louis: Missouri Botanical Garden Press, Beijing), 287.
- Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20, 525–527. doi: 10.1016/j.tplants.2015.06.012
- Dutta, B. (2015). Food and medicinal values of certain species of *Dioscorea* with special reference to Assam. *J. Pharmacognosy Phytochem.* 3, 15–18.
- Ezeocha, V., Nwogha, J., Oluoba, A., and Chukwu, L. (2014). Evaluation of poultry manure application rates on the nutrient composition of *Dioscorea bulbifera* (Aerial yam). *Nigerian Food J.* 32, 92–96. doi: 10.1016/S0189-7241(15)30122-3
- Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, 273–279. doi: 10.1093/nar/gkh458
- Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, 59–64. doi: 10.1093/nar/gkz238
- Guan, X. R., Zhu, L., Xiao, Z. G., Zhang, Y. L., Chen, H. B., and Yi, T. (2017). Bioactivity, toxicity and detoxification assessment of *Dioscorea bulbifera* L.: a comprehensive review. *Phytochem. Rev.* 16, 573–601. doi: 10.1007/s11101-017-9505-5
- Hu, K., Chen, M., Li, P., Sun, X. Q., and Lu, R. S. (2023b). Intraspecific phylogeny and genomic resources development for an important medicinal plant *Dioscorea nipponica*, based on low-coverage whole genome sequencing data. *Front. Plant Sci.* 14, 1320473. doi: 10.3389/fpls.2023.1320473
- Hu, K., Sun, X. Q., Chen, M., and Lu, R. S. (2023a). Low-coverage whole genome sequencing of eleven species/subspecies in *Dioscorea* sect. *Stenophora* (Dioscoreaceae): comparative plastome analyses, molecular markers development and phylogenetic inference. *Front. Plant Sci.* 14, 1196176. doi: 10.3389/fpls.2023.1196176
- Ikiriza, H., Okella, H., Tuyiringiye, N., Milton, A., Catherine, N., Wangalwa, R., et al. (2023). Diversity of *Dioscorea bulbifera* Linn in Uganda assessed by morphological markers and genotyping-by-sequencing technology (GBS). *J. Plant Breed. Crop Sci.* 15, 74–85. doi: 10.5897/JPBSCS2023.1013
- Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *P. Natl. Acad. Sci. U.S.A.* 104, 19369–19374. doi: 10.1073/pnas.0709121104
- Jayeola, A., and Oyebola, T. (2013). Morpho-molecular studies in the natural populations of *Dioscorea bulbifera* Linn. in Nigeria. *J. Exp. Mol. Biol.* 14, 19.
- Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 1–31. doi: 10.1186/s13059-020-02154-5
- Kaldate, R., Rana, M., Sharma, V., Hirakawa, H., Kumar, R., Singh, G., et al. (2017). Development of genome-wide SSR markers in horsegram and their use for genetic diversity and cross-transferability analysis. *Mol. Breed.* 37, 1–10. doi: 10.1007/s11032-017-0701-1
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kim, K. J., and Lee, H. L. (2004). Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* 11, 247–261. doi: 10.1093/dnares/11.4.247
- Kumar, M., Choi, J. Y., Kumari, N., Pareek, A., and Kim, S. R. (2015). Molecular breeding in *Brassica* for salt tolerance: importance of microsatellite (SSR) markers for molecular breeding in *Brassica*. *Front. Plant Sci.* 6, 688. doi: 10.3389/fpls.2015.00688
- Kumar, S., Das, G., Shin, H. S., and Patra, J. K. (2017). *Dioscorea* spp. (a wild edible tuber): a study on its ethnopharmacological potential and traditional use by the local people of Similipal Biosphere Reserve, India. *Front. Pharmacol.* 8, 52. doi: 10.3389/fphar.2017.00052
- Kuncari, E. S. (2022). Nutrition value and phytochemical screening of gembolo (*Dioscorea bulbifera* L.) bulbils and tubers from Bogor, West Java. *Jurnal Ilmu Pertanian Indonesia* 28, 18–25. doi: 10.18343/jipi.28.1.18
- Kundu, B. B., Vanni, K., Farheen, A., Jha, P., Pandey, D. K., and Kumar, V. (2021). *Dioscorea bulbifera* L. (Dioscoreaceae): a review of its ethnobotany, pharmacology and conservation needs. *S. Afr. J. Bot.* 140, 365–374. doi: 10.1016/j.sajb.2020.07.028
- Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633
- Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260
- Li, B., Lin, F., Huang, P., Guo, W., and Zheng, Y. (2020). Development of nuclear SSR and chloroplast genome markers in diverse *Liriodendron chinense* germplasm based on low-coverage whole genome sequencing. *Biol. Res.* 53, 1–12. doi: 10.1186/s40659-020-00289-0
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv 1303.3997*.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, X., Zhao, Y., Tu, X., Li, C., Zhu, Y., Zhong, H., et al. (2021). Comparative analysis of plastomes in Oxalidaceae: phylogenetic relationships and potential molecular markers. *Plant Diversity* 43, 281–291. doi: 10.1016/j.pld.2021.04.004
- Li, Y., Tan, C., Li, Z., Guo, J., Li, S., Chen, X., et al. (2022). The genome of *Dioscorea zingiberensis* sheds light on the biosynthesis, origin and evolution of the medicinally important diosgenin saponins. *Hortic. Res.* 9, uhac165. doi: 10.1093/hr/uhac165
- Lin, A. Q., Csorba, G., Li, L. F., Jiang, T. L., Lu, G. J., Thong, V., et al. (2014). Phylogeography of *Hipposideros armiger* (Chiroptera: Hipposideridae) in the Oriental Region: the contribution of multiple Pleistocene glacial refugia and intrinsic factors to contemporary population genetic structure. *J. Biogeogr.* 41, 317–327. doi: 10.1111/jbi.12163
- Liu, L., Wang, Y., He, P., Li, P., Lee, J., Soltis, D. E., et al. (2018). Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oreotrophe* and *Mukdenia* (Saxifragaceae), using genome skimming data. *BMC Genomics* 19, 235. doi: 10.1186/s12864-018-4633-x
- Liu, L., Zhang, Y., and Li, P. (2021). Development of genomic resources for the genus *Celtis* (Cannabaceae) based on genome skimming data. *Plant Diversity* 43, 43–53. doi: 10.1016/j.pld.2020.09.005
- Lu, R. S., Chen, M., Feng, Y., Yuan, N., Zhang, Y. M., Cao, M., et al. (2022). Comparative plastome analyses and genomic resource development in wild rice (*Zizania* spp., Poaceae) using genome skimming data. *Ind. Crop Prod.* 186, 115224. doi: 10.1016/j.indcrop.2022.115224
- Lu, R. S., Hu, K., Zhang, F. J., Sun, X. Q., Chen, M., and Zhang, Y. M. (2023). Pan-plastome of greater yam (*Dioscorea alata*) in China: intraspecific genetic variation, comparative genomics, and phylogenetic analyses. *Int. J. Mol. Sci.* 24, 3341. doi: 10.3390/ijms24043341
- Lu, R. S., Li, P., and Qiu, Y. X. (2016). The complete chloroplast genomes of three *Cardiocrinum* (Liliaceae) species: comparative genomic and phylogenetic analyses. *Front. Plant Sci.* 7, 2054. doi: 10.3389/fpls.2016.02054

- Lu, R. S., Yang, T., Chen, Y., Wang, S. Y., Cai, M. Q., Cameron, K. M., et al. (2021). Comparative plastome genomics and phylogenetic analyses of Liliaceae. *Bot. J. Linn. Soc.* 196, 279–293. doi: 10.1093/botlinnean/boaa109
- Magwé-Tindo, J., Wieringa, J. J., Sonké, B., Zapfack, L., Vigouroux, Y., Couvreur, T. L., et al. (2018). Guinea yam (*Dioscorea* spp., Dioscoreaceae) wild relatives identified using whole plastome phylogenetic analyses. *Taxon* 67, 905–915. doi: 10.12705/675.4
- Marakli, S. (2018). A brief review of molecular markers to analyze medicinally important plants. *Int. J. Life Sci. Biotechnol.* 1, 29–36. doi: 10.38001/ijlsb.438133
- Mulualem, T., and Weldemichel, G. (2013). Agronomical evaluation of aerial yam (*Dioscorea bulbifera*) accessions collected from South and Southwest Ethiopia. *Greener J. Agric. Sci.* 3, 693–704. doi: 10.15580/GJAS.2013.3.073113767
- Muraguri, S., Xu, W., Chapman, M., Muchugi, A., Oluwaniyi, A., Oyeibanji, O., et al. (2020). Intraspecific variation within Castor bean (*Ricinus communis* L.) based on chloroplast genomes. *Ind. Crop Prod.* 155, 112779. doi: 10.1016/j.indcrop.2020.112779
- Noda, H., Yamashita, J., Fuse, S., Pooma, R., Poopath, M., Tobe, H., et al. (2020). A large-scale phylogenetic analysis of *Dioscorea* (Dioscoreaceae), with reference to character evolution and subgeneric recognition. *Acta Phytotax. Geobot.* 71, 103–128. doi: 10.18942/apg.201923
- Ojinnaka, M., Okudu, H., and Uzosike, F. (2017). Nutrient composition and functional properties of major cultivars of aerial yam (*Dioscorea bulbifera*) in Nigeria. *Food Sci. Qual. Manage.* 62, 1–2.
- Osuagwu, A., and Edem, U. (2020). Evaluation of genetic diversity in Aerial Yam (*Dioscorea bulbifera* L.) using simple sequence repeats (SSR) markers. *Agrotechnology* 9, 202. doi: 10.35248/2168-9881.20.9.202
- Otegbayo, B., Oguniyan, D., Olunlade, B., Oroniran, O., and Atobatele, O. (2018). Characterizing genotypic variation in biochemical composition, anti-nutritional and mineral bioavailability of some Nigerian yam (*Dioscorea* spp.) land races. *J. F. Sci. Tech.* 55, 205–216. doi: 10.1007/s13197-017-2913-0
- Qu, Y., Song, G., Gao, B., Quan, Q., Ericson, P. G., and Lei, F. (2015). The influence of geological events on the endemism of East Asian birds studied through comparative phylogeography. *J. Biogeogr.* 42, 179–192. doi: 10.1111/jbi.12407
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029
- Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248
- Silva, D., Siqueira, M., Carrasco, N., Mantello, C., Nascimento, W., and Veasey, E. A. (2016). Genetic diversity among air yam (*Dioscorea bulbifera*) varieties based on single sequence repeat markers. *Genet. Mol. Res.* 15, 1–12. doi: 10.4238/gmr.15027929
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Straub, S. C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364. doi: 10.3732/ajb.1100335
- Terauchi, R., Terachi, T., and Tsunewaki, K. (1991). Intraspecific variation of chloroplast DNA in *Dioscorea bulbifera* L. *Theor. Appl. Genet.* 81, 461–470. doi: 10.1007/BF00219435
- Voris, H. K. (2000). Maps of Pleistocene sea levels in South-east Asia: shorelines, river systems and time durations. *J. Biogeogr.* 27, 1153–1167. doi: 10.1046/j.1365-2699.2000.00489.x
- Xia, E. H., Yao, Q. Y., Zhang, H. B., Jiang, J. J., Zhang, L. P., and Gao, L. Z. (2016). CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. *Front. Plant Sci.* 6, 1171. doi: 10.3389/fpls.2015.01171



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Xiao Ma,
Sichuan Agricultural University, China
Dayun Tao,
Yunnan Academy of Agricultural Sciences,
China

*CORRESPONDENCE

Shiyong Chen
✉ chengshi8827@163.com

RECEIVED 05 December 2023

ACCEPTED 26 February 2024

PUBLISHED 11 March 2024

CITATION

Li J, Li X, Zhang C, Zhou Q and Chen S (2024)
Phylogeographic analysis reveals extensive
genetic variation of native grass
Elymus nutans (Poaceae) on the
Qinghai-Tibetan plateau.
Front. Plant Sci. 15:1349641.
doi: 10.3389/fpls.2024.1349641

COPYRIGHT

© 2024 Li, Li, Zhang, Zhou 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.

Phylogeographic analysis reveals extensive genetic variation of native grass *Elymus nutans* (Poaceae) on the Qinghai-Tibetan plateau

Jin Li¹, Xinda Li², Changbing Zhang³, Qingping Zhou¹
and Shiyong Chen^{1,2*}

¹Sichuan Zoige Alpine Wetland Ecosystem National Observation and Research Station, Southwest Minzu University, Chengdu, China, ²College of Animal and Veterinary Sciences, Southwest Minzu University, Chengdu, China, ³Institute of Grass Plants, Sichuan Academy of Grassland Science, Chengdu, China

Introduction: *Elymus nutans* holds ecological and pastoral significance due to its adaptability and nutritional value, the Qinghai-Tibet Plateau (QTP) is a key hub for its genetic diversity. To conserve and harness its genetic resources in highland ecosystems, a thorough assessment is vital. However, a comprehensive phylogeographic exploration of *E. nutans* is lacking. The objective of this study was to unravel the genetic diversity, adaptation, and phylogenetics of *E. nutans* populations.

Methods: Encompassing 361 individuals across 35 populations, the species' genetic landscape and dynamic responses to diverse environments were decoded by using four chloroplast DNA (cpDNA) sequences and nine microsatellite markers derived from the transcriptome.

Results and discussion: This study unveiled a notable degree of genetic diversity in *E. nutans* populations at nuclear ($I = 0.46$, $H_e = 0.32$) and plastid DNA levels ($H_d = 0.805$, $\pi = 0.67$). Analysis via AMOVA highlighted genetic variation predominantly within populations. Despite limited isolation by distance (IBD), the Mekong-Salween Divide (MSD) emerged as a significant factor influencing genetic differentiation and conserving diversity. Furthermore, correlations were established between external environmental factors and effective alleles of three EST-SSRs (EN5, EN57 and EN80), potentially linked to glutathione S-transferases T1 or hypothetical proteins, affecting adaptation. This study deepens the understanding of the intricate relationship between genetic diversity, adaptation, and environmental factors within *E. nutans* populations on the QTP. The findings shed light on the species' evolutionary responses to diverse ecological conditions and contribute to a broader comprehension of plant adaptation mechanisms.

KEYWORDS

Elymus nutans, Qinghai-Tibet plateau, phylogeography, chloroplast DNA, EST-SSR

1 Introduction

The Qinghai-Tibet Plateau (QTP), known as the “Roof of the World,” holds significant importance in phylogeographic research due to its exceptional biodiversity and ecological significance (Yu et al., 2019). It serves as a center for genetic differentiation, supporting numerous endemic genera and species (Shahzad et al., 2017). The unique biodiversity of the plateau can be attributed to Quaternary climate oscillations and the uplift stages during the Miocene-Pliocene or Miocene-Quaternary periods (Zheng et al., 2021). Furthermore, compared to other regions, the QTP exhibits heightened sensitivity to Quaternary climate change, resulting in notable effects on the geographic and population distribution of plant species and leaving distinct genetic imprints (Du et al., 2020). However, the current rate of global climate change is unprecedented in history, raising concerns about its potential impact on species distribution and diversity (Proft et al., 2021). This risk is particularly alarming for the QTP, which is home to numerous endemic and threatened species highly susceptible to climate variations (Wang et al., 2018; Sun et al., 2021). Ensuring their survival is a matter of utmost concern. In response to rapid climate change, natural populations primarily exhibit adaptation, migration, or extinction (Stojanova et al., 2018). Genetic variation plays a pivotal role in species’ capacity to adapt to dynamic and ever-changing environments, when confronted with climate change, populations with higher genetic diversity possess an enhanced likelihood of survival (Gray et al., 2014). This heightened genetic diversity enables them to better withstand the intense selection pressures imposed by contemporary climate change (Jump et al., 2009).

Previous phylogeographic investigations have effectively compared and assessed the plant populations in the QTP, fostering explorations into the intricate interactions between lineage evolution, environmental heterogeneity, and genetic diversity (Aguilar-Melo et al., 2019; Liu et al., 2018). However, it’s noteworthy that while extensive phylogeographic surveys have been conducted on the flora of the QTP, a considerable portion of these studies has predominantly focused on arboreal and shrub species (Kou et al., 2014), often overlooking herbaceous plants (Xiong et al., 2021). Herbaceous plants, with their exceptional cold tolerance, temperature adaptability, and drought resistance, are widely distributed across the plateau, exhibiting unique adaptability and ecological relevance (Wang and Lu, 2014; Sun et al., 2020). Therefore, understanding the phylogenetic history and genetic status of these Poaceae plants is crucial for a comprehensive understanding of the floral evolution on the Qinghai-Tibet Plateau and for the formulation of effective germplasm conservation strategies (Lei et al., 2022).

The majority of herbaceous plants possess intricate evolutionary histories and large genomes, significantly impeding the application of genome scanning techniques and consequently imposing limitations on phylogeographic research efforts (Zwyrtková et al., 2022). Given the crucial role of grasslands in the QTP, this knowledge gap becomes even more consequential (Liu et al., 2018). In such scenarios, expressed sequence tag simple sequence repeats (EST-SSRs) derived from transcriptome sequencing data have emerged as a reliable alternative (Vu et al.,

2020; Li et al., 2023). Typically located in functional gene regions, EST-SSRs offer a plethora of advantages including high polymorphism, broad applicability to closely related species, repeatability, and cost-efficiency (Decroocq et al., 2003). Moreover, maternal chloroplast markers, characterized by their relatively low mutation rates and stable genetic structures, are deemed ideal tools for deciphering population genetic structures and evolutionary histories (Birky et al., 1983). These markers, due to their maternal inheritance, provide researchers with a means to circumvent complex genome structures and diverse genetic backgrounds, thereby facilitating more precise tracing of lineage relationships and genetic drift (Provan et al., 2001), and proving particularly invaluable in resolving phylogenetic admixture and reconstructing population dispersal paths (Petit et al., 1993). Consequently, in phylogeographic and evolutionary biological studies, the amalgamation of EST-SSRs and maternal chloroplast markers offers valuable genetic information for species with limited genomic resources. This genetic information can be effectively utilized across various research domains. For instance, Setsuko et al. (2020) combined 14 EST-SSRs with 3 cpDNA markers (*trnQ-rps16*, *matK*, and *trnL-trnF*) to investigate the population diversity and structure of the endemic plant *Pandanus boninensis* in the Ogasawara Islands, Japan, gathering evidence of migration between older and younger islands. Guo et al. (2022) utilized 38 EST-SSRs and 3 cpDNA markers (*ndhF-rpl132*, *rps16-trnQ*, and *trnE-trnT*) to test the refugia hypothesis for the subtropical vine plant *Actinidia eriantha* in Eastern China.

The current investigation delineates the geographic structure and population divergence of *Elymus nutans*, a perennial and allohexaploid ($2n = 6x = 42$) entity within the Triticeae (Poaceae). As an extensively distributed herbaceous plant, boasts a widespread presence across the QTP and the Himalayas, thriving at elevations ranging from 3000 to 4500 m (Chen et al., 2009). Owing to its prolific yield, nutritional richness, and robustness against a spectrum of abiotic stresses, this species assumes a critical role in animal husbandry and environmental conservation. These attributes position *E. nutans* as an exemplary subject for initiatives in ecological restoration, development of artificial grasslands, and advancement of agricultural and ecological studies in the QTP. However, the habitat of indigenous plants in the QTP, including *E. nutans*, has been threatened by climate change and overgrazing in recent decades (Pradheep et al., 2019). Therefore, understanding the genetic status of *E. nutans* germplasm resources is critical for the conservation and utilization of genetic resources, and maintaining the ecosystem stability. In the present study, the EST-SSR markers were used to assess the genetic diversity of *E. nutans* populations and obtain inter-population gene flow information, and the sequence variations in four chloroplast DNA regions were also applied to analyze the phylogeographic structuring of *E. nutans* populations. This research endeavors to demystify how the genetic variations identified contribute to the species’ capacity to adapt to the diverse environmental conditions prevalent across the QTP. Furthermore, it seeks to determine whether the relationships established between genetic markers and external environmental influences can provide a deeper understanding of the inherent adaptive mechanisms of the species. The findings of this investigation have the potential to

enhance our understanding of germplasm resources and inform breeding strategies for this species.

2 Methods

2.1 Plant sampling and DNA extraction

In the present study, 361 silica-dried leaf samples from 35 populations of *E. nutans* were collected in the southeastern QTP from August to September 2019. In order to reveal the genetic characteristics of this species in response to environmental change, populations sampling in this study was carried out on a longitude gradient. For each population, we sampled 7 to 15 individuals, ensuring that the sampling locations were at least 5 meters apart. The seeds of these specimens were planted at the experimental field of Sichuan Academy of Grassland Science, Hongyuan, China (32° 46.61'N, 102°32.63'E). [Supplementary Table 1](#) provides detailed information regarding the collection sites, geographical coordinates, and elevations for each of the source populations. Genomic DNA extraction was carried out using the DP350 Plant DNA Kit (Tiangen Biotechnology, Beijing, China), following the manufacturer's protocol. Subsequently, the quality and quantity of the extracted DNA samples were assessed using a NanoDrop-Lite instrument (Thermo Scientific, Waltham, MA, USA) and 1% agarose gel electrophoresis, respectively. Only the DNA samples meeting the required quality criteria were considered for further analysis. Subsequently, the qualified DNA samples were diluted to a concentration of 10 ng/μL for use.

2.2 Chloroplast DNA sequencing and EST-SSR genotyping

Four cpDNA fragments (*trnH-psbA*, *trnL-F*, *matK*, and *rbcl*) were amplified and sequenced across all *E. nutans* samples. These fragments were amplified using primers previously reported in the literature, known for showcasing polymorphisms in other *Elymus* species ([Xiong et al., 2022](#)). Amplification was conducted in a 30 μL reaction volume, comprising of 30 ng of genomic DNA, 1.5 μL of each primer, and 15 μL of 2× Es Taq MasterMix (CoWin Biosciences, Beijing, China), using a C1000 Touch Thermal Cycler (BIO-RAD, Foster City, CA, USA). The polymerase chain reaction (PCR) conditions throughout the experiment were consistent with those described by Shaw et al ([Shaw et al., 2005](#)), and the resulting products were sequenced by Tsingke Biotech (Beijing, China). The genotypes of all 361 samples were determined using nine pairs of EST-SSR primers developed for *E. nutans* in our previous study based on transcriptomic data, and the PCR procedure followed the same protocol as we described previously ([Li et al., 2023](#)). The PCR products were subjected to capillary electrophoresis in an ABI 3730xl DNA analyzer (Applied Biosystems, Foster, CA, USA) and analyzed using GeneMarker v. 2.2 software (SoftGenetics, Pennsylvania, USA).

2.3 Phylogeographic analysis based on cpDNA sequencing

All cpDNA data were edited and adjusted using DNAMAN to obtain consensus sequences. The sequences were then aligned using the ClustalW algorithm in MEGA 6.0 ([Tamura et al., 2013](#)). PhyloSuite v1.2.3 was utilized to concatenate the sequences serially ([Zhang et al., 2020](#)). DnaSP v5.0 was employed to calculate haplotype (H), haplotype diversity (*Hd*), nucleotide diversity (π), and population genetic distance based on the concatenated sequences ([Librado and Rozas, 2009](#)). A haplotype network was constructed using the median-joining method in PopART software ([Leigh and Bryant, 2015](#)). The haplotype phylogenetic tree was constructed by the maximum likelihood (ML) method in MEGA 6.0 software, *Elymus repens* and *Elymus ciliaris* were chosen as outgroups ([Bouckaert et al., 2014](#)). The maximum clade credibility tree and additional summary statistics were visualized using FigTree 1.3.1. This enabled the assessment of potential spatial expansion within the populations, Tajima's D and Fu's *F_s* were calculated using DnaSP version 5.0. Mismatch distribution analysis and neutrality tests were employed ([Tajima, 1989](#); [Fu and Li, 1993](#)). Analysis of molecular variance (AMOVA) was conducted using Arlequin v3.5 to estimate genetic variance within and between populations ([Excoffier and Lischer, 2010](#)).

2.4 Population structure and genetic barriers analysis based on microsatellite data

In the analysis, the raw data matrix generated by EST-SSR markers was examined using the GenAlEx 6.5102 project ([Peakall and Smouse, 2012](#)). This encompassed parameters such as Shannon's information index (*I*), observed number of alleles (*N_a*), and the number of effective alleles (*N_e*). The genetic structure of *E. nutans* populations was investigated via the Bayesian clustering method in Structure v.2.3.4 ([Falush et al., 2007](#)). Determination of the optimal number of clusters (*K*) involved 20 independent runs for each *K* value ranging from 1 to 18. Structure Harvester v0.6.94 ([Earl and Vonholdt, 2012](#)) facilitated this analysis, each run comprising 500,000 Monte Carlo Markov Chain (MCMC) replicates. An admixture model was utilized, incorporating a burn-in period of 10,000 replicates. The ΔK method ([Evanno et al., 2005](#)) was employed to estimate the most likely number of clusters. Subsequently, the 20 replicates underwent clustering and permutation using CLUMPP v1.1 with the LargeK Greedy algorithm ([Jakobsson and Rosenberg, 2007](#)). To explore genetic differentiation at the population level, an analysis of molecular variance (AMOVA) was conducted within the GenAlEx 6.5102 project. Additionally, Barrier v2.2, based on Monmonier's maximum difference algorithm ([Manni et al., 2004](#)), was employed to predict major genetic barriers' geographical locations between populations.

2.5 Genetic differentiation and polymorphism with climatic variables

In investigating the relationships between geographical and environmental factors and population genetic differentiation, population genetic distance indices for SSRs and cpDNA were computed using GenAlEx 6.5102 and DnaSP v5.0, respectively. Pairwise geographical distances were calculated utilizing the “geosphere” package in R, while environmental Euclidean distances were derived from nineteen bioclimatic variables obtained from the WorldClim website (<https://www.worldclim.org/>). Regression analyses were employed to evaluate isolation by distance (IBD) and isolation by environment (IBE) through the plotting of genetic distance against geographical and environmental Euclidean distances, respectively. These analyses facilitated an exploration of potential influences of geographical and environmental factors on population genetic differentiation.

To delve further into potential local adaptation due to environmental fluctuations, the correlation between climate variables and molecular markers was examined. Considering the high ploidy of *E. nutans*, variations in the detected allele numbers among different molecular markers across populations were anticipated. Assessing adaptive trends at specific sites involved establishing a linear regression relationship between the number of effective alleles and climate variables. This analysis offered insights into potential adaptive alterations in response to environmental factors at the identified loci.

3 Results

3.1 Haplotype distribution and phylogenetic relationship

The total length of the four cpDNA sequences was 4689 bp, and the lengths of the *trnH-psbA*, *trnL-trnF*, *matK*, and *rbcL* regions were 667, 1052, 1538, and 1432 bp, respectively. All sequences were deposited in GenBank under accession numbers: OR421574-OR423017. As cpDNA regions are uniparentally inherited markers, we used these four chloroplast fragments in our subsequent population genetics analysis. Among the 361 individuals sampled from 35 populations of *E. nutans*, a total of 20 variable sites were detected in the combined cpDNA. Genetic diversity indicators for various regions of *E. nutans* can be found in Table 1. The DQ4 population exhibited the highest values for number of haplotypes ($h = 6$) and haplotype diversity ($Hd = 0.911$), followed by population DQ3 ($h = 5$, $Hd = 0.833$). The lowest values were observed in populations RQ1, JD2, and SN1, with a single haplotype each ($h = 1$, $Hd = 0$). On the other hand, the GG4 population showed the highest nucleotide diversity ($\pi = 1.74 \times 10^{-3}$), followed by population DX2 ($\pi = 1.06 \times 10^{-3}$), these comprehensive information regarding sample locations, sample sizes, and descriptive statistics of genetic variation for the populations is available in Supplementary Table 1. Analyses of molecular variance (AMOVAs) based on cpDNA sequence data revealed that the variation among populations of *E. nutans* was

TABLE 1 Genetic diversity for four regions of *Elymus nutans*.

Region	cpDNA			Microsatellites			
	h	Hd	$\pi (10^{-3})$	N_a	N_e	I	He
Qamdo	15	0.827	0.63	2.52	2.17	0.44	0.28
Nyingchi	6	0.821	0.34	2.87	2.24	0.51	0.33
Lhasa	7	0.742	0.69	3.26	2.49	0.57	0.35
Nagqu	7	0.615	0.28	2.62	2.18	0.42	0.27
Total	19	0.805	0.67	2.65	2.21	0.46	0.32

equivalent to the variation within populations, accounting for 49.76% and 50.24%, respectively (Table 2).

A total of 19 distinct haplotypes (H1–H19) were identified, and the spatial distribution of each population, along with its corresponding haplotype composition, is illustrated in Figure 1. The differentiation among haplotype categories is visually discernible through color distinctions, as demonstrated in the haplotype network depicted in Figure 2. Both the ML tree and the haplotype network consistently exhibit correspondence in Figure 2. Based on the distribution patterns in the ML tree and haplotype network, the haplotypes can be systematically grouped into two categories. Group1 comprises 13 haplotypes, prominently featuring the most widely distributed and frequently occurring haplotype H4, alongside its subsequent haplotype H5. These two extensively distributed haplotypes hold central positions within the network diagram, suggesting their potential status as ancestral haplotypes. Additionally, Group1 encompasses 7 unique haplotypes, with H13, H14, H16, and H18 exclusive to Qamdo, while H11 exclusively appears in Nagqu. In contrast, Group2 consists of 6 haplotypes, none of which are observed in Nagqu. Notably, Group2 includes 3 haplotypes from Qamdo (H2, H15, H17) and 1 haplotype H8 from Nyingchi. To gain insight into the historical dynamics of *E. nutans*, mismatch distribution analysis revealed a multimodal pattern (Figure 3), and neutrality tests showed that both Fu and Li's F^* (Fu and Li's $F^* = 1.22427$, $p > 0.10$) and Tajima's D value (Tajima's $D = -0.07339$, $p > 0.10$) were not statistically significant. These results do not support population expansions in *E. nutans*.

3.2 Genetic diversity, population structure and genetic barriers analysis

Nine EST-SSR markers used to assess the diversity of *E. nutans* populations exhibited polymorphism (Supplementary Table 2). The observed allele number (N_a) detected by each primer ranged from 1.37 (EN91) to 3.57 (EN62), with an average of 2.65. The effective number of alleles (N_e) per locus ranged from 1.14 (EN91) to 3.30 (EN62), with a mean of 2.21. The average Shannon's information index (I) was 0.46, varying from 0.15 (EN91) to 0.74 (EN5). The observed heterozygosity (H_o) and expected heterozygosity (H_e) varied across primers, with the lowest values observed in EN91 (0.06 and 0.11, respectively) and the highest in EN67 (0.99 and 0.55, respectively). The overall average H_o and H_e were 0.32 and 0.29,

TABLE 2 AMOVA analysis based on the cpDNA and microsatellites for *E. nutans* populations.

Source of variation	df	cpDNA				Microsatellites			
		Sum of squares	Variance component	Total variance	<i>p</i>	Sum of squares	Variance component	Total variance	<i>p</i>
Among populations	34	1934.22	4.89	49.76%	<0.001	985.46	1.31	38%	<0.001
Within populations	326	1610.56	4.94	50.24%	<0.001	1433.16	2.08	62%	<0.001
Total	360	3544.78	9.83	100%		2418.62	3.39	100%	

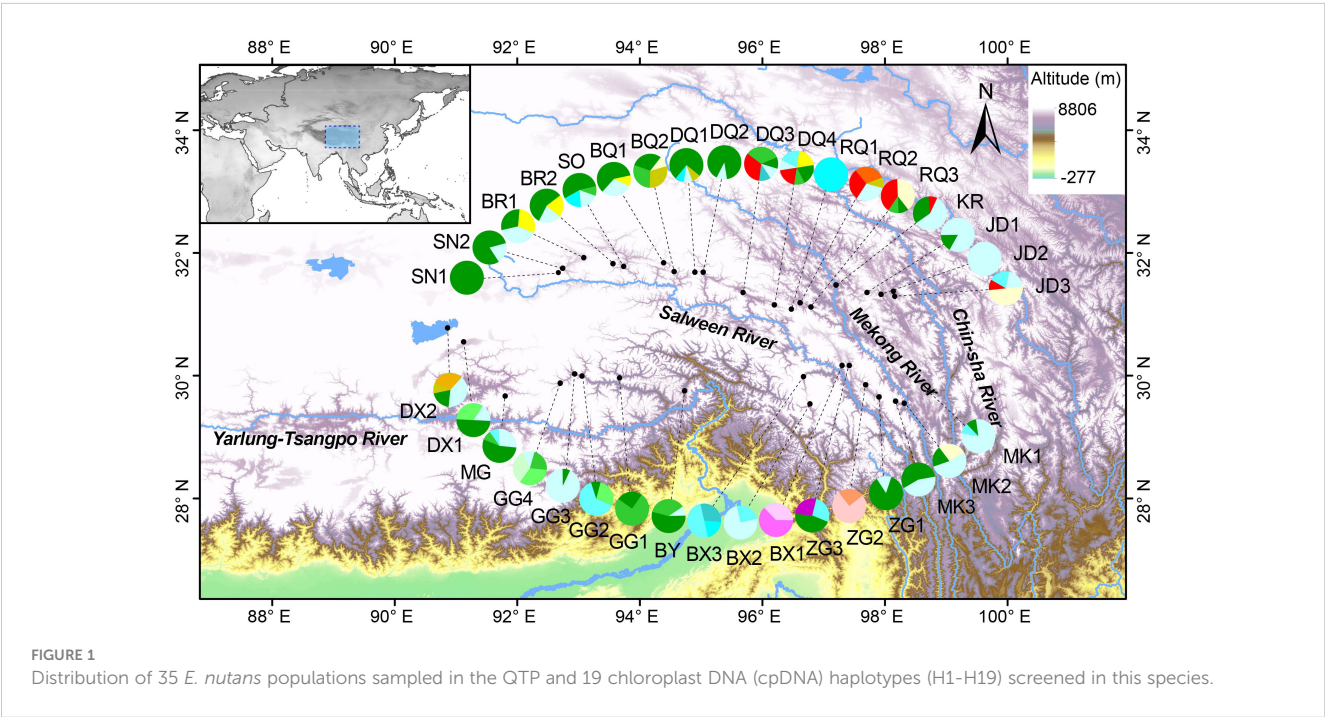
respectively. Table 1 provides an overview of the genetic variability at the population level. The *Na* and *Ne* per population ranged from 1.56 (MK2) to 3.56 (MG) and from 1.55 (MK2) to 2.82 (ZG3), with mean values of 2.65 and 2.21, respectively. Shannon’s information index (*I*) and expected heterozygosity (*He*) detected in each population ranged from MK2 (*I* = 0.12; *He* = 0.08) to ZG3 (*I* = 0.74; *He* = 0.46), with averages of 0.46 and 0.32, respectively.

The software Barrier was employed to identify genetic barriers among the 35 populations, represented by red lines in Figure 4A. The study confirmed that the Salween River serves as a significant genetic barrier in the region. Additionally, crisscrossing genetic barriers were observed near the Mekong River in the eastern part of the study area. The Bayesian cluster analysis (Structure) based on microsatellite data indicated the presence of three optimal clusters (Supplementary Figure S1), suggesting that the 361 individuals likely belong to two main genetic clusters. Using CLUMPP to determine the most optimal of the 20 replicates, a plot of the structure of the 35 populations was constructed (Figure 4B). Despite a considerable degree of hybridization between populations, their genetic backgrounds can still be differentiated through genetic barriers. Furthermore, the AMOVA results demonstrated that a significant proportion (62%)

of the genetic variation exists within the 35 *E. nutans* populations, while 38% of the genetic variation is found among populations (Table 2).

3.3 Effects of ecogeographical factors on genetic divergence and adaptation

Based on the Mantel test using pairwise distance values, our study revealed that geographic distance had a weak effect on the genetic differentiation detected by microsatellites (Figure 5A, *r*=0.181; *p*=0.008), while environmental distance did not show a significant impact (Figure 5B, *r*=0.064; *p*=0.128). However, there was no significant correlation between either geographic distance or environmental distance and the population genetic differentiation detected by cpDNA (Figures 5C, D). Nevertheless, we did observe a significant correlation between geographic distance and environmental distance among sampling points (Figure 5E, *r* = 0.495; *p* = 0.001), as well as a significant correlation between genetic distances among populations detected by cpDNA and microsatellites (Figure 5F, *r* = 0.209; *p* = 0.001).



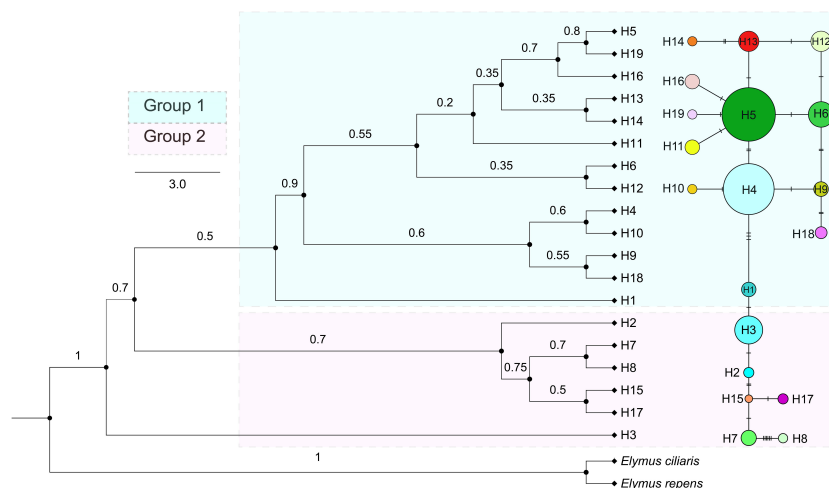


FIGURE 2

Phylogenetic tree and haplotype structure of *E. nutans* based on cpDNA sequences. Only bootstrap values higher than 0.5 are denoted above branches, the color combination of each circle represents the composition of different haplotypes (H1 – H19), and the size of the circle is proportional to the population size.

The variation of alleles is likely attributed to adaptation to external environmental factors (Ramírez-Valiente et al., 2010). Here, we examined the correlation between the number of effective alleles (N_e) in each population and the environmental factors. Three of the markers detected a significant correlation between the number of effective alleles detected and ecogeographical data (Figure 6). Specifically, the N_e detected by marker EN57 and EN80 exhibited a negative correlation with annual precipitation (Figures 6A, B), while the N_e detected by marker EN80 showed a significant positive correlation with the seasonal variability of precipitation (Figure 6C). Additionally, the N_e detected by marker EN5 displayed a significant positive correlation with the annual temperature range (Figure 6D). Upon annotating these loci with gene functions within the database, it was revealed that the marker EN57 locus is linked to glutathione S-transferase T1. In contrast, the marker EN5 and EN80 loci exhibited

homology with hypothetical proteins in species such as *Aegilops tauschii*, *Hordeum vulgare*, and *Setaria italica*, among others. Notably, the precise functions of these hypothetical proteins remain uncharacterized and warrant further accurate annotation.

4 Discussion

4.1 Population genetic variation

Genetic diversity, encompassing the total genetic variation among individuals within different populations of a species, serves as a crucial indicator of population adaptability to changing environments (Bell and Collins, 2008). Given the climate sensitivity of the QTP region, investigating the genetic diversity of species in this area holds significant importance (Wambulwa et al., 2021). *E. nutans*, a prominent herbaceous plant widely distributed in the QTP, has demonstrated exceptionally high levels of genetic variation through various approaches, including phenotype, ISSR, and SSR analyses (Chen et al., 2009; Li et al., 2023). However, prior studies on genetic variation primarily relied on individual or pooled genotyping resources, which might not sufficiently capture the spatial variation of genetic diversity among populations within this species.

To bridge the existing knowledge gap, a comprehensive analysis was undertaken, examining genetic variability among 361 specimens drawn from 35 distinct populations throughout the QTP region, with a specific focus on variations in nuclear and plastid fragments. The study unveiled substantial genetic diversity in both nuclear and plastid genes. Notably, the number of alleles ($N_a = 2.65$) and effective alleles ($N_e = 2.21$) identified through EST-SSR markers were in close alignment with those reported in a prior investigation of the closely associated and sympatric species, *E. brevistaratus* (Li et al., 2022; $N_a = 2.71$; $N_e = 2.28$). Nevertheless, it is imperative to acknowledge that the Shannon information index

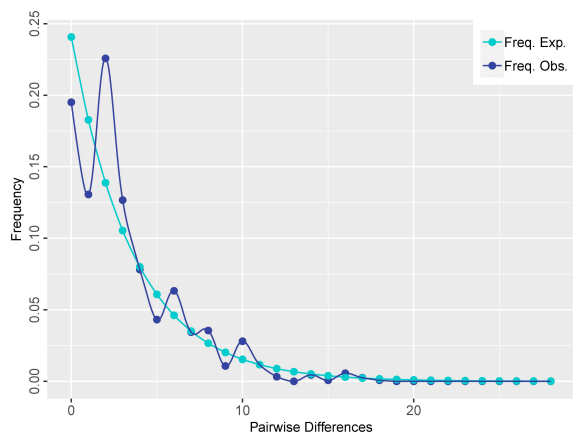
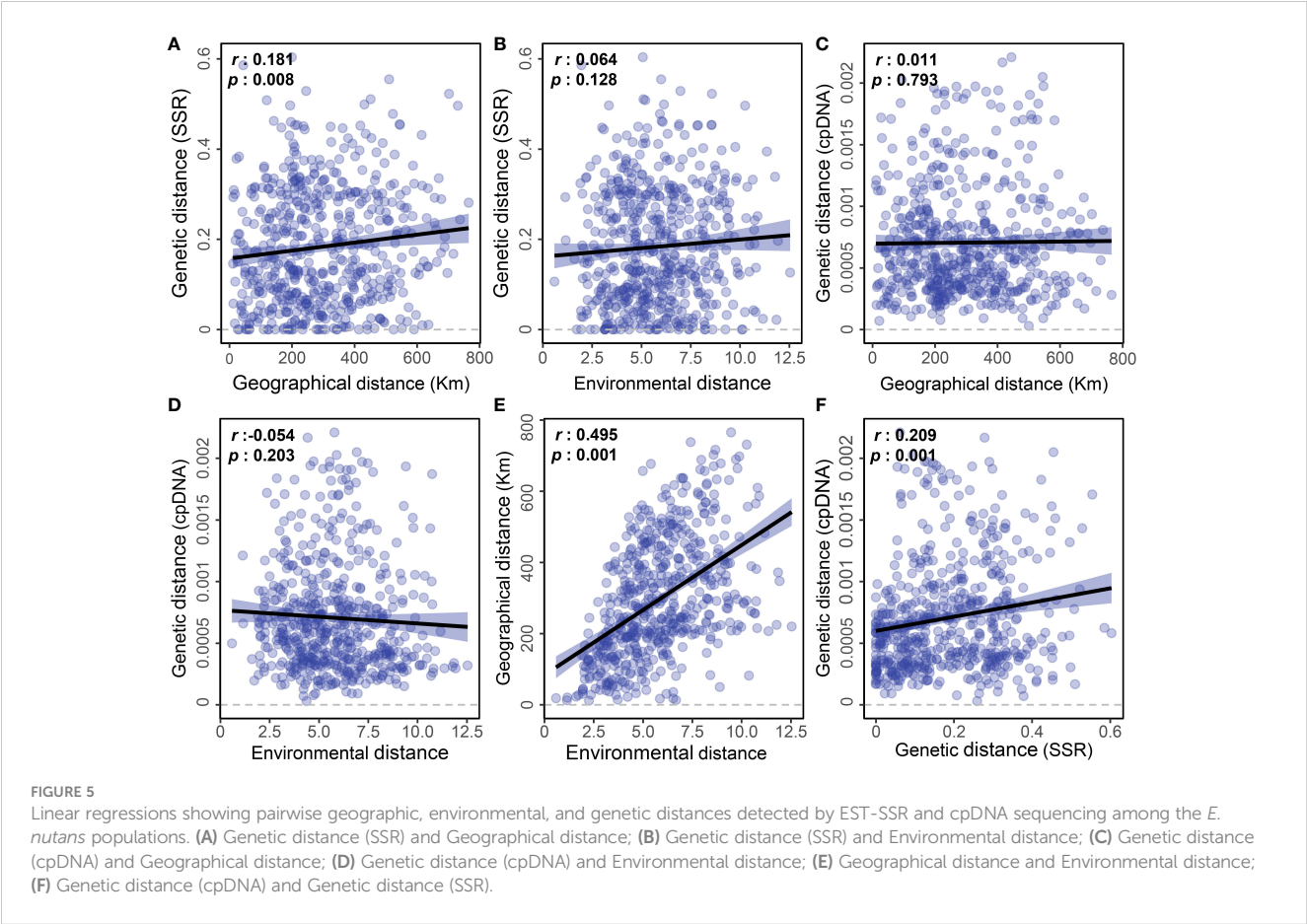
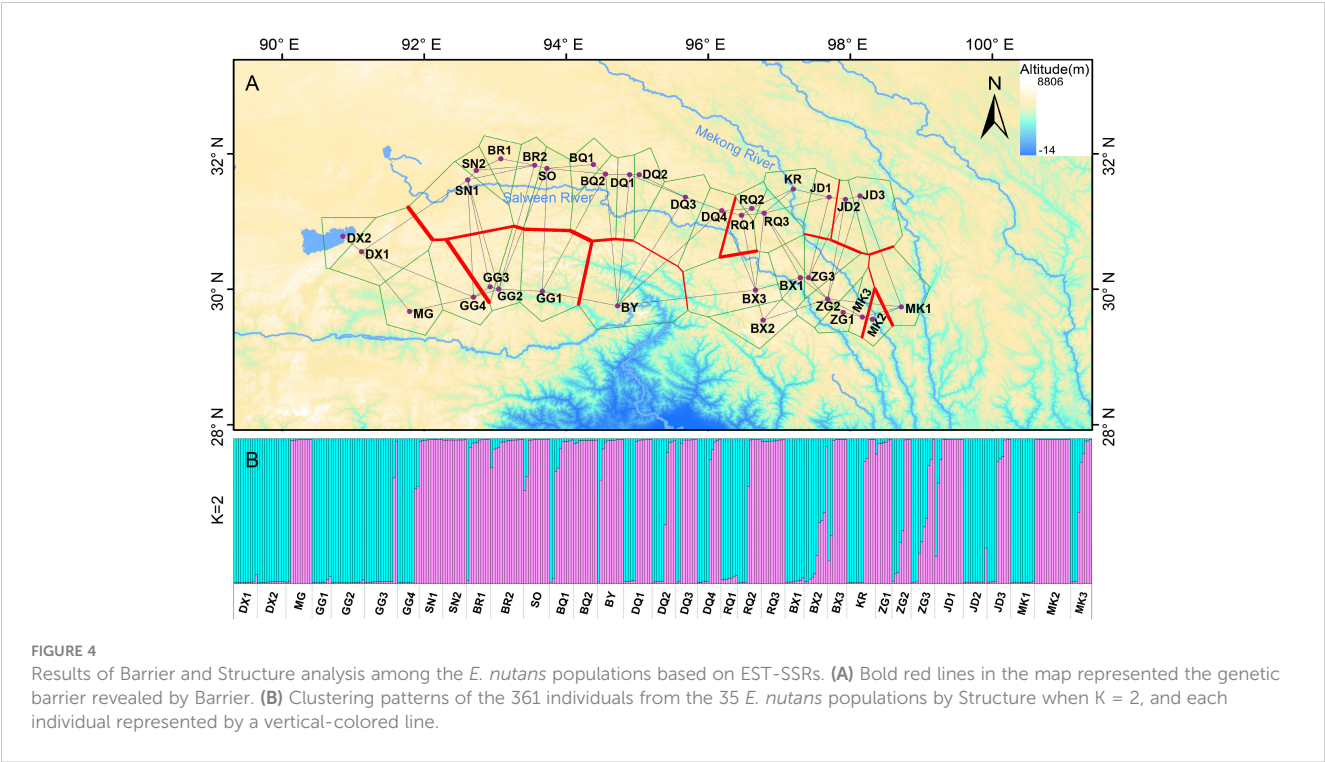


FIGURE 3

Mismatch distribution analysis plots based on cpDNA sequences for *E. nutans* populations.



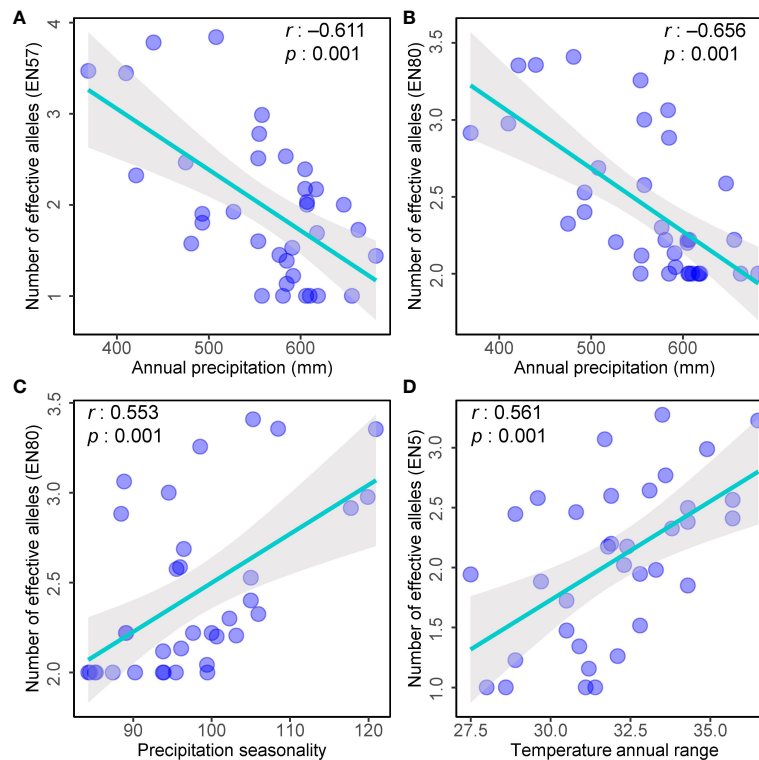


FIGURE 6

Scatter plots showing the relationships between the environmental factors and number of effective alleles (N_e) in each population. (A) Number of effective alleles (EN57) and Annual precipitation; (B) Number of effective alleles (EN80) and Annual precipitation; (C) Number of effective alleles (EN80) and Precipitation seasonality; (D) Number of effective alleles (EN5) and Temperature annual range.

(I) recorded for the *E. nutans* population ($I = 0.46$) was significantly elevated in comparison to that of *E. breviaristatus* ($I = 0.13$), underscoring a distinctive and intriguing pattern of genetic diversity. As a species with an extensive distribution, *E. nutans* demonstrates a more homogeneous allele distribution, potentially attributed to its propensity to sustain larger population sizes and engage in more substantial gene flow across various populations (Magota et al., 2021). Conversely, the endangered status of *E. breviaristatus* may exert limitations on gene flow and genetic exchange, culminating in a diminished Shannon's information index (I). This stark contrast elucidates the impact of species distribution and conservation status on genetic diversity and gene flow within these species.

A high H_d (0.805) and π (0.67×10^{-3}) were revealed in our study, indicating values lower than those observed in the related and widespread species of *Elymus sibiricus* in the QTP region detected based on cpDNA ($H_d = 0.834$, $\pi = 1.08 \times 10^{-3}$). Conversely, these values surpassed those documented for *E. sibiricus* populations in Xinjiang and northern China, which exhibit respective H_d values of 0.527 and 0.543, and π values of 0.042×10^{-3} and 0.53×10^{-3} , respectively (Xiong et al., 2022). Similarly, this high degree of genetic diversity has also been confirmed in other plants on the Tibetan Plateau, such as *Stellera chamaejasme* ($H_d = 0.834$) (Zhang et al., 2010), *Iris lozyi* ($H_d = 0.820$) (Zhang et al., 2021). This observation corroborates the notion that the QTP serves as a focal point for the differentiation in most species, including *Elymus*

species. Further, on a population scale, specific groups displaying elevated haplotype counts and notable haplotype diversity, as exemplified by populations such as DQ3, DQ4, RQ2, and others, are strategically positioned within the geographical span delimited by the Salween River and the Mekong River. These observations substantiate the hypothesis that this geographic region holds substantial significance as a sanctuary during ice ages for *E. nutans*. Furthermore, the locale has historically held a pivotal role as a biogeographic demarcation within the East Asian floral context, commonly recognized as the Ward Line-Mekong-Salween Divide (MSD) (Ward, 1921; Luo et al., 2017), which is highly suggestive of its potential as a hub for *E. nutans* diversification. In addition, this hypothesis is supported by the previous findings of Yu et al (Yu et al., 2019), who identified this field as one of the nine evolutionary hotspots in the QTP, bolsters the proposition of its paramount significance in the evolutionary trajectory of *E. nutans*, namely the eastern part of Nyenchen Tanglha Mountains.

4.2 Genetic structure and barriers

The analysis of cpDNA haplotypes and microsatellite structures in this study suggests that the genetic differentiation among the 35 populations of *E. nutans* lacks distinct definition. This conclusion is supported by the presence of individuals displaying mixed genetic backgrounds within the majority of populations. Moreover, the

AMOVA results emphasize that genetic variation primarily arises within populations, with comparatively lower levels of genetic divergence observed among populations. We attribute this observed divergence to *E. nutans*' wider ecological niche and larger effective population size, which likely enhance individuals' migration and dispersion across diverse habitats, facilitating gene flow among populations. Such occurrences are common among widely distributed plant species and are pivotal in maintaining genetic diversity and adaptability within populations (Broadhurst et al., 2018; Kahl et al., 2021; Veto et al., 2023).

Nonetheless, there still exist pronounced genetic barriers among certain adjacent populations. These barriers could be attributed to specific factors such as geographic isolation, environmental differences, or ecological niche divergence (Ryan et al., 2017; Yu et al., 2021), which collectively restrict gene flow among populations. As indicated by the results obtained from the Barrier analysis (Figure 3), our study reveals that the genetic barriers detected are primarily aligned with the Salween River and the Mekong River, impeding gene flow between populations situated on opposite sides of these rivers. The Ward Line-MSD, formed by these two rivers, holds significance as a major biogeographic boundary within the East Asian plant region (Ward, 1921). It has been extensively studied and confirmed to exist in various species, such as *Marmoritis complanatum*, *Koenigia forrestii*, and *Sinopodophyllum hexandrum* (Li et al., 2011; Luo et al., 2017; Rana et al., 2023). Our research, focusing on the intraspecific genetic differentiation of populations and operating at a finer scale, further underscores the importance of MSD in the dynamic process of species differentiation. The elucidation of these intricacies significantly enriches our comprehension of species evolution and ecological dynamics. Such an enhanced understanding, in turn, facilitates the implementation of more refined and effective strategies for the collection, conservation, and management of the germplasm resources pertaining to this species.

4.3 The role of ecogeographical factors on genetic divergence and adaptation

Although our study highlights the extensive effect of the MSD concerning the genetic separation of *E. nutans* populations, which restricts gene flow and leads to prompt genetic divergence, a noteworthy phenomenon presents itself in the form of a weak yet statistically significant pattern of isolation by distance (IBD) among populations only demonstrated via microsatellite data. In contrast to the robust IBD pattern uncovered in our earlier study of *E. breviaristatus* populations (Li et al., 2022), the relatively weaker or even absent IBD pattern discerned in *E. nutans* populations further reinforces the notion of heightened dispersal capabilities within these populations. However, the insights gleaned from the cpDNA analysis do not corroborate this phenomenon. It is posited that this disparity may originate from the fundamental differences in transmission modes, rates of genetic drift, and migration patterns inherent to nuclear and plastid DNA, culminating in the divergent genetic configurations manifested in these two distinct genetic substrates (Soltis and Kuzoff, 1995).

Perennial plants strategically accumulate diverse allelic variants in response to various environmental conditions, reflecting their adaptive mechanism to external selection pressures (Castillo et al., 2010; Sork

et al., 2010; Raschke et al., 2015). In our investigation, we identified significant divergence within three analyzed loci among populations situated in regions marked by more extreme climatic conditions, such as low precipitation or imbalances in precipitation and temperature. Of particular significance, notable correlations were observed between the number of effective alleles and specific environmental factors, indicating a discernible influence of natural selection on these genetic markers, thereby facilitating localized genetic differentiation. While two of these loci lack prior annotations, it is noteworthy that marker EN67 is unequivocally associated with a glutathione S-transferase T1 in wheat and *Arabidopsis*. Glutathione S-transferase, an essential component of the glutathione antioxidant system, plays a pivotal role in managing oxidative stress and detoxifying harmful compounds within plants (Sappl et al., 2009; Nianiou-Obeidat et al., 2017). We posit that this functional attribute may contribute to allelic divergence observed in arid regions. Additionally, although the marker EN5 and EN80 motifs lack established annotations, we hypothesize that they are linked to mechanisms enabling adaptation to climatic extremes. These motifs hold promise as potential candidate genes for genetic breeding in wheat plants, particularly to enhance resilience against environmental challenges. Further research is warranted to unravel the precise functional implications of these loci and their roles in plant adaptation.

Collectively, our findings, coupled with our prior results, implies that concerning gene loci affected by environmental selection, the lack of a clear-cut isolation by environment (IBE) pattern signifies a more intricate association between genetic differentiation and environmental adaptation than previously presumed. This complexity likely arises from a confluence of interacting factors, highlighting the multifaceted nature of the evolutionary processes governing population differentiation and adaptation.

5 Conclusions

In this comprehensive study, the genetic dynamics of *E. nutans* populations across the QTP were explored using a combination of cpDNA and microsatellite analyses. Significant genetic diversity within and among populations was revealed through the analysis of haplotype distribution and phylogenetic relationships. Mismatch distribution analysis and neutrality tests indicated a complex demographic history with no evidence of recent population expansion. The examination of population genetic differentiation unveiled the significant role of geographic barriers in shaping the genetic landscape of *E. nutans*, with the Salween River and Mekong River identified as potent genetic boundaries that impede gene flow between populations. Although the impact of geographic distance on genetic differentiation appears to be minimal, significant correlations have been identified between certain microsatellite loci and environmental factors, suggesting potential adaptability of these loci to climatic challenges. In summary, this study unveils the intricate genetic pathways of *E. nutans* within the dynamic QTP. The findings underscore the directive influence of geographical barriers and ecological factors on genetic differentiation and adaptation. The insights garnered from this research hold substantial importance for the conservation of germplasm resources and resistance breeding in the context of an ever-changing environment.

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/genbank/>, OR421574–OR423017.

Author contributions

JL: Conceptualization, Investigation, Methodology, Writing – original draft. XL: Investigation, Methodology, Writing – original draft. CZ: Methodology, Resources, Validation, Writing – original draft. QZ: Conceptualization, Validation, Writing – review & editing. SC: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Natural Science Foundation of China (No. 31900280), the Key Research & Development Program of Sichuan province (2021YFYZ0013) and

the Double First-Class program of Southwest Minzu University (CX2023017).

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.2024.1349641/full#supplementary-material>

References

- Aguiar-Melo, C., Zanella, C. M., Goetze, M., Palma-Silva, C., Hirsch, L. D., Neves, B., et al. (2019). Ecological niche modeling and a lack of phylogeographic structure in *Vriesea incurvata* suggest historically stable areas in the southern Atlantic Forest. *Am. J. Bot.* 106, 971–983. doi: 10.1002/ajb2.1317
- Bell, G., and Collins, S. (2008). Adaptation, extinction and global change. *Evol. Appl.* 1, 3–16. doi: 10.1111/j.1752-4571.2007.00011.x
- Birky, C. W., Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics* 103, 513–527. doi: 10.1093/genetics/103.3.513
- Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., et al. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537. doi: 10.1371/journal.pcbi.1003537
- Broadhurst, L. M., Mellick, R., Knerr, N., Li, L., and Supple, M. A. (2018). Land availability may be more important than genetic diversity in the range shift response of a widely distributed eucalypt, *Eucalyptus melliodora*. *For. Ecol. Manage.* 409, 38–46. doi: 10.1016/j.foreco.2017.10.024
- Castillo, A., Dorado, G., Feuillet, C., Sourdille, P., and Hernandez, P. (2010). Genetic structure and ecogeographical adaptation in wild barley (*Hordeum chilense* Roemer et Schultes) as revealed by microsatellite markers. *BMC Plant Biol.* 10, 266. doi: 10.1186/1471-2229-10-266
- Chen, S. Y., Ma, X., Zhang, X. Q., and Chen, Z. H. (2009). Genetic variation and geographical divergence in *Elymus nutans* Griseb. (Poaceae: Triticeae) from West China. *Biochem. Syst. Ecol.* 37, 716–722. doi: 10.1016/j.bse.2009.12.005
- Decroocq, V., Favé, M. G., Hagen, L., Bordenave, L., and Decroocq, S. (2003). Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor. Appl. Genet.* 106, 912–922. doi: 10.1007/s00122-002-1158-z
- Du, F. K., Wang, T., Wang, Y., Ueno, S., and de Lafontaine, G. (2020). Contrasted patterns of local adaptation to climate change across the range of an evergreen oak, *Quercus aquifolioides*. *Evol. Appl.* 13, 2377–2391. doi: 10.1111/eva.13030
- Earl, D. A., and Vonholdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7
- Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x
- Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x
- Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x
- Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709. doi: 10.1093/genetics/133.3.693
- Gray, M. M., St Amand, P., Bello, N. M., Galliard, M. B., Knapp, M., Garrett, K. A., et al. (2014). Ecotypes of an ecologically dominant prairie grass (*Andropogon gerardii*) exhibit genetic divergence across the U.S. Midwest grasslands' environmental gradient. *Mol. Ecol.* 23, 6011–6028. doi: 10.1111/mec.12993
- Guo, R., Zhang, Y. H., Zhang, H. J., Landis, J. B., Zhang, X., Wang, H. C., et al. (2022). Molecular phylogeography and species distribution modelling evidence of 'oceanic' adaptation for *Actinidia eriantha* with a refugium along the oceanic-continental gradient in a biodiversity hotspot. *BMC Plant Biol.* 22, 89. doi: 10.1186/s12870-022-03464-5
- Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233
- Jump, A. S., Marchant, R., and Penuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends Plant Sci.* 14, 51–58. doi: 10.1016/j.tplants.2008.10.002
- Kahl, S. M., Kappel, C., Joshi, J., and Lenhard, M. (2021). Phylogeography of a widely distributed plant species reveals cryptic genetic lineages with parallel phenotypic responses to warming and drought conditions. *Ecol. Evol.* 11, 13986–14002. doi: 10.1002/ecs3.8103
- Kou, Y. X., Wu, Y. X., Jia, D. R., Li, Z. H., and Wang, Y. J. (2014). Range expansion, genetic differentiation, and phenotypic adaption of *Hippophae neurocarpa* (Elaeagnaceae) on the Qinghai-Tibet Plateau. *J. Syst. Evol.* 52, 303–312. doi: 10.1111/jse.12063
- Lei, Y. X., Fan, X., Sha, L. N., Wang, Y., Kang, H. Y., Zhou, Y. H., et al. (2022). Phylogenetic relationships and the maternal donor of *Roegneria* (Triticeae: Poaceae) based on three nuclear DNA sequences (ITS, Acc1, and Pgk1) and one chloroplast region (*trnL-F*). *J. Syst. Evol.* 60, 305–318. doi: 10.1111/jse.12664

- Leigh, J. W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410
- Li, J., Ma, S. E., Jiang, K. K., Zhang, C. B., Liu, W. H., and Chen, S. Y. (2022). Drivers of population divergence and genetic variation in *Elymus brevibractatus* (Keng) Keng f. (Poaceae: Triticeae), an endemic perennial herb of the Qinghai-Tibet plateau. *Front. Ecol. Evol.* 10, 1068739. doi: 10.3389/fevo.2022.1068739
- Li, J., Tian, H. Q., Ji, W. Q., Zhang, C. B., and Chen, S. Y. (2023). Inflorescence trait diversity and genotypic differentiation as influenced by the environment in *Elymus nutans* griseb. from Qinghai-Tibet Plateau. *Agronomy* 13, 1004. doi: 10.3390/agronomy13041004
- Li, Y., Zhai, S. N., Qiu, Y. X., Guo, Y. P., Ge, X. J., and Comes, H. P. (2011). Glacial survival east and west of the 'Mekong-Salween Divide' in the Himalaya-Hengduan Mountains region as revealed by AFLPs and cpDNA sequence variation in *Sinopodophyllum hexandrum* (Berberidaceae). *Mol. Phylogenet. Evol.* 59, 412–424. doi: 10.1016/j.ympev.2011.01.009
- Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187
- Liu, Y. P., Ren, Z. M., Harris, A. J., Peterson, P. M., Wen, J., and Su, X. (2018). Phylogeography of *Orinus* (Poaceae), a dominant grass genus on the Qinghai-Tibet Plateau. *Bot. J. Linn. Soc.* 186, 202–223. doi: 10.1093/botlinnean/box091
- Luo, D., Xu, B., Li, Z.-M., and Sun, H. (2017). The 'Ward Line-Mekong-Salween Divide' is an important floristic boundary between the eastern Himalaya and Hengduan Mountains: evidence from the phylogeographical structure of subnival herbs *Marmoritis complanatum* (Lamiaceae). *Bot. J. Linn. Soc.* 185, 482–496. doi: 10.1093/botlinnean/box067
- Magota, K., Sakaguchi, S., Hirota, S. K., Tsumamoto, Y., Suyama, Y., Akai, K., et al. (2021). Comparative analysis of spatial genetic structures in sympatric populations of two riparian plants, *Saxifraga acerifolia* and *Saxifraga fortunei*. *Am. J. Bot.* 108, 680–693. doi: 10.1002/ajb2.1644
- Manni, F., Guerard, E., and Heyer, E. (2004). Geographic patterns of (genetic, morphological, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Hum. Biol.* 76, 173–190. doi: 10.1353/hub.2004.0034
- Nianiou-Obeidat, I., Madesis, P., Kissoudis, C., Voulgari, G., Chronopoulou, E., Tsafiris, A., et al. (2017). Plant glutathione transferase-mediated stress tolerance: functions and biotechnological applications. *Plant Cell Rep.* 36, 791–805. doi: 10.1007/s00299-017-2139-7
- Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460
- Petit, R. J., Kremer, A., and Wagner, D. B. (1993). Geographic structure of chloroplast DNA polymorphisms in European oaks. *Theor. Appl. Genet.* 87, 122–128. doi: 10.1007/BF00223755
- Pradheep, K., Singh, M., Sultan, S. M., Singh, K., Parimalan, R., and Ahlawat, S. P. (2019). Diversity in wild relatives of wheat: an expedition collection from cold-arid Indian Himalayas. *Genet. Resour. Crop Ev.* 66, 275–285. doi: 10.1007/s10722-018-0706-6
- Proft, K. M., Bateman, B. L., Johnson, C. N., Jones, M. E., Pauza, M., and Burridge, C. P. (2021). The effects of weather variability on patterns of genetic diversity in Tasmanian bettongs. *Mol. Ecol.* 30, 1777–1790. doi: 10.1111/mec.15847
- Provan, J., Powell, W., and Hollingsworth, P. M. (2001). Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol. Evol.* 16, 142–147. doi: 10.1016/S0169-5347(00)02097-8
- Ramírez-Valiente, J. A., Lorenzo, Z., Soto, A., Valladares, F., Gil, L., and Aranda, I. (2010). Natural selection on cork oak: allele frequency reveals divergent selection in cork oak populations along a temperature cline. *Evol. Ecol.* 24, 1031–1044. doi: 10.1007/s10682-010-9365-6
- Rana, H. K., Rana, S. K., Luo, D., and Sun, H. (2023). Existence of biogeographic barriers for the long-term Neogene-Quaternary divergence and differentiation of *Koenigia forrestii* in the Himalaya-Hengduan Mountains. *Bot. J. Linn. Soc.* 201, 230–253. doi: 10.1093/botlinnean/boac045
- Raschke, A., Ibanez, C., Ullrich, K. K., Anwer, M. U., Becker, S., Glockner, A., et al. (2015). Natural variants of ELF3 affect thermomorphogenesis by transcriptionally modulating PIF4-dependent auxin response genes. *BMC Plant Biol.* 15, 197. doi: 10.1186/s12870-015-0566-6
- Ryan, S. F., Fontaine, M. C., Scriber, J. M., Pfreder, M. E., O'Neil, S. T., and Hellmann, J. J. (2017). Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climatic gradient. *Mol. Ecol.* 26, 4725–4742. doi: 10.1111/mec.14236
- Sappl, P. G., Carroll, A. J., Clifton, R., Lister, R., Whelan, J., Millar, A. H., et al. (2009). The Arabidopsis glutathione transferase gene family displays complex stress regulation and co-silencing multiple genes results in altered metabolic sensitivity to oxidative stress. *Plant J.* 58, 53–68. doi: 10.1111/j.1365-3113.2008.03761.x
- Setsuko, S., Sugai, K., Tamaki, I., Takayama, K., Kato, H., and Yoshimaru, H. (2020). Genetic diversity, structure, and demography of *Pandanus boninensis* (Pandanaceae) with sea drifted seeds, endemic to the Ogasawara Islands of Japan: Comparison between young and old islands. *Mol. Ecol.* 29, 1050–1068. doi: 10.1111/mec.15383
- Shahzad, K., Jia, Y., Chen, F. L., Zeb, U., and Li, Z. H. (2017). Effects of mountain uplift and climatic oscillations on phylogeography and species divergence in four endangered *Notopterygium* herbs. *Front. Plant Sci.* 8, 1929. doi: 10.3389/fpls.2017.01929
- Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., et al. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* 92, 142–166. doi: 10.3732/ajb.92.1.142
- Soltis, D. E., and Kuzoff, R. K. (1995). Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). *Evolution* 49, 727–742. doi: 10.2307/2410326
- Sork, V. L., Davis, F. W., Westfall, R., Flint, A., Ikegami, M., Wang, H., et al. (2010). Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Née) in the face of climate change. *Mol. Ecol.* 19, 3806–3823. doi: 10.1111/j.1365-294X.2010.04726.x
- Stojanova, B., Surinova, M., Klapšte, J., Kolarikova, V., Hadincova, V., and Munzbergova, Z. (2018). Adaptive differentiation of *Festuca rubra* along a climate gradient revealed by molecular markers and quantitative traits. *PloS One* 13, e0194670. doi: 10.1371/journal.pone.0194670
- Sun, J., Fu, B. J., Zhao, W. W., Liu, S. L., Liu, G. H., Zhou, H. K., et al. (2021). Optimizing grazing exclusion practices to achieve Goal 15 of the sustainable development goals in the Tibetan Plateau. *Sci. Bull.* 66, 1493–1496. doi: 10.1016/j.scib.2021.03.014
- Sun, M., Huang, D. J., Zhang, A. L., Khan, I., Yan, H. D., Wang, X. S., et al. (2020). Transcriptome analysis of heat stress and drought stress in pearl millet based on Pacbio full-length transcriptome sequencing. *BMC Plant Biol.* 20, 323. doi: 10.1186/s12870-020-02530-0
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197
- Veto, N. M., Postolache, D., Escudero, F. L. G., Vajana, E., Braga, R. B., Salgueiro, F., et al. (2023). Population structure and signals of local adaptation in *Eugenia uniflora* (Myrtaceae), a widely distributed species in the Atlantic Forest. *Bot. J. Linn. Soc.* 201, 100–113. doi: 10.1093/botlinnean/boac012
- Vu, D. D., Shah, S. N. M., Pham, M. P., Bui, V. T., Nguyen, M. T., and Nguyen, T. P. T. (2020). De novo assembly and Transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* Ha et Grushv., including the development of EST-SSR markers for population genetics. *BMC Plant Biol.* 20, 358. doi: 10.1186/s12870-020-02571-5
- Wambulwa, M. C., Milne, R., Wu, Z. Y., Spicer, R. A., Provan, J., Luo, Y. H., et al. (2021). Spatiotemporal maintenance of flora in the Himalaya biodiversity hotspot: Current knowledge and future perspectives. *Ecol. Evol.* 11, 10794–10812. doi: 10.1002/ece3.7906
- Wang, R. R. C., and Lu, B. R. (2014). Biosystematics and evolutionary relationships of perennial Triticeae species revealed by genomic analyses. *J. Syst. Evol.* 52, 697–705. doi: 10.1111/jse.12084
- Wang, Y., Liang, Q., Hao, G., Chen, C., and Liu, J. (2018). Population genetic analyses of the endangered alpine *Sinadoxa corydalis* (Adoxaceae) provide insights into future conservation. *Biodivers. Conserv.* 27, 2275–2291. doi: 10.1007/s10531-018-1537-7
- Ward, F. K. (1921). The Mekong-Salween Divide as a geographical barrier. *Geographical J.* 58, 49–56. doi: 10.2307/1780720
- Xiong, Y., Lei, X., Bai, S., Xiong, Y., Liu, W., Wu, W., et al. (2021). Genomic survey sequencing, development and characterization of single- and multi-locus genomic SSR markers of *Elymus sibiricus* L. *BMC Plant Biol.* 21, 3. doi: 10.1186/s12870-020-02770-0
- Xiong, Y., Xiong, Y., Shu, X., Yu, Q., Lei, X., Li, D., et al. (2022). Molecular phylogeography and intraspecific divergences in siberian wildrye (*Elymus sibiricus* L.) wild populations in China, inferred from chloroplast DNA sequence and cpSSR markers. *Front. Plant Sci.* 13, 862759. doi: 10.3389/fpls.2022.862759
- Yu, H., Favre, A., Sui, X., Chen, Z., Qi, W., Xie, G., et al. (2019). Mapping the genetic patterns of plants in the region of the Qinghai-Tibet Plateau: Implications for conservation strategies. *Divers. Distrib.* 25, 310–324. doi: 10.1111/ddi.12847
- Yu, Y. L., Wang, H. C., Yu, Z. X., Schinnerl, J., Tang, R., Geng, Y. P., et al. (2021). Genetic diversity and structure of the endemic and endangered species *Aristolochia delavayi* growing along the Jinsha River. *Plant Diversity* 43, 225–233. doi: 10.1016/j.pld.2020.12.007
- Zhang, D., Gao, F. L., Jakovlic, I., Zou, H., Zhang, J., Li, W. X., et al. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* 20, 348–355. doi: 10.1111/1755-0998.13096
- Zhang, G., Han, Y., Wang, H., Wang, Z., Xiao, H., and Sun, M. (2021). Phylogeography of *Iris loczyi* (Iridaceae) in Qinghai-Tibet Plateau revealed by chloroplast DNA and microsatellite markers. *Acta Bot. Sin.* 47, 1607–1616. doi: 10.1093/aobpla/plab070
- Zhang, Y. H., Volis, S., and Sun, H. (2010). Chloroplast phylogeny and phylogeography of *Stellera chamaejasme* in Qinghai-Tibet Plateau and in adjacent regions. *Mol. Phylogenet. Evol.* 57, 1162–1172. doi: 10.1016/j.ympev.2010.08.033
- Zheng, H. Y., Guo, X. L., Price, M., He, X. J., and Zhou, S. D. (2021). Effects of mountain uplift and climatic oscillations on phylogeography and species divergence of *Chamaesium* (Apiaceae). *Front. Plant Sci.* 12, 673200. doi: 10.3389/fpls.2021.673200
- Zwyrková, J., Blavet, N., Doležalová, A., Čápal, P., Said, M., Molnár, I., et al. (2022). Draft sequencing crested wheatgrass chromosomes identified evolutionary structural changes and genes and facilitated the development of SSR markers. *Int. J. Mol. Sci.* 23, 3191. doi: 10.3390/ijms23063191



OPEN ACCESS

EDITED BY

Maarten Van Zonneveld,
World Vegetable Center, Taiwan

REVIEWED BY

Photini V. Mylona,
Hellenic Agricultural Organisation (HAO),
Greece
Katherine Steele,
Bangor University, United Kingdom

*CORRESPONDENCE

Sónia Negrão
✉ sonia.negrão@ucd.ie

RECEIVED 28 July 2023

ACCEPTED 28 February 2024

PUBLISHED 20 March 2024

CITATION

Bernád V, Al-Tamimi N, Langan P, Gillespie G,
Dempsey T, Henchy J, Harty M, Ramsay L,
Houston K, Macaulay M, Shaw PD, Raubach S,
McDonnell KP, Russell J, Waugh R,
Khodaeiaminjan M and Negrão S (2024)
Unlocking the genetic diversity and
population structure of the newly
introduced two-row spring European
Heritage Barley collection (ExHIBiT).
Front. Plant Sci. 15:1268847.
doi: 10.3389/fpls.2024.1268847

COPYRIGHT

© 2024 Bernád, Al-Tamimi, Langan, Gillespie,
Dempsey, Henchy, Harty, Ramsay, Houston,
Macaulay, Shaw, Raubach, McDonnell, Russell,
Waugh, Khodaeiaminjan and Negrão. 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.

Unlocking the genetic diversity and population structure of the newly introduced two-row spring European Heritage Barley collection (ExHIBiT)

Villő Bernád¹, Nadia Al-Tamimi¹, Patrick Langan¹,
Gary Gillespie², Timothy Dempsey¹, Joey Henchy¹, Mary Harty²,
Luke Ramsay³, Kelly Houston³, Malcolm Macaulay³,
Paul D. Shaw⁴, Sebastian Raubach⁴, Kevin P. McDonnell^{2,5},
Joanne Russell³, Robbie Waugh^{3,6}, Mortaza Khodaeiaminjan¹
and Sónia Negrão^{1*}

¹School of Biology and Environmental Science, University College Dublin, Dublin, Ireland, ²School of Agriculture and Food Science, University College Dublin, Dublin, Ireland, ³Cell and Molecular Sciences, The James Hutton Institute, Dundee, United Kingdom, ⁴Department of Information and Computational Sciences, The James Hutton Institute, Dundee, United Kingdom, ⁵School of Biosystems Engineering, University College Dublin, Dublin, Ireland, ⁶Division of Plant Sciences, University of Dundee at The James Hutton Institute, Dundee, United Kingdom

In the last century, breeding programs have traditionally favoured yield-related traits, grown under high-input conditions, resulting in a loss of genetic diversity and an increased susceptibility to stresses in crops. Thus, exploiting understudied genetic resources, that potentially harbour tolerance genes, is vital for sustainable agriculture. Northern European barley germplasm has been relatively understudied despite its key role within the malting industry. The European Heritage Barley collection (ExHIBiT) was assembled to explore the genetic diversity in European barley focusing on Northern European accessions and further address environmental pressures. ExHIBiT consists of 363 spring-barley accessions, focusing on two-row type. The collection consists of landraces (~14%), old cultivars (~18%), elite cultivars (~67%) and accessions with unknown breeding history (~1%), with 70% of the collection from Northern Europe. The population structure of the ExHIBiT collection was subdivided into three main clusters primarily based on the accession's year of release using 26,585 informative SNPs based on 50k iSelect single nucleotide polymorphism (SNP) array data. Power analysis established a representative core collection of 230 genotypically and phenotypically diverse accessions. The effectiveness of this core collection for conducting statistical and association analysis was explored by undertaking genome-wide association studies (GWAS) using 24,876 SNPs for nine phenotypic traits, four of which were associated with SNPs. Genomic regions overlapping with previously characterised flowering genes (HvZTLb) were identified, demonstrating the utility of the ExHIBiT core collection for locating genetic regions that determine important traits. Overall,

the ExHIBiT core collection represents the high level of untapped diversity within Northern European barley, providing a powerful resource for researchers and breeders to address future climate scenarios.

KEYWORDS

barley, genetic resources, agronomic characterization, germplasm collection, genome-wide association studies, plant phenotyping

1 Introduction

Food security is a pressing global issue, with agricultural production facing severe yield penalties due to abiotic stresses caused by climate change (Prosekov & Ivanova, 2018; Masson-Delmotte et al., 2021; Sasidharan et al., 2021). Barley is the fourth most produced cereal crop, and is used for human food, animal feed and malting alcoholic drinks. Therefore, it is a key crop for food security as it can be grown in many marginal communities where very few other crops survive (Newton et al., 2011).

Barley was domesticated around 10,000 years ago in the Fertile Crescent (Badr et al., 2000) and has since undergone continuous selection and breeding. Traditional breeding methods and recent advances in genetic engineering have significantly increased barley yields (Harwood, 2019). However, due to past population bottlenecks, genetic drift, and inbreeding, modern barley cultivars have a narrow genetic basis compared to landraces and wild ancestors (Schmidt et al., 2023). This has resulted in modern cultivars becoming more susceptible to stress and less adaptable to changing environments (Tanksley and McCouch, 1997; Caldwell et al., 2006). Landraces and cultivars bred before the Green Revolution exhibit improved stress resilience and adaptation to their environments (Slama et al., 2018; Marone et al., 2021), but their exceptional genetic potential remains largely uncharacterized (Newton et al., 2010, 2011; Monteagudo et al., 2019).

Natural genetic diversity is a key pillar of plant breeding, with most breeding techniques heavily relying on it as the canvas for breeders to improve and develop new cultivars. To screen for genetic diversity, it is important to have a diverse collection with a large number of accessions. Collections with higher levels of diversity are more likely to harbour resilience, that has of yet remained untapped. In addition, in association methods such as genome-wide association studies (GWAS), variation is essential for the establishment of a relationship between markers and traits (Korte and Farlow, 2013). However, phenotyping of a large collection can be costly and time consuming, and some accessions can be highly similar, resulting in additional work without the associated benefits (Frankel and Brown, 1984; Berger et al., 2012; Araus et al., 2018). The core collection strategy minimises repetitiveness within a collection while preserving genetic

diversity and reducing phenotyping costs (Brown, 1989a, b). To establish a core collection, highly similar accessions within the collection need to be identified and removed, which can be achieved using the passport data of the accessions, population structure and phenotypic characterization of the collection. The barley research community has established several core collections to study diversity and evolution as well as to screen for biotic and abiotic stress responses. For instance, the International Barley Core Collection (BCC) was established in 1989 (van Hintum, 1994; van Treuren et al., 2006) to reflect the diversity of barley worldwide, and since its inception, it has been used to better understand barley's diversity and evolution (Liu et al., 1999, 2000, 2001, 2002). The Spanish Core Collection was assembled to show the diversity of barley found within the Spanish Germplasm Bank (Igartua et al., 1998), and the Czech winter barley collection (Dreiseitl, 2021; Dreiseitl and Nesvadba, 2021) has been recently established to screen for disease resistance.

The majority of barley collections have been primarily focused on accessions from Southern Europe, disregarding the specific needs and challenges of barley production in Northern Europe (Pasam et al., 2014; Selçuk et al., 2015). In Northern Europe, which accounts for 25% of the EU's malting production capacity (Euromalt, 2021), the impact of climate change on cultivation is becoming increasingly challenging. Malting is the premium use product for barley (Hertrich, 2013), with two-row spring barley being the main target of selective breeding for malting quality and yield (Tondelli et al., 2013). The growing craft brewing (Guido, 2019) and distilling markets (Umego and Barry-Ryan, 2022) have created an increased demand for barley with favourable malting quality and unique taste. The Northern European region is expected to experience deteriorating agricultural conditions due to climate change, characterized by more frequent extreme weather events, excessive precipitation, and even drought (Uleberg et al., 2014; Nolan and Flanagan, 2022), leading to a reduction in barley production (Xie et al., 2018). Assembling, characterising and utilising genetic resources capable of overcoming these threats will help ensure the resilience and productivity of barley, even in changing climatic conditions. Moreover, fostering barley production using low carbon emission methods will enhance the sustainability of the critically important malting industry (European Commission, 2020; Sleight, 2022).

This study aims to: (i) assemble a natural and diverse two-row spring barley collection (ExHIBiT) focusing on Northern European accessions and investigate its genetic and phenotypic diversity (ii) establish a core-collection of two-row spring barley for multiple purposes, and (iii) analyse the role of geographic origin and breeding history in the formation of the ExHIBiT genetic structure using the 50k iSelect SNP array (Bayer et al., 2017). The ExHIBiT collection is predominantly composed of Northern and Central European accessions due to their historical contribution for the malting industry in these regions. Characterisation of the ExHIBiT collection will promote the use of heritage barley as an untapped reservoir of genetic variation for breeders and support the identification of quantitative trait loci (QTL), facilitating the advance genetic and breeding research and tackling barley sustainability in Northern and Central Europe.

2 Materials and methods

2.1 Plant material

The ExHIBiT collection comprised of 363 two-row spring barley accessions, from several gene banks and collections, namely from the Department of Agriculture, Food and the Marine (DAFM), Ireland, Nordic Genetic Resource Centre (NordGen), Norway, the James Hutton Institute (JHI), UK. The germplasm material from JHI includes accessions from i) *Bustos-Korts et al., 2019*, namely from the Wheat and barley legacy for breeding improvement (WHEALBI) project (<https://www.whealbi.eu/>); ii), IMPROMALT project (Looseley et al., 2020); iii) 9k project and Heritage collection (Schmidt et al., 2019), as well as the Germplasm Resource Unit of the John Innes Centre (GRU-JIC) and JHI stocks. Accessions were selected based on their passport information, to reflect their genetic diversity and breeding history. Particular focus was placed on Northern and Central European accessions, taking into consideration previous population structure results of a European two-row spring barley collection (Saade et al., 2016). In this work, the term heritage is used to encompass the intricate crop breeding history of a region, which is in turn influenced by historic and ever-changing factors such as source material, climate, land type, evolving agricultural equipment, manufacturing processes and market demands. We further characterise these endemic resources into three groups. (i) Landraces: highly diverse material continually selected by producers over time and therefore well adapted to local environments; (ii) Old Cultivars: cultivars actively selected by formal breeding programs prior to the Green Revolution (Pre-1960s) and; (iii) Elite Cultivars: cultivars developed in the modern era of plant breeding (post-1960) and after the introduction of the Distinctiveness, Uniformity and Stability (DUS) testing in the United Kingdom. The completed list of accessions, including year of release and geographical origin based on information from JHI, the European search catalogue for plant genetic resources (EURISCO) and previously published works (Maxted et al., 2014; Weise et al., 2017; Faccini et al., 2021) are given in *Supplementary Table S1*.

2.2 Genotyping of the ExHIBiT collection

DNA was extracted from an individual plant utilising two-week-old leaf tissue of the accessions within the ExHIBiT collection using the DNeasy Plant Mini Kit (Qiagen, Germany) following the manufacturer's protocol. Plants for DNA extraction were grown in controlled conditions prior to any field experiment. Accessions were genotyped at the JHI using the Illumina Infinium *iSelect* HD 50k chip, which was designed to capture the most representative set of barley germplasm (Bayer et al., 2017). Physical positions of markers were based on the pseudo-molecule assembly of the most recently updated barley reference assembly- "Morex" V3 (Mascher et al., 2021). Markers with a call rate value lower than 90% and minor allele frequency (MAF) lower than 5% were removed using TASSEL (Bradbury et al., 2007). To identify any duplicate lines within the collection, standard R v3.6.0 0 (R Core Team, 2022) was used to calculate the similarity between the accessions by examining the percentage of markers sharing the same nucleotide. The SNP marker data for this study have been deposited in the European Variation Archive (EVA) at EMBL-EBI under accession number PRJEB67728 (<https://www.ebi.ac.uk/eva/?eva-study=PRJEB67728>).

2.3 Population structure and pedigree analysis

To determine ExHIBiT population structure, genotypic data was analysed using STRUCTURE V.2.3.4 software, which uses a Bayesian clustering approach to assign individuals to K subgroups (Pritchard et al., 2000). Five independent runs (K = 1 to 10) were performed with 50,000 burn-in periods, and 10,000 Markov Chain Monte Carlo iterations for each value of K. The best number of K was chosen using the ΔK method (Evanno et al., 2005) by running the Structure Harvester software (Earl and VonHoldt, 2012). Accessions were classified as belonging to a group if more than 50% of the markers belong to that group, otherwise they were classified as admixture.

To construct the Neighbour Joining Tree (Saitou and Nei, 1987) from the 363 barley accessions, using simple matching of markers, the R package APE: Analysis of Phylogenetics and Evolution (Paradis et al., 2004) was employed. The phylogenetic tree was visualised using R package phytools (Revell, 2012). Phylogenetic distances between accessions were calculated using R package 'adephylo' (Jombart et al., 2010). Principal Component Analysis (PCA) was conducted on the same set of data using R package pcaMethods (Stacklies et al., 2007) and visualised using R package ggplot2 (Wickham, 2016).

Pedigree data of the collection was gathered from various sources including historical records held at JHI, breeder supplied pedigree definitions, manuscripts and other written communications, pedigree definitions obtained from the Agriculture and Horticulture Development Board (AHDB) pocketbooks and finally the AHDB Recommended Lists app (<https://ahdb.org.uk/knowledge-library/recommended-lists-for-cereals-and-oilseeds-rl-app>). The collated data was checked for inconsistencies between data sources and finally formatted in Helium format files that can be visualised using

Helium (<https://helium.hutton.ac.uk>), where pedigree structure can be explored and additional data types uploaded and overlaid on the pedigree for visualisation and analysis.

2.4 Phenotyping the ExHIBiT collection

The ExHIBiT collection was studied during 2020 under field conditions at University College Dublin (UCD) Lyons Estate Research Farm, Ireland (53.18322, -6.31398) along with three checks, namely Golden Promise (top old cultivar in Europe), RGT Planet (top cultivated elite cultivar in Europe in 2020's) and Propino (top cultivated elite cultivar in Europe in 2010's). Checks were used to detect and correct for spatial variation across the trial blocks, ensuring that the partial replication provided an estimate of the trial error. The field was divided into 15 rows and 30 columns, which formed a grid of 50 blocks with nine (3 by 3) plots in each block. Each plot contained a primary check (RGT planet) at its centre and a secondary check (RGT planet, Golden Promise or Propino) was randomly placed around the field. In total, 52 RGT planet, 17 Propino and 16 Golden Promise plost were used. Each accession plot measured 4m by 0.45m, containing four rows of plants with 15 cm spacing between them. Accessions were grown according to local management practices in terms of sowing rate, weed and disease control, and fertiliser inputs. The sowing rate of 140.8 kg Ha⁻¹ was maintained consistently across all accessions following the recommendations of Teagasc, the Agricultural and Food Development Authority of Ireland. A full outline of the trial dates, fertilisation and weed control practices are provided in [Supplementary Table S2](#). Description of weather including temperature, relative humidity and precipitation during the growing season is presented in [Supplementary Table S3](#). During sowing, the accessions with ID number from 338 to 363 in the last

row of the field were phenotypically unreliable due to sowing equipment malfunction; hence, phenotypic data for these accessions was not collected and row-type was labelled as unknown. However, despite the lack of their phenotypic data, the accessions 338-363 were included in the genotypic analysis to investigate the overall genetic variability of the ExHIBiT collection as this information was collected prior to field trials.

A total of nine phenotypic traits were recorded for each accession ([Table 1](#)), with the timing of main stages being recorded according to the Zadoks growth stage scale ([Zadoks et al., 1974](#)). Flowering time (FLT) was recorded according to [Alqudah and Schnurbusch \(2017\)](#). During the harvest, plot samples were manually cut at ground level from one of the middle rows (linear metre) and stored in a glasshouse prior to processing. To determine the Shoot Fresh Mass (SFM), grain yield (YLD), and harvest index of each accession, the samples were threshed and cleaned using machinery from Almaco, Nevada, USA. The machinery used included the thresher model SBT (serial number 99005) and the seed cleaner model ABSC (serial number 99006). The harvest index (HI) is defined as the ratio of harvested grain to total shoot dry matter. The spikes of all accessions were photographed to create a spike image library which is available from the ExHIBiT Germinate ([Raubach et al., 2021](#)) database (<http://ics.hutton.ac.uk/germinate-exhibit>).

2.5 Power analysis and selection of ExHIBiT core collection

To reduce the number of accessions while preserving the diversity of the collection, the ExHIBiT core collection was established using a power analysis, in which the number of accessions was determined to achieve sufficient association power

TABLE 1 List of nine traits recorded during the 2020 and 2021 field trials. Includes type of trait, name of trait, method of measurement and unit of measurement.

	Trait	Abbreviation	Method of measurement	Unit
Pre-harvest traits	Tiller count	TN	Number of tillers per plant (average from three plants per plot)	–
	Flowering time	FLT	Number of days from sowing to flowering	days
	Ripening period	RIP	Number of days from sowing to ripening	days
	Height	HEI	Distance from soil surface to tip of the spike (excluding awns), averaged from four measurements	cm
Post-harvest traits	Shoot Fresh Mass	SFM	Weight of fresh shoot per linear metre	kg
	Grain Yield	YLD	Weight of grains per linear metre	g
	Harvest Index	HI	Ratio between grain yield (YLD) and shoot fresh mass (SFM)	%
Seed traits	Thousand kernel weight	TKW	Weight of 1000 kernels	g
	Protein content	PRO	Protein content of seeds after drying	%

using both genotypic and phenotypic data (which included all recorded traits from 2020 except Protein Content-PRO). Effective sample size and statistical power was computed using the R package “pwr” (Champely, 2020). The power and sample sizes were calculated under different ranges of factors, including MAF of 0.5, 0.2, 0.01. The core collection was set up by identifying the most diverse genotypes using the Core Hunter 3 software (De Beukelaer et al., 2018), in which subsets on the bases of multiple genetic and phenotypic measures, including both distance measures and allelic diversity indices were assembled. To further confirm the conservation of genetic diversity in the core collection, the results of the genetic principal components (PCs) of the whole ExHIBiT collection were compared to the PCs of the core collection. Wilcox test was performed by running the `wilcox.test` function from R package `stats` v4.2.2 (R Core Team, 2022) between the phenotypic data from 2020 of the whole ExHIBiT collection versus the core collection.

2.6 Phenotyping the ExHIBiT core collection

A total of 230 accessions comprising the ExHIBiT core collection, along with two checks (RGT planet and Golden Promise), were trialled at UCD Lyons Estate Research Farm in 2021. The experimental layout consisted of two blocks arranged in a completely randomized design. The core collection was fully replicated, and checks were randomly replicated across the field, in summary each block included a fully replicated instance of the ExHIBiT core collection and 22 replicates of each check randomly distributed. Trial dates, fertilisation and weed control practices together with weather conditions during growth season are provided in Supplementary Table S2, S3, respectively. The sowing rate of 140.8 kg Ha⁻¹ was maintained consistently across all accessions. The field trial comprised a total of 504 plots, distributed by 24 rows and 21 columns. The plot size was 7m by 0.60m, containing 5 rows of plants with a row spacing of 0.15m; thus, enabling the retrieval of agronomic data. For sowing, the Wintersteiger (A-4910 Reid, Austria machinery with the serial number 2270-4014-PDS-E and the machinery type Plotseed XL) was used. The same traits were recorded as previously described for 2020 (Table 1).

2.7 Statistical analysis of phenotypic data

An initial step of data cleaning and processing included outlier removal (for both 2020 and 2021) using Tukey’s method (Anscombe and Tukey, 1963) with outliers removed from both within and across the years according to Khodaeiaminjan et al. (2023). The data was then adjusted for the spatial variation in the field using a mixed-model analysis for each trait in each year using ASReml-R v4.0 (Butler et al., 2017) and `asremlPlus` (Brien, 2023) packages for the R statistical computing environment R v3.6.0 (R Core Team, 2022). In brief, the formula of the maximal mixed model for this analysis is:

$$y = X\beta + Zu + e;$$

where y is the vector of values of the trait analysed and β , u and e are the vectors for the fixed, random, and residual effects, respectively. The design matrices corresponding to ‘ β ’ and ‘ u ’ are denoted by X and Z , respectively. The checks, blocks, position, and genotype effects were all accounted for in this model. To identify the environmental terms that were sources of variation and needed to be included in the analysis model, variograms were examined following recommendations outlined by Gilmour et al. (1997). From these analyses, the best linear unbiased estimates (BLUEs) were obtained and used as an input for the subsequent association analysis. The heritability was calculated according to Cullis et al. (2006). Full details about the spatial correction of the field data can be found in Saade et al. (2016).

Correlation analysis was performed using the Pearson correlation with R package `Hmisc` (Harrell, 2023), and the correlation matrix figure was generated with R package `corrplot` (Wei and Simko, 2021). PCA of the phenotypic data was conducted using R package `pcaMethods` (Stacklies et al., 2007) and visualised using R package `ggplot2` (Wickham, 2016). All scripts used are available in Germinate at <http://ics.hutton.ac.uk/germinate-exhibit>.

2.8 Genome-wide association studies

GWAS was performed using the Genome Association and Prediction Integrated Tool package (GAPIT) (Wang and Zhang, 2021). Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK), a state-of-the-art multivariate model was employed for GWAS as a suitable model for smaller populations (~200) (Huang et al., 2019). To cope with population structure, kinship matrix and PCs were included in the model. The optimal number of PCs for each trait was determined using the results from the population structure analysis, and by analysing quantile-quantile (QQ) plots, created by the `qqPlot()` function in the “car” package in R (Fox and Weisberg, 2019), which are commonly used to effectively determine false positives and negative associations (Riedelsheimer et al., 2012; Kristensen et al., 2018). False discovery rate (FDR<0.05) was considered as the significant association threshold between markers and traits (Storey and Tibshirani, 2003; Storey et al., 2022), together with the Bonferroni-adjusted threshold of $\alpha=0.05$. Phenotypic distribution of significant SNPs identified in GWAS was analysed using t-test.

2.9 Linkage Disequilibrium (LD) analysis and candidate gene selection

Pairwise Linkage disequilibrium (LD) analysis was carried out according to Khodaeiaminjan et al. (2023). In short, LD-decay and LD-blocks were analysed using the ‘`Ldheatmap`’ R package (Shin et al., 2006) for each chromosome. Regions of interest for candidate genes were considered: i) genome region containing a significant marker in which flanking markers displayed strong LD ($r^2>0.5$), and neighbouring markers on either side and ii) genome region

containing significant markers outside of LD block defined by flanking marker. The allelic diversity of the SNP markers previously identified by Bustos-Korts et al. (2019) was examined by quantifying the percentage makeup of two alleles at those SNP regions.

3 Results

The ExHIBiT collection, comprising 363 barley accessions, from 22 European countries (including the former Yugoslavia) was assembled in this study. The collection specifically focused on Northern Europe (70%), with most accessions coming from the UK (30%). Four accessions originate from outside of Europe due to mislabelling in genebanks. The ExHIBiT collection includes elite cultivars (~67%), old cultivars (~18%), landraces (~14%), and accessions with unknown breeding history (~1%), representing the genetic diversity and breeding history of two-row spring barley in Europe, with the majority of accessions being released before the 90's (Figure 1). Full passport data for genotypes (based on information from JHI, the European search catalogue for plant genetic resources (EURISCO), Weise et al. (2017) and Faccini et al. (2021) is provided in Supplementary Table S1. Despite only accessions labelled as 'two-row' types being selected for this study, 16 out of the 363 accessions in the ExHIBiT were identified as six-row upon phenotypic analysis. These accessions remained as part of the ExHIBiT collection but were not considered for inclusion in the core collection and not included in subsequent studies. The entire collection was genotyped using the 50k iSelect SNP array (Bayer et al., 2017). In total 35,968 markers were mapped to a physical position on the "Morex" V3 genome sequence (Mascher et al., 2021). Full 50k SNP array data for the collection can be found on

the germinate website (<http://ics.hutton.ac.uk/germinate-exhibit>). After MAF and missing data filtering 26,585 robust markers remained for genotypic analysis. This data set is available at EMBL-EBI under accession number PRJEB67728 (<https://www.ebi.ac.uk/eva/?eva-study=PRJEB67728>) (Supplementary Figure S1).

3.1 Phylogenetic relationship & population structure of the ExHIBiT collection

The genetic structure of the ExHIBiT collection showed three main groups (Figure 2A), which can be further divided into six smaller sub-groups, as shown by the ΔK peaks at K3 and K6 (Figure 2B). The fixation index (F_{st}) showed significant divergence within the groups, with values of 0.32, 0.48 and 0.35 for three groups of K3.1, K3.2, and K3.3 respectively. K3.1, K3.2, K3.3 contained 151, 122 and 64 accessions respectively, with the remaining 26 accessions being classified as admixture. These groups can also be distinguished in both PC & phylogenetic trees (Figures 2C, D). The division of these three groups was investigated using geographical data in three main regions of origin (UK and Ireland, Northern and Southern Europe) (Figures 3A, B). However, the results showed that the population structure was not influenced by geographical origin. To further explore the ExHIBiT population structure, breeding history was investigated, and the results showed that in K3.1, 56% of the accessions were released post-1990 and 43% pre-1990. While in K3.2, 86% of the accessions had been released pre-1990 and in K3.3, 63% of accessions were landraces according to known breeding history. In these six groups, the K3.3 group split into K6.1 and K6.2 where K6.1 contains the two-row landraces and K6.2 all the six-row

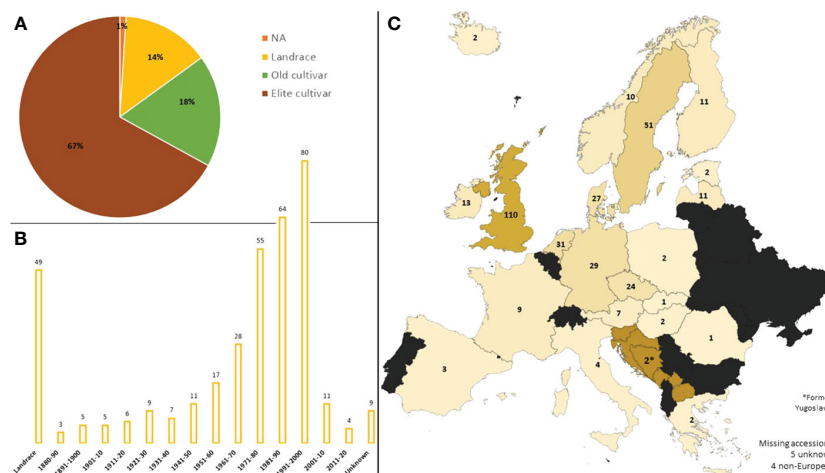


FIGURE 1

The Background of the ExHIBiT collection. (A) Breeding history of accessions in the ExHIBiT collection, divided into landraces, old cultivars (released before 1960) and elite cultivars (released after 1960). (B) The decade of release of accessions in the ExHIBiT collection, showing the number of accessions released in each decade in the collection. Landraces are separated as they do not have a specific year of release. (C) Country of origin of the accessions in the ExHIBiT collection. In the case of landraces, this is the country where the accession was found. In the case of the old and elite cultivars, this is the location of the institute or company where breeding took place. The origin of one accession is unknown and four lines originate from outside of Europe.

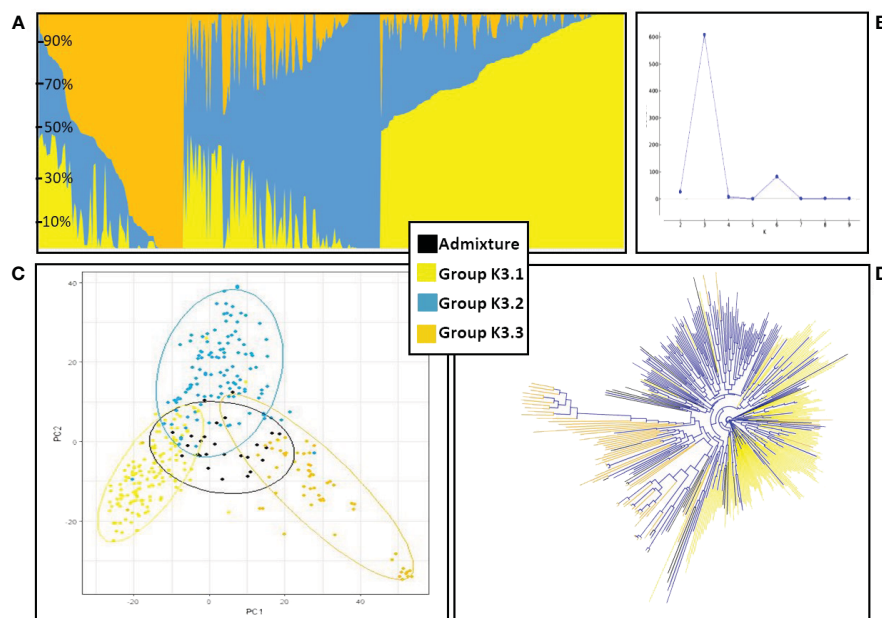


FIGURE 2

Population structure and phylogenetic analysis of ExHIBiT collection. **(A)** Results of structure analysis for $K=3$, each vertical line represents an accession, y-axis shows percentage content of each accession to the three different groups. **(B)** Graph of Delta K values for $K=2$ to $K=10$. Maximum DeltaK was reached at $K=3$ and another peak at $K=6$. **(C)** Principal Component Analysis based on 50K genotypic data coloured according to the three groups identified in population structure analysis. **(D)** phylogenetic tree coloured according to the three groups identified in population structure analysis.

landraces (Supplementary Figure S4). In the population structure analysis, the accessions categorised as “admixture” were predominantly elite cultivars, comprising approximately 69% of the group. The remaining 31% of this category was divided into 12% landraces and 19% old cultivars. This distribution closely mirrors the overall composition of the collection.

In principal component analysis of genotypic data, PC1 and PC2 explained 8% and 6% of variation respectively. Clustering by

PCA and phylogenetic analysis (Supplementary Figure S2) results are consistent with a population structure of three groups representing barley breeding history (Figures 4A, B), with landraces, pre-1990 and post-1990 accessions being distinct. PCA confirms the distinction between the different row types with two- and six-row accessions clearly clustering away from each other (Supplementary Figure S3), and the six-row accessions perfectly overlapping with cluster K6.2 when analysing the PCA results

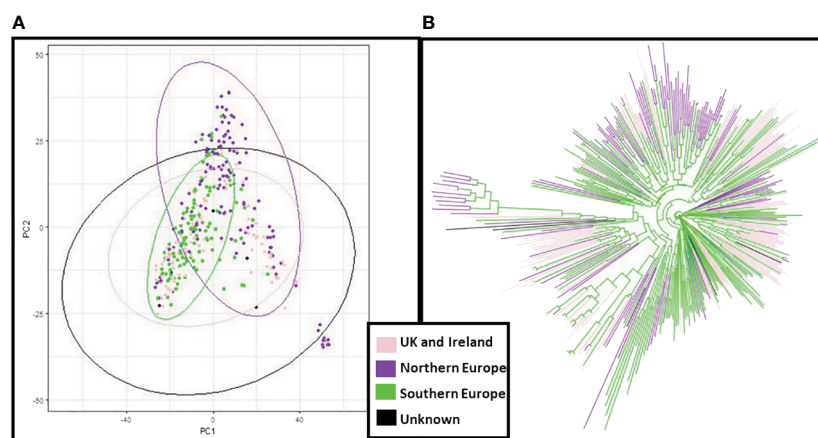


FIGURE 3

Principal Component Analysis (PCA) and neighbour joining phylogenetic analysis based on the region of origin of the ExHIBiT collection. **(A)** PCA coloured by the region of origin **(B)** phylogenetic tree coloured by region of origin. In pink shown accessions from the UK and Ireland, in purple accessions from Northern Europe, in green accessions from Southern Europe and in black accessions whose region of origin is unknown.

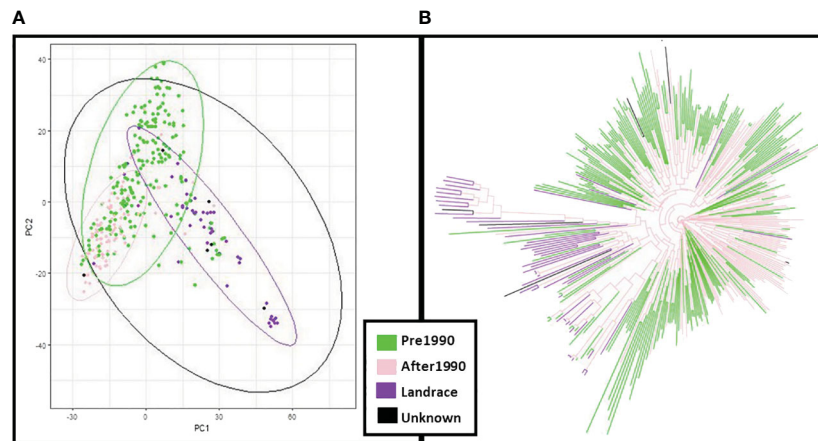


FIGURE 4

Principal Component Analysis (PCA) and neighbour joining phylogenetic analysis based on the year of release of the ExHIBiT collection **(A)** PCA coloured by year of release. **(B)** phylogenetic tree coloured by year of release. In green shown accessions released before 1990, in pink accessions released after 1990, in purple accessions landraces and in black accessions whose year of release is unknown.

according to population structure with six sub-groups (Supplementary Figure S4). The splitting of K3.2 into subgroups K6.3 and K6.4, and K3.3 into K6.5 and K6.6, is unclear, with no distinct overlap to either geographical origin or year of release.

To assess how effectively the ExHIBiT collection captures the genetic diversity of European two-row spring barley, the genotypic diversity of the collection was compared to wider IPK Barley Core 1000 collection (Milner et al., 2019). The IPK collection reflects worldwide barley diversity and contains a large number of European accessions. Principle component analysis results

combining the two collections revealed that ExHIBiT distinctly clusters with the European two-row spring barley accessions within the IPK collection (Figure 5A) with PC1 explaining 13% and PC2 8% of the variation in the data (Figure 5B).

The pedigree of lines within the ExHIBiT collection overlap with the pedigree data assembled from historical records held at JHI and supplemented with breeder declared pedigrees from AHDB and genotypes maintained at JHI for 1,847 European barley varieties (<https://helium.hutton.ac.uk>). The pedigree data for ExHIBiT can be visualised using the Helium pedigree visualisation platform

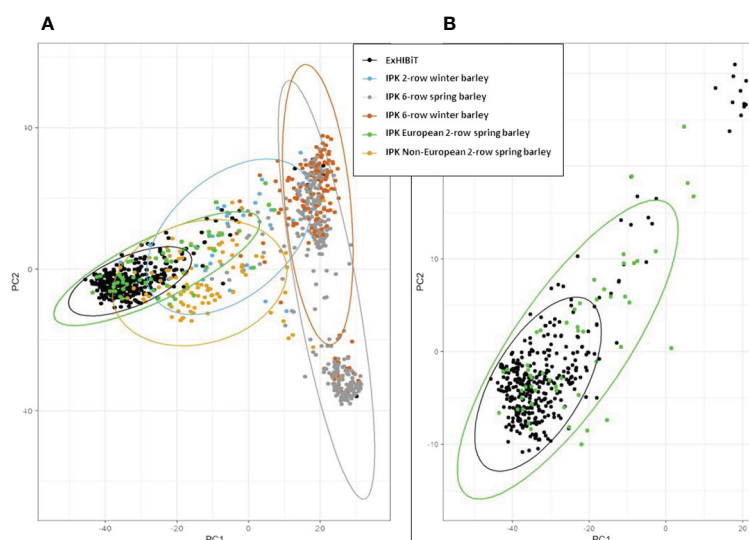


FIGURE 5

Principal Component Analysis (PCA) of ExHIBiT collection against IPK Barley Core 1000 collection. **(A)** PCA of the ExHIBiT collection (in black) clustered against the IPK Barley Core 1000 collection with European 2-row spring barley accessions from the IPK collection shown in green, non-European two-row spring barley accessions in yellow, six-row spring barley accessions in orange, two-row winter barley accessions in blue and six-row winter barley in grey. **(B)** PCA of ExHIBiT collection (in black) clustered with only the European two-row spring barley accession from the IPK Barley Core 1000 shown in green.

(Shaw et al., 2014; <https://helium.hutton.ac.uk/#/pedigree/exhibit>) where along with pedigree definitions, additional characterization data related to the ExHIBiT collection is openly available.

3.2 Phenotypic variation of the ExHIBiT collection

Nine agronomic traits were measured (Table 1) for all accessions in the ExHIBiT collection, in 2020. The collection showed considerable diversity in the field, with plant height (HEI) varying from 69.3cm (Kria) to 122.9cm (Irish Goldthorpe) with an average HEI of 90.6 cm. The YLD per linear metre varied from 79.8g (Craigs Triumph) to 155g (Primus) with an average of 120g (Supplementary Table S5). In 2020 flowering time (FLT) and ripening period (RIP) data collection was limited by weather conditions and lockdown restrictions associated with COVID19, resulting in unavoidable missing information. RIP particularly suffered from this resulting in data with nearly no variation, therefore this trait was excluded from any further analysis (Supplementary Table S4). The results indicate that FLT spanned

from 66 to 83 days from sowing to flowering with an average of 70 days. After a visual analysis of the histograms (Figure 6), all traits, except FLT and RIP, appear to be normally distributed. To further illustrate the phenotypic diversity of the collection, a photo library showing the variability in shape and size of the spikes, has been included with the genotypic data on the Germinate database (<http://ics.hutton.ac.uk/germinate-exhibit>), and is exemplified in Supplementary Figure S5.

3.3 Construction and phenotypic characterization of the ExHIBiT core collection

To assemble a representative ExHIBiT core population, a power analysis was undertaken using phenotypic and genotypic data from the 363 accessions. Power analysis revealed that genetic effect would be detectable with 230 accessions at a power of 0.8. Regarding phenotypic data from 2020, it was observed that there was no statistical difference between the entire ExHIBiT and core collection (363 vs. 230 accessions) with *p*-values ranging from 0.96 to 0.20.

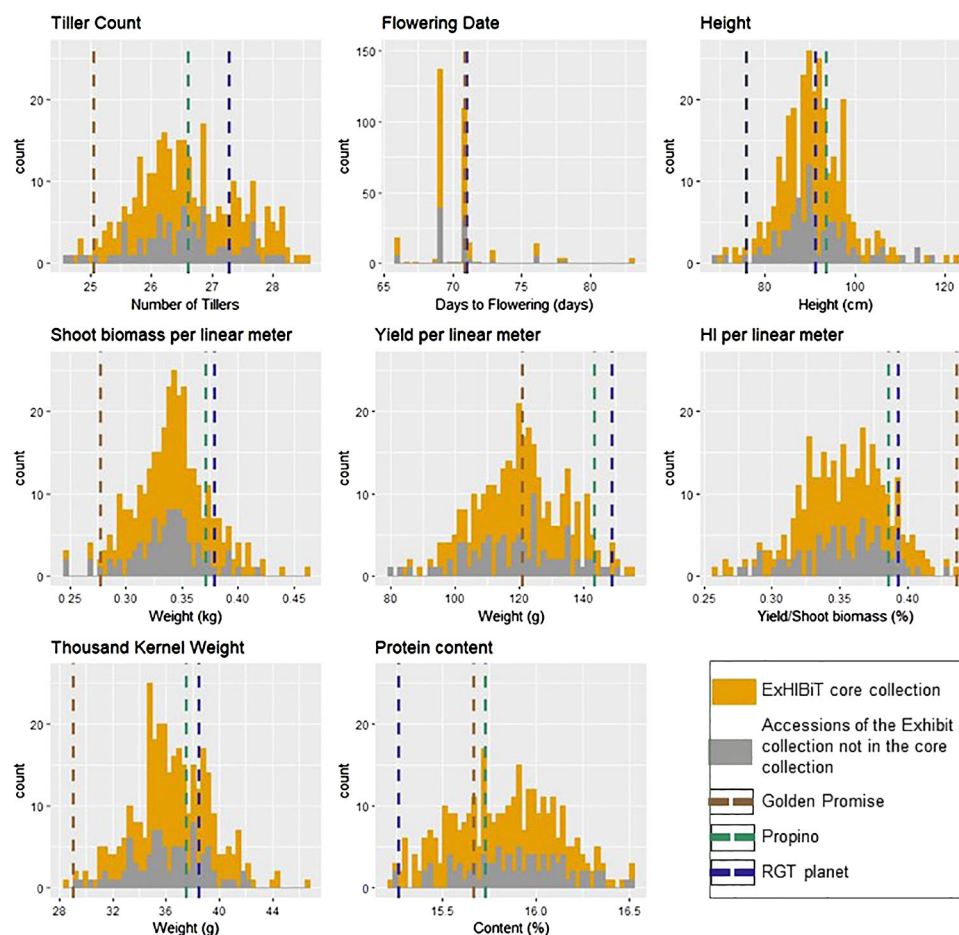


FIGURE 6

Histogram presenting the diversity difference between two sets of accessions in the 2020 field trials. The ExHIBiT core collection, consisting of 230 accessions, is represented in yellow, while the ExHIBiT collection accessions not included in the core collection, totalling 123 accessions, are shown in grey. The dashed lines represent the average values of the checks: RGT planet (in black), Propino (in green), and Golden Promise (in brown).

The range of phenotypic values between the core collection and the entire ExHIBiT collection is highly similar (Figure 6). PCA results in a near perfect overlap between the whole and core collection in terms of genetic diversity, with the exception of the six-row barley accessions which were not included in core collection (Supplementary Figure S6). The core collection contains accessions from 20 European countries, with around 70% coming from Northern Europe. It includes elite cultivars (~74%), old cultivars (~15%), and landraces (~10%), and accessions with unknown breeding history (1%) (Supplementary Figure S7). The makeup of the core collection highly resembles that of the ExHIBiT collection.

The core collection was sown in April 2021 at UCD Lyons Estate Research Farm and phenotyped during the growing season

until harvesting in August. The same ten agronomic traits were recorded as in 2020, and similar phenotypic diversity was observed in both years with the full dataset presented in Supplementary Table S6 and correlation in Supplementary Table S8. HEI varied from 80.3cm (Wren) to 150.7cm (Chevalier Tystofte) with an average of 105.5cm. YLD per linear meter varied from 75g (Irish Goldthorpe) to 117g (Canasta) (Table 2).

PCA results for phenotypic data from 2021 were examined to compare to PCA results from the genotypic data and verify similar patterns in the clustering (Figure 7). The results show that in the PCA of 2021 phenotypic data, PC1 explained 7% of variation and PC2 3% of variation. PCA obtained from the phenotypic and genotypic data in the core collection is consistent with PCA patterns of the whole ExHIBiT collection. These results

TABLE 2 Agronomic data from ExHIBiT core collection field traits in 2021. Shows minimum, maximum, average and standard deviation for all nine collected traits¹.

	Trait	Unit	Minimum	Maximum	Average	Standard deviation
Pre-harvest traits	TN		6.57	9.71	7.35	0.411
	FLT	days	53.86	79.97	65.75	3.91
	RIP	days	118.94	125.83	122.98	1.33
	HEI	cm	80.3	150.65	105.51	13.05
Post-harvest traits	SFM	kg	0.2308	0.3415	0.27	0.01895
	YLD	g	75.04	116.7	97.86	8.38
	HI	%	0.259	0.458	0.356	0.0276
Seed traits	TKW	g	41.26	58.89	50.25	3.048
	PRO	%	11.8	17.7	13.95	0.909

¹See Table 1 for list of abbreviations.

TABLE 3 Genome-Wide Association Studies results for nine traits in 230 ExHIBiT core accessions using BLINK.

Trait	MAF	PCs	Associated SNP	Chromosome	Position	P.value	LD block
FLT	0.38938	3	JHI-Hv50k-2016-383902	6H	37,699,708	9.30E-08	LD Block H6 - 87
FLT	0.0996	3	JHI-Hv50k-2016-460460	7H	43,608,409	2.71E-07	LD Block H7 - 133
HEI	0.1659	3	JHI-Hv50k-2016-168497	3H	174,958,294	1.51E-10	LD Block H3 - 155
HEI	0.3362	3	JHI-Hv50k-2016-205137	3H	563,141,095	1.15E-09	LD Block H3 - 296
HEI	0.0568	3	JHI-Hv50k-2016-453491	7H	25,872,760	2.16E-07	LD Block H7 - 85
HEI	0.1288	3	JHI-Hv50k-2016-227194	4H	3,218,930	2.99E-06	LD Block H4 - 17
PRO	0.0526	4	JHI-Hv50k-2016-102790	2H	546,982,003	1.03E-07	LD Block H2 - 224
PRO	0.0614	4	JHI-Hv50k-2016-283903	5H	16,822,952	2.94E-08	LD Block H5 - 77
PRO	0.0746	4	JHI-Hv50k-2016-504314	7H	598,982,074	2.67E-10	LD Block H7 - 287
TKW	0.1288	4	JHI-Hv50k-2016-57915	1H	516,436,801	8.88E-08	JHI.Hv50k.2016.57915
TKW	0.1900	4	JHI-Hv50k-2016-117361	2H	610,739,002	1.01E-06	LD Block H2 - 287
TKW	0.0917	4	JHI-Hv50k-2016-200892	3H	545,834,674	2.91E-12	LD Block H3 - 270
TKW	0.1572	4	JHI-Hv50k-2016-358877	5H	573,661,301	1.52E-06	LD Block H5 - 492

The table includes the list of significant associations detected, the identified markers for each trait, the optimum number of PCs used for GWAS, p-values and, MAF and LD blocks are given. The positions are based on “Morex” V3 (Mascher et al., 2021).

demonstrate that the region of geographical origin has no visible effect on clustering of the core collection while the year of release shows three distinct groups; landraces and accessions released before and after 1990.

Relationships between the traits and year of release were explored using correlation analysis. Correlations between years varied from 0.44 for HEI to 0.045 for SFM. Positive correlation between HEI and PRO (0.45 in 2021 and 0.16 in 2020), FLT (0.33 in 2021 and 0.36 in 2020) and SFM (0.28 in 2021 and 0.3 in 2020) were observed. Positive correlation between SFM and FLT (0.31 in 2021 and 0.14 in 2022) and YLD (0.6 in 2021 and 0.48 in 2020) were observed. Year of release negatively correlated with HEI and FLT, and positively correlated with HI and YLD (Supplementary Figure S8).

3.4 Genome-wide association studies

To confirm the value of the ExHIBiT core collection for genetic analyses, GWAS was performed on all phenotypic traits collected during the 2021 field trial (Table 3). The 50K SNP data was re-filtered on the core collection, resulting in 24,876 high quality markers for the 230 accessions.

GAPIT was used to test BLINK models. To account for the population structure, PC3 was initially tested due to the identification of three groups in the population structure analysis. After QQ plot analysis, PC4 was used as a fixed effect in BLINK whenever the results indicated that PC3 was not suitable. The QQ plots for all traits indicated that either PC3 or PC4 was appropriate, eliminating the need to run BLINK with higher PC numbers. Significant SNPs were identified for several traits: FLT, HEI, Thousand Kernel Weight (TKW), and PRO. GWAS results with Manhattan plots and phenotypic distribution of significant SNPs for HEI, TKW and PRO can be found in Supplementary Figure S9 while the complete list of significant SNPs is presented in Table 3.

For validation purposes, FLT was selected as an example because it is a very well described trait in barley (Alqudah et al., 2014; Bustos-Korts et al., 2019; Fernández-Calleja et al., 2021) and presented the highest heritability (0.912).

3.5 Genetic regions underlying FLT and diversity analysis of flowering genes

The GWAS results identified two markers to be significantly associated with FLT on chromosomes 6H (*JHI-Hv50k-2016-383902* located at chr6: 37,699,708) and 7H (*JHI-Hv50k-2016-460460* located at chr7: 43,608,409) (Figure 8A). Significant *p*-values for associations between markers and phenotypic traits were determined using the FDR (Storey and Tibshirani, 2003; Storey et al., 2022) and the Bonferroni-adjusted threshold of $\alpha=0.05$, which corresponded to a logarithm of odds (LOD) score of 5.69. Pairwise marker LD matrices and LD-decay were estimated for each of the chromosomes separately based on the SNP data showing lowest *p*-values. The LD regions expanded the genomic regions of interest to chr6:36,884,125 - 115,649,880 bp and chr7:40,532,291 - 44,676,872 bp. The significant SNPs identified on 6H overlapped with two previously identified flowering genes: *HvZTLb* (Bustos-Korts et al., 2019) and *Eam7* (Stracke and Borrner, 1998). Additionally, other flowering genes *HvZTLa*, *Vrn-H3* (chr7:39,680,381), *HvCO8* (chr7:50,187,671) and *HvLHY* were found to be in close proximity to the significant SNP on 7H, although they were not located within the same LD block (Alqudah et al., 2014; Bustos-Korts et al., 2019). Analysis of the phenotypic distribution of the two significant SNP markers shows statistical differences between the alleles among the population (Figure 8B).

To verify the genetic diversity of the ExHIBiT core collection at previously identified flowering genes, 11 flowering genes and 24 related SNP markers were examined following the findings of

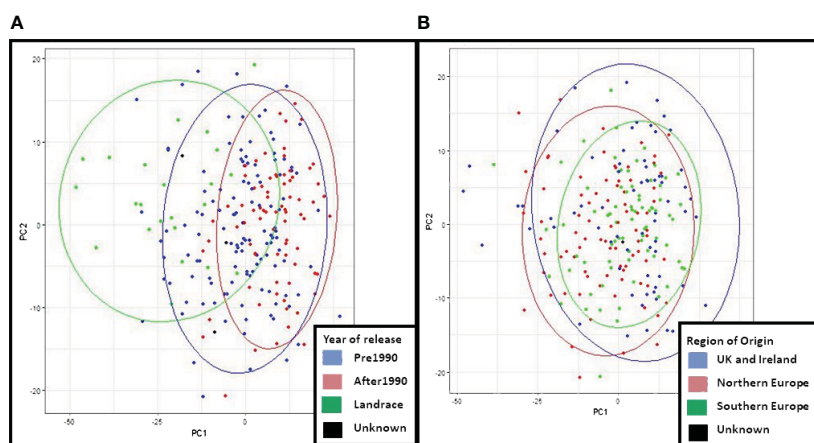


FIGURE 7

Principal Component Analysis (PCA) of 2021 phenotypic data for the ExHIBiT core collection. (A) PCA coloured by year of release of accessions. In green are shown landraces, in blue pre-1990 accessions and in red accession released after 1990. (B) PCA coloured by region of origin, in red are shown accessions from Northern Europe, in green Southern European accessions and in blue accessions from UK and Ireland.

Bustos-Korts et al. (2019). Out of these 24 alleles, two were filtered out due to a high number of missing data. The results show that eight SNP markers within four genes have been fixed in the ExHIBiT core collection, including the *Ppd-H1* gene. The remaining 14 SNP markers and seven flowering genes showed allelic diversity in the collection. Ratio of alleles at these SNPs ranged from a near equal allelic make up (49% vs. 51% at SNP JHI-Hv50k-2016-494493 associated with gene *HvZTLA*) to one allele being predominant at SNP JHI-Hv50k-2016-411622 associated with gene *HvCO2* (13% vs. 87%).

4 Discussion

Here the ExHIBiT collection is introduced for the first time. The collection comprises 363 European spring barley accessions and was used to assemble the core ExHIBiT collection comprising 230 two-row accessions. This collection was created to reflect European barley diversity, and to be of use in future screening and association mapping studies to address the molecular and physiological mechanism underlying biotic and abiotic stresses.

4.1 Diversity & population structure of the ExHIBiT collection

Barley has a large genetic diversity, with landraces and older cultivars representing great untapped diversity; which, when bred with modern cultivars can create high yielding and resilient accessions (Ceccarelli et al., 1995; Lakew et al., 1997; Kumar et al., 2020). Several barley collections have been created and introduced previously, but their majority are either global with only a subset of European lines (Liu et al., 2000; Milner et al., 2019), or focus on specific parts of Europe (Igartua et al., 1998; Milotova et al., 2008). Northern European germplasm tends to be under-represented in diversity panels compared to Southern accessions (Pasam et al., 2014; Selçuk et al., 2015). Therefore, there are still aspects of European and in particular Northern European barley diversity that have not been fully explored. The ExHIBiT collection contains many historically important malting varieties from Northern Europe, ranging from landraces to elite cultivars and reflecting the diverse breeding practices used throughout the history of Northern Europe (e.g., Chevalier, RGT Planet, Kenia, Quench, Proctor, Carlsberg (Plarr et al., 1963; Hagenblad & Leino, 2022; Nejat, 2022).

To ensure that the collection's skew towards Northern European accessions did not introduce a bias and that the genetic diversity of European barley was successfully captured, the ExHIBiT collection was compared to the IPK Barley Core 1000 collection (Milner et al., 2019). IPK Barley Core 1000 is a large collection that aims to capture global barley diversity, including Europe. PCA results of the ExHIBiT collection together with the IPK collection showed that the ExHIBiT accessions fit well with other two-row spring barley accessions present in IPK collection. Specifically, ExHIBiT accessions showed the same diversity to the European two-row spring barley accessions in the IPK collection (Figure 5). This result indicates that, the ExHIBiT collection has successfully

captured European barley diversity despite its focus on Northern European accessions.

Based on genomic data, population structure of the ExHIBiT collection appears to be mainly defined by the year of release, while the region of origin plays a minimal role. Similar patterns, found in the PCA of the phenotypic data from 2021, further confirms this population structure. This is consistent with results of previous research with non-landrace accessions of European origin (Malysheva-Otto et al., 2006, 2007; Tondelli et al., 2013; Brbaklić et al., 2021). This result contrasts with research focusing on global collections where population structure tends to be defined by country/region of origin (Jones et al., 2011; Muñoz-Amatriáin et al., 2014; Pasam et al., 2014; Russell et al., 2016).

The population structure of the ExHIBiT collection was found to consist of three main groups, which can be broadly described as landraces, pre-1990's cultivars and post-1990's released cultivars (Figure 2). The ExHIBiT population structure is in good agreement with previous studies focusing on European two-row spring barley. Saade et al. (2020), found that the collection of 377 European two-row spring barley had a population structure made up of three groups, whereas Tondelli et al. (2013), using a collection of 216 European two-row spring barley, found a population structure made up of only two groups. However, the latter collection from Tondelli et al. (2013), does not contain landraces and the two identified groups could be broadly described as pre- and post-1990 released cultivars. The population structure results are consistent with breeding practices in Europe because of the interchange of germplasm between breeding programs, creating a diverse germplasm without major division in the population structure (Rostoks et al., 2006), with most contemporary barley cultivars having four preeminent accessions in their pedigree, these pedigrees being Spratt Archer (from Ireland), Gull (from Sweden), Binder (from Moravia) and Isaria (from Bavaria) (Fischbeck, 1992, 2003, Russell et al., 2000). These results are consistent with the findings of Schreiber et al. (2024), reporting that many European elite barley cultivars are descendant of a small number of "founder" genotypes, namely Kenia, Maja and Gull. The pedigree of the ExHIBiT collection was visualized using the Helium pedigree visualisation platform (Shaw et al., 2014; <https://helium.hutton.ac.uk/#/pedigree/exhibit>) where along with pedigree definitions, additional characterization data related to the ExHIBiT collection is openly available and can be explored, visualised, and exported should additional downstream analysis be required. These example datasets not only shows the ExHIBiT collection, but also shows where this collection sits in relation to other European barley varieties and how its constituents are related to other varietal material.

Despite the importance of six-row spring barley, the ExHIBiT collection exclusively focuses on two-row spring barley as this is the main type used by the malting industry making it of premium value (Newton et al., 2011; Hertrich, 2013). A small number of the accessions in the collection that were labelled as two-row in their passport data, upon planting were identified as six-row barley. Mislabelling and duplication are the most frequent problems within genebanks, this is due to the large number of accessions maintained (Mascher et al., 2019; Dreiseitl and Nesvadba, 2021).

Although the 16 six-row accessions were considered as part of the ExHIBiT collection, they were not considered for inclusion in the core collection. The inclusion of the six-row accessions further supports the idea that the population structure is defined by year of release instead of row-type as breeding history appears to be the most dominant factor. The population structure of the remaining subgroups is not fully explained by row number nor year of release, implying some influence from region of origin or other factors. Malysheva-Otto et al. (2006), found that accessions from Northern Europe and the Soviet Union tend to form subgroups. However, due to the small representation of accessions from some countries and regions, no further assumptions on the reasons behind the structure of the subgroups can be made.

4.2 Core collection construction and phenotyping

The core collection with 230 accessions was created to be used in genetic and association mapping studies, effectively reducing the cost and time investment required, while still preserving the diversity of the ExHIBiT collection. One of the main objectives in the establishment of the ExHIBiT core collection was to ensure that the population has sufficient accession numbers with maximum diversity for genetic studies. To create the core collection, a distance-based method called Core Hunter was applied, relying on both genetic and phenotypic data. The choice of appropriate method depended on the purpose of the study (i.e., to capture as much diversity as possible with the smallest number of accessions), in addition to the computational speed of the method, and the information required (i.e., sample size). The distance-based method was ideal as its main purpose was maximising the combination of allelic diversity at genome level, which is a key factor for breeding programs (Leroy et al., 2014). The phenotypic and genotypic diversity in the core and ExHIBiT collection were compared and results showed that the diversity of the whole collection was preserved (Figures 6 and 7).

Field data for both ExHIBiT (2020) and core collection (2021) assessment, showed variation in all phenotypic traits, including HEI, FLT, SFM, YLD, TKW and PRO. The checks; RGT Planet, Propino (modern malting barley) and Golden Promise (prominent malting barley in the 1960s). In both years several accessions from the ExHIBiT core collection outperformed the checks in terms of YLD, SFM, HI and PRO. Specifically, Clansman (ID 220), Drost (ID 88), Ladik (ID 253), Primus (ID 260) and Tyne (ID 205) accessions consistently showed superior performance compared to the checks in terms of YLD and HI. Clansman germplasm exhibited exceptional agronomical qualities with consistently higher yields than RGT Planet. Four of these accessions are from Northern Europe (UK, Denmark or Sweden). These results indicate that some of these accessions have a great potential in future breeding programs. However, the ExHIBiT material has not been fully characterised, for biotic and abiotic stresses. The correlation between agronomic performance and year of release is consistent with knowledge of barley breeding priorities, which through time have favoured shorter and early flowering type plants with higher YLD and HI (Monteagudo et al., 2019; Brbaklić et al., 2021).

4.3 Association mapping

To confirm the effectiveness of the ExHIBiT collection for future genetic studies, association mapping was performed but not with the aim of identifying new QTLs. Similar validation analysis has been previously carried out by Wang et al. (2021). Due to its high heritability and extensive knowledge, FLT was selected as the validation trait, leading to the identification of significant markers on chromosome 6H and 7H. Previous studies have located several flowering genes on these chromosomes (Alqudah et al., 2014; Bustos-Korts et al., 2019; He et al., 2019). The identified significant SNP marker on the chromosome 6H overlaps with the previously identified flowering gene *HvZTLb*

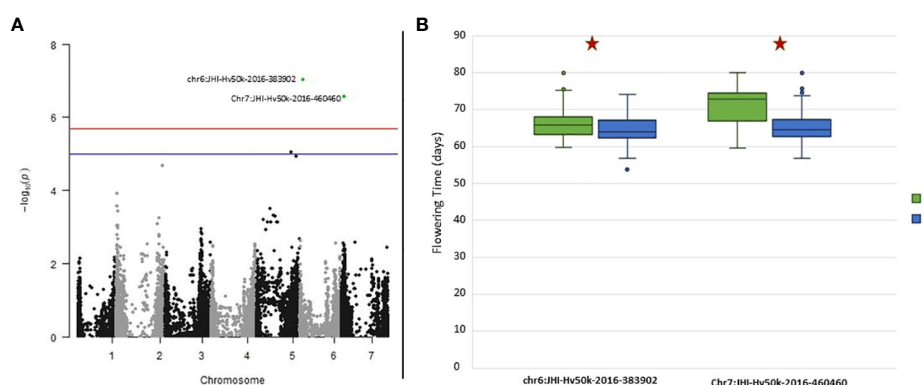


FIGURE 8

Genome-Wide Association Studies (GWAS) results for Flowering Time for the ExHIBiT core collection (A) Manhattan plot of GWAS analysis using BLINK with three principal components (PCs) in 2021 flowering time. In green are highlighted significant markers by Bonferroni correction (logarithm of odds (LOD) score of 5.69). In blue is the suggestive line (LOD score of 5) and in red the genome-wide significance line (LOD score of 5.5). (B) Boxplot showing phenotypic distribution of flowering time of two significant single nucleotide polymorphisms (SNPs) identified by GWAS with the BLINK model as observed on (A) results highlighted in green. Red star shows significance according to t-test (p-value < 0.05).

(Bustos-Korts et al., 2019). Furthermore, well-known genes *HvZTLA*, *Vrn-H3*, *HvCO8*, and *HvLHY* are in close proximity to the significant SNP markers on chromosome 7H, although they are not situated within the same LD block (Alqudah et al., 2014; Bustos-Korts et al., 2019). The significant marker identified on chromosome 6 (JHI-Hv50k-2016-383902) is also located in close proximity to previously identified QTL in chr6:13,136,770 (Alqudah et al., 2014) that might underlie the earlier described *Eam7* (Stracke and Borrner, 1998). However, some of the most well-described flowering genes including the photoperiod response gene (*Ppd-H1*) and vernalization gene *VRN-H1* (Turner et al., 2005), were not identified in GWAS. This could be explained by previous research suggesting that Northern European barley (which makes up 70% of the ExHIBiT collection) is quite homogenous in terms of flowering genes particularly *Ppd-H1* (Aslan et al., 2015). This observation was confirmed by examining the diversity at SNP markers associated with flowering genes from previous research (Bustos-Korts et al., 2019).

5 Conclusion

Barley is an essential crop for food security and has a large genetic diversity, including landraces, old and elite cultivars. In Europe, barley has a high market value with most of its use and growth due to malt processing and breweries. Landraces and old cultivars, which pre-date the Green Revolution represent a great, yet untapped diversity. When bred with elite cultivars, it is possible that high yielding and resilient accessions can be generated. To utilise old accessions in breeding, first their diversity must be explored. The ExHIBiT collection was created to reflect two-row spring European barley diversity, and to be of use in future screening studies and association mapping studies. The ExHIBiT collection provides a better understanding of the genetics of European heritage barley, contributing to improve barley yield, stress resistance, and to promote sustainable barley production in Northern European climates. The public availability of this new, and fully characterised, European heritage collection that contains historically important malting varieties will be useful for breeders, geneticists, physiologists and pathologists around the world, providing a valuable resource for a flourishing malting industry. The 230 ExHIBiT core collection is manageable for field studies and can contribute to the development of barley germplasm as well as to the identification of genomic regions associated with traits of economic importance.

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: PRJEB67728 (ENA; <https://www.ebi.ac.uk/ena/browser/view/PRJEB67728>) and Germinate (<http://ics.hutton.ac.uk/germinate-exhibit>).

Author contributions

VB: performed field experiments and data collection, analysed and validated genotypic and phenotypic data, performed GWAS and participated in writing of the original draft. NAT: performed field experiments and data collection, analysed and validated genotypic and phenotypic data, performed the power analysis and participated in writing of the original draft. PL: performed field experiments and data collection. GG: performed agronomic design for the field and performed field experiments and data collection. TD: performed field experiments and data collection. JH: performed field experiments and data collection. MH: performed agronomic design for the field. LR: assembled the ExHIBiT collection. KH: assembled the ExHIBiT collection and performed the power analysis. MM: performed genotyping and provided genomic data. PS: assembled pedigree data and created the Germinate databases associated with this work and participated in writing of the original draft. SR: assembled pedigree data and created the Germinate databases associated with this work and participated in writing of the original draft. KM: performed agronomic design for the field. JR assembled the ExHIBiT collection and performed genotyping and provided genomic data. RW: assembled the ExHIBiT collection. MK: analysed and validated genotypic and phenotypic data, performed GWAS, and participated in the writing of the original draft and editing of the manuscript. SN: designed and supervised the project, assembled the ExHIBiT collection, performed field experiments and data collection, analysed and validated genotypic and phenotypic data and reviewed and edited the manuscript. All authors contributed to the review and editing of the manuscript.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research reported in this publication was supported by Science Foundation Ireland centre by the SFI President of Ireland Future Research Leaders to SN under Grant No. 18/FRL/6197.

Acknowledgments

Seed material was kindly provided by the James Hutton Institute, Department of Agriculture Food and Marine (DAFM) and NordGen germplasm bank. We thank Cara Mac Aodháin from DAFM for all the info related with Irish lines held by DAFM. We thank the UCD Lyons Research farm and staff for their technical support, namely Eddie Jordan, Eugene Brennan and Ian Keating as well as other CSI-Dublin lab members who helped with the harvesting: Jason Walsh, Katie O'Dea. We thank Dr. Julio Isidro Sánchez for input on initial field discussions. We thank Luke Daly for proof-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.2024.1268847/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Flow diagram showing genotypic and phenotypic data analysis steps. Single nucleotide polymorphisms (SNP) marker data is available for 363 accessions and 35,968 markers were mapped to 'Morex' V3. After quality control (filtering for minor allele frequency of >0.05 and 0.1 individual SNP missing) 9,383 SNP markers were removed. Phenotypic data was lost for 26 genotypes due to unreliable field data and lack of germination, and another 13 genotypes were identified as six-row barley. Genotypic data (26,585 markers) were used for population structure, Principal Component Analysis (PCA) and neighbour joining phylogenetic tree construction. Phenotypic data was used for core selection and genome-wide association study.

SUPPLEMENTARY FIGURE 2

Phylogenetic tree of the ExHIBiT collection. Neighbour Joining Tree of the 363 accessions in the ExHIBiT collection using 26,585 SNP markers. Bootstrap values are displayed on the branches.

SUPPLEMENTARY FIGURE 3

Examination of the effect of row type on Principal Component Analysis (PCA) and neighbour joining phylogenetic tree of the ExHIBiT collection (A) PCA coloured by row type. (B) phylogenetic tree coloured by row type. In purple are shown two-row barley accessions, green represents the six-row barley accessions and pink is used for accessions with unknown row type.

SUPPLEMENTARY FIGURE 4

Examination of the effect of six subgroups identified in population structure analysis on Principal Component Analysis (PCA) and neighbour joining

phylogenetic tree of the ExHIBiT collection (A) PCA coloured by six subgroups. (B) phylogenetic tree coloured by six subgroups.

SUPPLEMENTARY FIGURE 5

Diversity panel of ExHIBiT spikes. Selection of spikes from the ExHIBiT spike library to represent the range of phenotypes including the variation in sizes and shapes. Accessions from left to right: Chevalier Tystofte 2, Wisa, Cocktail, Kinnan, Karat, Thuringia, Athos, Alis, Draught, and Beavans 35.

SUPPLEMENTARY FIGURE 6

Comparison of Principal Component Analysis between the whole ExHIBiT collection and the ExHIBiT core collection. The ExHIBiT core collection is shown in blue and the rest of the ExHIBiT collection in red. The range of values in both data sets have similar values confirming the preservation of genotypic diversity in the selection of the core collection.

SUPPLEMENTARY FIGURE 7

The Background of the ExHIBiT core collection. (A) Breeding history of accessions in the ExHIBiT core collection, divided into landraces, old cultivars (released before 1960) and elite cultivars (released after 1960). (B) The decade of release of accessions in the ExHIBiT collection, showing the number of accessions released in each decade in the collection. Landraces are separated as they do not have a specific year of release. (C) Country of origin of the accessions in the ExHIBiT core collection. In the case of landraces, this is the country where the accession was found. In the case of the old and elite cultivars, this is the location of the institute or company where breeding took place. The origin of one accession is unknown and four lines originate from outside of Europe.

SUPPLEMENTARY FIGURE 8

Correlation matrix between phenotypic data for ExHIBiT core collection (from 2021) and ExHIBiT collection (from 2020), both within and between the years. Blue shows positive correlation and red negative correlation. Trait name followed by 20 denotes data from 2020 field trial and trait name followed by 21 denotes data from 2021 field trial.

SUPPLEMENTARY FIGURE 9

Genome-Wide Association Studies (GWAS) mapping results for Plant Height, Thousand Kernel Weight and Protein content for the ExHIBiT core collection (A) Manhattan plot of GWAS analysis using BLINK, with three principal components (PCs) in 2021 Plant Height. In green highlighted significant markers by Bonferroni correction (logarithm of odds (LOD) score of 5.69). In blue is the suggestive line (LOD score of 5) and in red the genome-wide significance line (LOD score of 5.5). (B) Boxplot showing phenotypic distribution of plant height of two significant single nucleotide polymorphisms (SNP) identified by GWAS with BLINK model as seen on (A) highlighted in green. Red star shows significance according to t-test (p-value < 0.05). (C) Manhattan plot of GWAS analysis using BLINK (PC4) in 2021 thousand kernel weight. (D) Boxplot showing phenotypic distribution of Thousand Kernel Weight of four significant SNPs identified by GWAS with BLINK model as seen on (C) highlighted in green. (E) Manhattan plot of GWAS analysis using BLINK (PC4) in 2021 protein content. (F) boxplot showing phenotypic distribution of protein content of three significant SNPs identified by GWAS with BLINK model as seen on (E) highlighted in green.

References

- Alqudah, A. M., and Schnurbusch, T. (2017). Heading date is not flowering time in spring barley. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00896
- Alqudah, A. M., Sharma, R., Pasam, R. K., Graner, A., Kilian, B., and Schnurbusch, T. (2014). Genetic dissection of photoperiod response based on GWAS of pre-anthesis phase duration in spring barley. *PLoS One* 9, e1131120. doi: 10.1371/journal.pone.0113120
- Anscombe, F. J., and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics* 5, 141–160. doi: 10.1080/00401706.1963.10490071
- Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23, 451–466. doi: 10.1016/j.tplants.2018.02.001
- Aslan, S., Forsberg, N. E. G., Hagenblad, J., and Leino, M. W. (2015). Molecular genotyping of historical barley landraces reveals novel candidate regions for local adaption. *Crop Sci.* 55, 1031–1034. doi: 10.2135/cropsci2015.02.0119
- Badr, A., Müller, K., Schäfer-Pregl, R., el Rabey, H., Effgen, S., Ibrahim, H. H., et al. (2000). On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.* 17, 499–510. doi: 10.1093/oxfordjournals.molbev.a026330
- Bayer, M. M., Rapazote-Flores, P., Ganai, M. W., Hedley, P. E., Macaulay, M., Plieske, J., et al. (2017). Development and evaluation of a barley 50k iSelect SNP array. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01792
- Berger, B., de Regt, B., and Tester, M. (2012). "High-Throughput Phenotyping of plant shoots," in *High-Throughput Phenotyping in Plants: Methods and Protocols Methods in Molecular Biology* ed. J. Normanly (New York, NY: Humana Press), 9–20.
- 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

- Brbaklić, L., Trkulja, D., Mikić, S., Miroslavić, M., Momčilović, V., Dudić, B., et al. (2021). Genetic diversity and population structure of Serbian barley (*Hordeum vulgare* L.) collection during a 40-year long breeding period. *Agronomy* 11, 118. doi: 10.3390/agronomy11010118
- Brien, C. (2023). *asremlPlus: Augments "ASReml-R" in Fitting Mixed Models and Packages Generally in Exploring Prediction Differences*. Available online at: <https://CRAN.R-project.org/package=asremlPlus> (Accessed March 14, 2023).
- Brown, A. H. D. (1989a). Core collections: a practical approach to genetic resources management. *Genome* 31, 818–824. doi: 10.1139/g89-144
- Brown, A. H. D. (1989b). The case for core collections. In A. H. D. Brown, et al (ed.) *Use Plant Genet. Resour.* Cambridge, England: Cambridge Univ. Press, 136–156.
- Bustos-Korts, D., Dawson, I. K., Russell, J. R., Tondelli, A., Guerra, D., Ferrandi, C., et al. (2019). Exome sequences and multi-environment field trials elucidate the genetic basis of adaptation in barley. *Plant J.* 99, 1172–1191. doi: 10.1111/tpj.14414
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2017). *ASReml-R Reference Manual Version 4*. (Hempstead: VSN International Ltd). Available at: <http://www.homepages.ed.ac.uk/iwhite/asreml/uop>.
- Caldwell, K. S., Russell, J. R., Langridge, P., and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172, 557–567. doi: 10.1534/genetics.104.038489
- Ceccarelli, S., Grando, S., and Van Leur, J. A. G. (1995). Barley landraces of the fertile crescent offer new breeding options for stress environments. *Diversity* 11, 112–113.
- Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. Available online at: <https://CRAN.R-project.org/package=pwr> (Accessed March 14, 2023).
- Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* 11, 381–393. doi: 10.1198/108571106X154443
- De Beukelaer, H., Davenport, G. F., and Fack, V. (2018). Core Hunter 3: Flexible core subset selection. *BMC Bioinf.* 19, 1–12. doi: 10.1186/s12859-018-2209-z
- Dreiseitl, A. (2021). Genotype heterogeneity in accessions of a winter barley core collection assessed on postulated specific powdery mildew resistance genes. *Agronomy* 11, 513. doi: 10.3390/agronomy11030513
- Dreiseitl, A., and Nesvadba, Z. (2021). Powdery mildew resistance genes in single-plant progenies derived from accessions of a winter barley core collection. *Plants* 10, 1–9. doi: 10.3390/plants10101988
- Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7
- Euromalt (2021). *Euromalt Statistics*. Available online at: <https://www.euromalt.be/euromalt-statistics> (Accessed June 6, 2023).
- European Commission (2020). A Farm to Fork Strategy for a fair, healthy and environmentally-friendly food system COM(2020)/381 final. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. COM(2020) 381 final. Brussels 20.5.2020. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0381>. [Accessed 03 March 2024]
- Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x
- Faccini, N., Delbono, S., Oğuz, A. Ç., Cattivelli, L., Vale, G., and Tondelli, A. (2021). Resistance of European spring 2-row barley cultivars to pyrenophora graminea and detection of associated loci. *Agronomy* 11, 374. doi: 10.3390/agronomy11020374
- Fernández-Calleja, M., Casas, A. M., and Igartua, E. (2021). Major flowering time genes of barley: allelic diversity, effects, and comparison with wheat. *Theor. Appl. Genet.* 134, 1867–1897. doi: 10.1007/s00122-021-03824-z
- Fischbeck, G. (1992). *Barley Genetics VI: Proceedings of the 6th International Barley Genetics Symposium* (Sweden: Helsingborg).
- Fischbeck, G. (2003). “Diversification through breeding,” in *Diversity in Barley (Hordeum Vulgare)*. Eds. R. V. Bothmer, T. V. Hintum, H. Knuepfer and K. Sato (Elsevier, Amsterdam), 29–52.
- Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression* Vol. 16 (Vienna: R Foundation for Statistical Computing). Available at: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>. Sage.
- Frankel, O. H., and Brown, A. H. D. (1984). “Plant genetic resources today: a critical appraisal,” in *Crop genetic resources: conservation and evaluation* ed. J. H. W. Holden and J. T. Williams (London: George Allen and Unwind), 249–257.
- Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Source: J. Agricultural Biological Environ. Stat.* 2, 269–293. doi: 10.2307/1400446
- Guido, L. F. (2019). Brewing and craft beer. *Beverages* 5, 51. doi: 10.3390/beverages5030051
- Hagenblad, J., and Leino, M. W. (2022). Chevalier barley: The influence of a world-leading malting variety. *Crop Sci.* 62, 235–246. doi: 10.1002/csc.2.20668
- Harrell, F. E. (2023). *Hmisc: Harrell Miscellaneous*. Available online at: <https://CRAN.R-project.org/package=Hmisc> (Accessed March 14, 2023).
- Harwood, W. A. (2019). *Barley*. (Humana New York, NY). doi: 10.1007/978-1-4939-8944-7
- He, T., Hill, C. B., Angessa, T. T., Zhang, X. Q., Chen, K., Moody, D., et al. (2019). Gene-set association and epistatic analyses reveal complex gene interaction networks affecting flowering time in a worldwide barley collection. *J. Exp. Bot.* 70, 5603–5616. doi: 10.1093/jxb/erz332
- Hertrich (2013). Topics in brewing: malting barley. *Tech. Q.* 50, 29–41. doi: 10.1094/TQ-50-1-0331-01
- Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* 8, giy154. doi: 10.1093/gigascience/giy154
- Igartua, E., Gracia, M. P., Lasa, J. M., Medina, B., Molina-Cano, J. L., Montoya, J. L., et al. (1998). The Spanish barley core collection. *Genet. Resour. Crop Evol.* 45, 475–481. doi: 10.1023/A:1008662515059
- Jombart, T., Balloux, F., and Dray, S. (2010). adephylo: New tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* 26, 1907–1909. doi: 10.1093/bioinformatics/btq292
- Jones, H., Civián, P., Cockram, J., Leigh, F. J., Smith, L. M., Jones, M. K., et al. (2011). Evolutionary history of barley cultivation in Europe revealed by genetic analysis of extant landraces. *BMC Evolutionary Biol.* 11, 1–12. doi: 10.1186/1471-2148-11-320
- Khodaeiaminjan, M., Knoch, D., Ndella, T. M. R., Marchetti, C. F., Kořínková, N., Techer, A., et al. (2023). Genome-wide association study in two-row spring barley landraces identifies QTLs associated with plantlets root system architecture traits in well-watered and osmotic stress conditions. *Front. Plant Sci.* 14, 1120. doi: 10.3389/fpls.2023.1125672
- Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* 9, 1–9. doi: 10.1186/1746-4811-9-29
- Kristensen, P. S., Jahoor, A., Andersen, J. R., Cericola, F., Orabi, J., Janss, L. L., et al. (2018). Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. *Front. Plant Sci.* 9, 69. doi: 10.3389/fpls.2018.00069
- Kumar, A., Verma, R. P. S., Singh, A., Kumar Sharma, H., and Devi, G. (2020). Barley landraces: Ecological heritage for edaphic stress adaptations and sustainable production. *Environ. Sustainability Indic.* 6, 100035. doi: 10.1016/j.indic.2020.100035
- Lakew, B., Semeane, Y., Alemayehu, F., Gebre, H., Grando, S., Van Leur, J. A. G., et al. (1997). Exploiting the diversity of barley landraces in Ethiopia. *Genet. Resour. Crop Evol.* 44, 109–116. doi: 10.1023/A:1008644901982
- Leroy, T., De Bellis, F., Legnate, H., Musoli, P., Kalonji, A., Looor Solorzano, R. G., et al. (2014). Developing core collections to optimize the management and the exploitation of diversity of the coffee *Coffea canephora*. *Genetica* 142, 185–199. doi: 10.1007/s10709-014-9766-5
- Liu, F., Sun, G. L., Salomon, B., and Von Bothmer, R. (2001). Distribution of allozymic alleles and genetic diversity in the American Barley Core Collection. *Theor. Appl. Genet.* 102, 606–615. doi: 10.1007/s001220051687
- Liu, F., Sun, G. L., Salomon, B., and von Bothmer, R. (2002). Characterization of genetic diversity in core collection accessions of wild barley, *Hordeum vulgare* ssp. spontaneum. *Heredity* 136, 67–73. doi: 10.1034/j.1601-5223.2002.1360110.x
- Liu, F., von Bothmer, R., and Salomon, B. (1999). Genetic diversity among East Asian accessions of the barley core collection as revealed by six isozyme loci. *Theor. Appl. Genet.* 98, 1226–1233. doi: 10.1007/s001220051188
- Liu, F., Von Bothmer, R., and Salomon, B. (2000). Genetic diversity in European accessions of the Barley Core Collection as detected by isozyme electrophoresis. *Genet. Resour. Crop Evol.* 47, 571–581. doi: 10.1023/A:1026532215990
- Looseley, M. E., Ramsay, L., Bull, H., Swanson, J. S., Shaw, P. D., Macaulay, M., et al. (2020). Association mapping of malting quality traits in UK spring and winter barley cultivar collections. *Theor. Appl. Genet.* 133, 2567–2582. doi: 10.1007/s00122-020-03618-9
- Malysheva-Otto, L. V., Ganal, M. W., Law, J. R., Reeves, J. C., and Röder, M. S. (2007). Temporal trends of genetic diversity in European barley cultivars (*Hordeum vulgare* L.). *Mol. Breed.* 20, 309–322. doi: 10.1007/s11032-007-9093-y
- Malysheva-Otto, L. V., Ganal, M. W., and Röder, M. S. (2006). Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet.* 7, 1–14. doi: 10.1186/1471-2156-7-6
- Marone, D., Russo, M. A., Mores, A., Ficco, D. B. M., Laidò, G., Mastrangelo, A. M., et al. (2021). Importance of landraces in cereal breeding for stress tolerance. *Plants* 10, 1267. doi: 10.3390/plants10071267
- Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* 51, 1076–1081. doi: 10.1038/s41588-019-0443-6
- Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., et al. (2021). Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* 33, 1888–1906. doi: 10.1093/plcell/koab077
- Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (2021). *IPCC 2021: Climate Change 2021: The Physical Science Basis* (New York, NY, USA: Cambridge University Press, Cambridge, United Kingdom).
- Masted, N., Scholten, M., Ford-Lloyd, B., Allender, C., Astley, D., Vincent, H., et al. (2014). *Landrace conservation strategy for the United Kingdom* (Birmingham, UK: The University of Birmingham).
- Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., et al. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51, 319–326. doi: 10.1038/s41588-018-0266-x

- Milotova, J., Martynov, S. P., Dobrotvorskaya, T. V., and Vaculova, K. (2008). Genealogical analysis of the diversity of spring barley cultivars released in former Czechoslovakia and modern Czech Republic. *Russ J. Genet.* 44, 51–59. doi: 10.1134/S1022795408010079
- Monteagudo, A., Casas, A. M., Cantalapiedra, C. P., Contreras-Moreira, B., Gracia, M. P., and Igartua, E. (2019). Harnessing novel diversity from landraces to improve an elite barley variety. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00434
- Muñoz-Amatrián, M., Cuesta-Marcos, A., Endelman, J. B., Comadran, J., Bonman, J. M., Bockelman, H. E., et al. (2014). The USDA barley core collection: Genetic diversity, population structure, and potential for genome-wide association studies. *PLoS One* 9, 1–13. doi: 10.1371/journal.pone.0094688
- Nejat, N. (2022). Gene editing of the representative WRKY family members in an elite malting barley cultivar RGT Planet by CRISPR/Cas9. *Diss. Murdoch University*. doi: 10.13140/RG.2.2.20399.10408
- Newton, A. C., Flavell, A. J., George, T. S., Leat, P., Mullholland, B., Ramsay, L., et al. (2011). Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. *Food Secur.* 3, 141–178. doi: 10.1007/s12571-011-0126-3
- Newton, A. C., Gravouil, C., and Fountaine, J. M. (2010). Managing the ecology of foliar pathogens: Ecological tolerance in crops. *Ann. Appl. Biol.* 157, 343–359. doi: 10.1111/j.1744-7348.2010.00437.x
- Nolan, P., and Flanagan, J. (2022). High-resolution climate projections for Ireland - A multi-model ensemble approach. Available from: <https://www.epa.ie/publications/research/climate-change/research-339-high-resolution-climate-projections-for-ireland.php>. [Accessed 3rd March 2024]
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412
- Pasam, R. K., Sharma, R., Walther, A., Özkan, H., Graner, A., and Kilian, B. (2014). Genetic diversity and population structure in a legacy collection of spring barley landraces adapted to a wide range of climates. *PLoS One* 9, e116164. doi: 10.1371/journal.pone.0116164
- Plarr, W., Hoffmann, W., Göpp, K., and Broekhuizen, S. (1963). “Barley growing and breeding in Europe,” in S. Broekhuizen (Ed.), *International Barley Genetics Symposium*, Wageningen.
- 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
- Prosekov, A. Y., and Ivanova, S. A. (2018). Food security: The challenge of the present. *Geoforum* 91, 73–77. doi: 10.1016/j.geoforum.2018.02.030
- Raubach, S., Kilian, B., Dreher, K., Amri, A., Bassi, F. M., Boukar, O., et al. (2021). From bits to bites: Advancement of the Germinate platform to support prebreeding informatics for crop wild relatives. *Crop Sci.* 61, 1538–1566. doi: 10.1002/csc2.20248
- R Core Team (2022) *R: A Language and Environment for Statistical Computing*. Available online at: <https://www.R-project.org/> (Accessed November 22, 2022).
- Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x
- Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., et al. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci.* 109, 8872–8877. doi: 10.1073/pnas.1120813109
- Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P. R., Roose, M. L., et al. (2006). Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc. Natl. Acad. Sci.* 103, 18656–18661. doi: 10.1073/pnas.0606133103
- Russell, J. R., Ellis, R. P., Thomas, W. T., Waugh, R., Provan, J., Booth, A., et al. (2000). A retrospective analysis of spring barley germplasm development from foundation genotypes to currently successful cultivars. *Mol. Breed.* 6, 553–568. doi: 10.1023/A:1011372312962
- Russell, J., Mascher, M., Dawson, I. K., Kyriakidis, S., Calixto, C., Freund, F., et al. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* 48, 1024–1030. doi: 10.1038/ng.3612
- Saade, S., Brien, C., Pailles, Y., Berger, B., Shahid, M., Russell, J., et al. (2020). Dissecting new genetic components of salinity tolerance in two-row spring barley at the vegetative and reproductive stages. *PLoS One* 15, 1–19. doi: 10.1371/journal.pone.0236037
- Saade, S., Maurer, A., Shahid, M., Oakey, H., Schmöckel, S. M., Negrão, S., et al. (2016). Yield-related salinity tolerance traits identified in a nested association mapping (NAM) population of wild barley. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep32586
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Sasidharan, R., Voisenek, L. A. C. J., and Perata, P. (2021). Plant performance and food security in a wetter world. *New Phytol.* 229, 5–7. doi: 10.1111/nph.17067
- Schmidt, S. B., Brown, L. K., Booth, A., Wishart, J., Hedley, P. E., Martin, P., et al. (2023). Heritage genetics for adaptation to marginal soils in barley. *Trends Plant Sci.* 28, 544–551. doi: 10.1016/j.tplants.2023.01.008
- Schmidt, S. B., George, T. S., Brown, L. K., Booth, A., Wishart, J., Hedley, P. E., et al. (2019). Ancient barley landraces adapted to marginal soils demonstrate exceptional tolerance to manganese limitation. *Ann. Bot.* 123, 831–843. doi: 10.1093/aob/mcy215
- Schreiber, M., Wonneberger, R., Haaning, A. M., Coulter, M., Russell, J., Himmelbach, A., et al. (2024). Genomic resources for a historical collection of cultivated two-row European spring barley genotypes. *Sci. Data* 11, 66. doi: 10.1038/s41597-023-02850-4
- Selçuk, A., Forsberg, N., Hagenblad, J., and Leino, M. W. (2015). Molecular genotyping of historical barley landraces reveals novel candidate regions for local adaptation. *Crop Sci.* 55, 2766–2776. doi: 10.2135/cropsci2015.02.0119
- Shaw, P. D., Graham, M., Kennedy, J., Milne, I., and Marshall, D. F. (2014). Helium: visualization of large scale plant pedigrees. *BMC Bioinf.* 15, 1–15. doi: 10.1186/1471-2105-15-259
- Shin, J.-H., Blay, S., McNeeney, B., and Graham, J. (2006). LDheatmap: an R function for graphical display of pairwise linkage disequilibrium between single nucleotide polymorphisms. *J. Stat. Soft.* 16, 1–9. doi: 10.18637/jss.v016.c03
- Slama, A., Mallek-Malej, E., Mohamed, H., Rhim, T., and Radhouane, L. (2018). A return to the genetic heritage of durum wheat to cope with drought heightened by climate change. *PLoS One* 13, e0196873. doi: 10.1371/journal.pone.0196873
- Sleight, J. (2022) Cutting emissions in malting barley by 50% in five years. In: *The Scottish Farmer*. Available online at: <https://www.thescottishfarmer.co.uk/news/23165268.cutting-emissions-malting-barley-50-five-years/> (Accessed June 6, 2023).
- Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods - A bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23, 1164–1167. doi: 10.1093/bioinformatics/btm069
- Storey, J. D., Bass, A. J., Dabney, A., and Robinson, D. (2022) *qvalue: Q-value estimation for false discovery rate control*. Available online at: <http://github.com/jdstorey/qvalue>.
- Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100, 9440–9445. doi: 10.1073/pnas.1530509100
- Stracke, S., and Borrner, A. (1998). Molecular mapping of the photoperiod response gene *ea7* in barley. *Theor. Appl. Genet.* 97, 797–800. doi: 10.1007/s001220050958
- Tankley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild narrow genetic base of crop plants. *Sci.* (1979) 277, 1063–1066. doi: 10.1126/science.277.5329.1063
- Tondelli, A., Xu, X., Moragues, M., Sharma, R., Schnaithmann, F., Ingvarsdson, C., et al. (2013). Structural and temporal variation in genetic diversity of European spring two-row barley cultivars and association mapping of quantitative traits. *Plant Genome* 6, 1–14. doi: 10.3835/plantgenome2013.03.0007
- Turner, A., Beales, J., Faure, S., Dunford, R. P., and Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Sci.* (1979) 11, 1031–1034. doi: 10.1126/science.1117682
- Uleberg, E., Hanssen-Bauer, I., van Oort, B., and Dalmannsdottir, S. (2014). Impact of climate change on agriculture in Northern Norway and potential strategies for adaptation. *Clim. Change* 122, 27–39. doi: 10.1007/s10584-013-0983-1
- Umego, E. C., and Barry-Ryan, C. (2022). Overview of the Irish brewing and distilling sector: Processing inputs supply and quality requirements. *BrewingScience* 75, 9–16. doi: 10.23763/BrSc21-19umego
- van Hintum, T. J. L. (1994). Comparison of marker systems and construction of a core collection in a pedigree of European spring barley. *Theor. Appl. Genet.* 89, 991–997. doi: 10.1007/BF00224529
- van Treuren, R., Tchoudinova, I., van Soest, L. J. M., and van Hintum, T. J. L. (2006). Marker-assisted acquisition and core collection formation: A case study in barley using AFLPs and pedigree data. *Genet. Resour. Crop Evol.* 53, 43–52. doi: 10.1007/s10722-004-0585-x
- Wang, X., Ando, K., Wu, S., Reddy, U. K., Tamang, P., Bao, K., et al. (2021). Genetic characterization of melon accessions in the U.S. National Plant Germplasm System and construction of a melon core collection. *Mol. Horticulture* 1, 1–13. doi: 10.1186/s43897-021-00014-9
- Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinf.* 19, 629–640. doi: 10.1016/j.gpb.2021.08.005
- Wei, T., and Simko, V. (2021) *Package “corplot”: Visualization of a Correlation Matrix*. Available online at: <https://github.com/taiyun/corplot> (Accessed March 14, 2023).
- Weise, S., Oppermann, M., Maggioni, L., Van Hintum, T., and Knupffer, H. (2017). EURISCO: The European search catalogue for plant genetic resources. *Nucleic Acids Res.* 45, D1003–D1008. doi: 10.1093/nar/gkw755
- Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). Available online at: <https://ggplot2.tidyverse.org> (Accessed March 14, 2023).
- Xie, W., Xiong, W., Pan, J., Ali, T., Cui, Q., Guan, D., et al. (2018). Decreases in global beer supply due to extreme drought and heat. *Nat. Plants* 4, 964–973. doi: 10.1038/s41477-018-0263-1
- Zadoks, J. C., Chang, T. T., and Konzak, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Res.* 14, 415–421. doi: 10.1111/j.1365-3180.1974.tb01084.x



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Siraj Ismail Kayondo,
International Institute of Tropical Agriculture
(IITA), Nigeria
Javaid Akhter Bhat,
Nanjing Agricultural University, China

*CORRESPONDENCE

Brian M. Irish
✉ brian.irish@usda.gov

[†]These authors have contributed
equally to this work and share
first authorship

RECEIVED 15 November 2023

ACCEPTED 18 March 2024

PUBLISHED 03 April 2024

CITATION

Zhao D, Sapkota M, Lin M, Beil C, Sheehan M,
Greene S and Irish BM (2024) Genetic
diversity, population structure, and taxonomic
confirmation in annual medic (*Medicago* spp.)
collections from Crimea, Ukraine.
Front. Plant Sci. 15:1339298.
doi: 10.3389/fpls.2024.1339298

COPYRIGHT

© 2024 Zhao, Sapkota, Lin, Beil, Sheehan,
Greene and Irish. 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.

Genetic diversity, population structure, and taxonomic confirmation in annual medic (*Medicago* spp.) collections from Crimea, Ukraine

Dongyan Zhao^{1†}, Manoj Sapkota^{1†}, Meng Lin¹, Craig Beil¹,
Moira Sheehan¹, Stephanie Greene² and Brian M. Irish^{3*}

¹Breeding Insight, Cornell University, Ithaca, NY, United States, ²Agricultural Genetic Resources Preservation Research Unit, United States Department of Agriculture (USDA), Agricultural Research Service (ARS), Prosser, WA, United States, ³Plant Germplasm Introduction and Testing Research Unit, United States Department of Agriculture (USDA), Agricultural Research Service (ARS), Prosser, WA, United States

Annual medic (*Medicago* spp.) germplasm was collected from the Crimean Peninsula of Ukraine in 2008 to fill gaps in geographic coverage in the United States department of Agriculture, Agricultural Research Service, National Plant Germplasm System (NPGS) temperate-adapted forage legume collection. A total of 102 accessions across 10 *Medicago* species were collected. To assess genetic diversity, population structure, and to confirm taxonomic identities, the collections were phenotypically and genetically characterized. Phenotyping included the use of 24 descriptor traits while genetic characterization was accomplished using a 3K Diversity Array Technologies (DARtag) panel developed for alfalfa (*Medicago sativa* L.). For both field and molecular characterizations, a reference set of 92 geographically diverse and species-representative accessions were obtained from the NPGS collection. Phenotypic descriptors showed consistency among replicated plants within accessions, some variation across accessions within species, and evident distinctions between species. Because the DARtag panel was developed for cultivated alfalfa, the transferability of markers to the species being evaluated was limited, resulting in an average of ~1,500 marker loci detected per species. From these loci, 448 markers were present in 95% of the samples. Principal component and phylogenetic analysis based on a larger set of 2,396 selected markers clustered accessions by species and predicted evolutionary relationships among species. Additionally, the markers aided in the taxonomic identity of a few accessions that were likely mislabeled. The genotyping results also showed that sampling individual plants for these mostly self-pollinating species is sufficient due to high reproducibility between single (n=3) and pooled (n=7) biological replicate leaf samples. The phenotyping and the 2,396 Single Nucleotide Polymorphism (SNP) marker set were useful in estimating population structure in the Crimean and reference accessions, highlighting novel and unique genetic diversity captured in the Crimean accessions. This research not only demonstrated the

utility of the DArTag marker panel in evaluating the Crimean germplasm but also highlighted its broader application in assessing genetic resources within the *Medicago* genus. Furthermore, we anticipate that our findings will underscore the importance of leveraging genetic resources and advanced genotyping tools for sustainable crop improvement and biodiversity conservation in annual medic species.

KEYWORDS

diversity, DArTag genotyping, marker, germplasm, legume

1 Introduction

The genus *Medicago* L. includes agriculturally significant crops like alfalfa (*M. sativa* L.) and the model legume *M. truncatula* L. Other important perennial and annual species useful as forages and cover crops are also in the genus. Many of the species are valuable because they cover the ground effectively preventing erosion and weeds, can help build soil organic matter content, are often a nutrient rich source of fodder, and can fix atmospheric nitrogen because of the ability to form a symbiosis with root nodulating bacteria. Most of the 87 currently described taxonomic species in the genus are of Eurasian or Mediterranean origin with ranges extending into Europe and North Africa (Small, 2011). Furthermore, due to the weedy nature of some species and extensive human-mediated dissemination, especially in the case of alfalfa, many of these species can now be found across the globe.

Alfalfa or lucerne is the most widely grown and important perennial forage legume crop worldwide (Undersander et al., 2021). In the United States, it was the fourth most cultivated crop in 2021 with an estimated direct value of \$11.6 billion (Putnam and Meccage, 2022) and ranked first among forage crops by planting area with a total of 14.9 million acres in 2022 (<https://www.nass.usda.gov/>). Alfalfa is a key nutritional component for dairy and beef production because it contains a high amount of crude protein, provides dietary fiber needed to maintain rumen health, and is an excellent source of vitamins and minerals. In addition, it is unparalleled as a component of sustainable agricultural systems because of its perennialism, ability to fix nitrogen, protect water quality, interrupt pest and pathogen cycles in annual crops, and improve soil carbon storage (Fernandez et al., 2019). Alfalfa is a widely adapted plant that can grow in a range of environments, but its performance and adaptability can be influenced by factors such as climate, soil type, location, and management (Undersander et al., 2021). Alfalfa is a tetraploid, insect-pollinated, outcrossing crop, so it is highly heterogeneous and heterozygous with inherent genetic diversity. Improving and diversifying alfalfa cultivars is challenging, particularly concerning the effective enhancement of quantitative traits like yield. There is need to improve breeding strategies that harness the crop's existing

diversity to enhance key attributes. Furthermore, the exploration and utilization of novel genetic resources, including those of wild relatives, and the assessment of genetic diversity are of utmost importance for the sustainable development of improved alfalfa. Traits have been introgressed into cultivated alfalfa from several wild relatives (Mizukami et al., 2006; Humphries et al., 2021). However, challenges remain in utilizing biotechnology and molecular marker techniques to improve breeding efforts for alfalfa and related annual medic crops.

Of the thirty-five annual medic mostly diploid species described, several including barrel medic (*M. truncatula*), black medic (*M. lupulina*), burr medic (*M. polymorpha*), snail medic (*M. scutellata*), and strand medic (*M. litoralis*) have been used as forage crops and cover crops (Zhu et al., 1996; Fisk et al., 2001; Muir et al., 2006). Many of these annuals are an essential component of native Mediterranean region flora and as a fodder source (Piano and Francis, 1992; Small, 2011). They also have been introduced and used in southern Australian pasture systems with breeding efforts focusing on germination, vigor, adaptation to those edaphic conditions, and forage quality (Crawford et al., 1989; Nichols et al., 2012). Ideal growing conditions for many of the annual medics include warm dry summers with cooler winter periods, with well-drained light soils with neutral to basic pH levels (Zhu et al., 1996; Muir et al., 2006). Many of these species persist by producing hard seeds that germinate the following growing season (i.e., self-reseeding) leading to regenerating or 'perennial' pastures.

Genetic diversity conserved in plant germplasm collections is crucial for crop improvement, serving as the foundation for selection and breeding programs (Govindaraj et al., 2015). Novel genetic resources, characterized by unique traits and allelic variations, have significant potential for developing improved cultivars with enhanced productivity, stress tolerance, and nutritional quality. These resources are acquired through the collection projects and from other germplasm collections. Genebanks preserve and make this germplasm representing wild relatives, landraces, and varieties available (Gabriel, 1992; Guzzon et al., 2022). Genetic diversity within collections plays a crucial role in long-term crop sustainability. It provides resilience against environmental changes, pests, and diseases by maintaining a

genepool of potentially advantageous alleles that can be selected under diverse pressures. Additionally, diverse populations are less susceptible to genetic erosion and offer a buffer against the loss of genetic adaptability and vulnerability to emerging threats (Mundt, 2002; Bijlsma and Loeschcke, 2012; King and Lively, 2012).

In the United States, the National Plant Germplasm System (NPGS) of the Department of Agriculture (USDA), Agricultural Research Service (ARS), is responsible for conservation and facilitation of access to a wide range of cultivated and wild plant genetic resources that are essential for the continued sustainability and advancement of plant agriculture (Postman et al., 2006; Byrne et al., 2018). A significant temperate-adapted forage legume (TFL) collection is managed as part of the NPGS collections in Pullman, WA. The collection includes larger subsets of accessions for important crops like alfalfa (>4,000 unique accessions), cultivated clovers (*Trifolium* spp. - e.g., red, and white clover), birdsfoot trefoil (*Lotus corniculatus* L.) as well as many wild relatives (Irish and Greene, 2021). Within the *Medicago* genus, large subsets of accessions are represented by many perennial and annual species, many of which are wild relatives of alfalfa and/or are crops themselves.

Assessing genetic diversity, population structure, reducing redundancy, and correctly assigning taxonomic identities in plant germplasm collections are important approaches to implement when managing plant genetic resource collections. Historically, highly heritable phenotypic traits have been used to address this need, with descriptor traits developed for many crops species (<https://alliancebioversityciat.org/publications-data> “Descriptors”) including the annual medics (International Board for Plant Genetic Resources (IBPGR), 1991). However, modern molecular marker approaches have evolved significantly and exceed as a tool for addressing these needs. Molecular marker techniques have been used for decades to study population structure and genetic diversity in plant germplasm collections in addition to enhance breeding efforts and accelerate genetic gains for major staple food crops (Tanksley, 1983; Helentjaris et al., 1985; Feuerstein et al., 1990; and reviewed in Hasan et al., 2021).

The development and use of molecular markers in forage crops like alfalfa and related *Medicago* species has been much slower than in other important agricultural food crops (Barker and Warnke, 2001). Various molecular and genomic techniques have been applied in alfalfa and other *Medicago* research. Molecular markers, such as random amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), and simple-sequence repeats (SSRs), have been utilized to assess genetic diversity within different *Medicago* germplasm collections (Brummer et al., 1991; Kidwell et al., 1994; Musial et al., 2002; Maureira et al., 2004), construct linkage maps (Brummer et al., 1993; Brouwer and Osborn, 1999; Kaló et al., 2000), and conduct association studies (Barcaccia et al., 2000; Brouwer et al., 2000; Tavoletti et al., 2000). Nevertheless, these methods are labor-intensive and have become somewhat outdated. The emergence of next-generation sequencing

technologies has introduced novel molecular approaches, including whole-genome sequencing and genotyping-by-sequencing (GBS), which facilitates the identification of genome-wide variants like single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) less than 20 bp. However, the computational demands are notably high, particularly for polyploid crops such as alfalfa. Considering these challenges, it is crucial to develop genomic tools tailored to the specific biological and logistical complexities of these crops. This tailored approach will have the potential to conveniently streamline the management of genetic resources and enhance breeding outcomes. One recently developed genotyping platform is DArTag (Diversity Array Technologies - DArT), which employs an amplicon-based targeted genotyping approach (Blyton et al., 2023; <https://www.diversityarrays.com/services/targeted-genotyping/>). Oligos are designed to anneal to the target known genetic variants (SNPs and InDels) to produce sequencing products of 54 bp (legacy technology) or 81 bp (current technology) in length. The reads can be used to call SNPs, or in the case of complex genomes like alfalfa, used to identify microhaplotypes because sequencing reads can contain variants beyond the target SNP, which allows for the detection of more than two alleles at each of the 3,000 loci (Zhao et al., 2023). As the amplicons are very short, variants found within these reads are assumed to be in very strong linkage disequilibrium and therefore can be used for phasing genotyping calls and determining allele dosage. The practical maximum number of probes on a DArTag panel is ~3,000 loci, though, it's worth noting that the optimal maximum may vary depending on the species and genome complexity, as well as the required read depths to accurately call genotypes (Andrzej Kilian, DArT, personal communication). One limitation of using targeted-amplicon sequencing platforms is the potential ascertainment bias (Heslot et al., 2013), which may restrict their applicability in diversity studies or when exploring species beyond those for which the marker panel was initially designed. Recognizing this issue, we view this study as a valuable test case to assess the panel's effectiveness on other *Medicago* species related to alfalfa.

In this study, our goals were to comprehensively characterize the diversity of a subset of annual medic (*Medicago* spp.) accessions obtained from a plant germplasm collecting mission to the Crimean Peninsula of Ukraine. More specifically, we aimed to assess the phenotypic variation within and among different species, shedding light on the unique traits and characteristics exhibited by each accession. Secondly, we viewed this study as a valuable test case to assess the effectiveness of a 3K DArTag alfalfa marker panel for genotyping annual medics. Additionally, we described and characterized the genetic diversity and population structure of this collection, using the markers to unravel the relationships among the various annual medic species. By accomplishing these objectives, we aim to provide breeders access to valuable, previously unexplored genetic resources, along with a useful genotyping tool for studying population structure in annual medic species, thereby facilitating the enhancement of their breeding programs.

2 Materials and methods

2.1 Plant materials

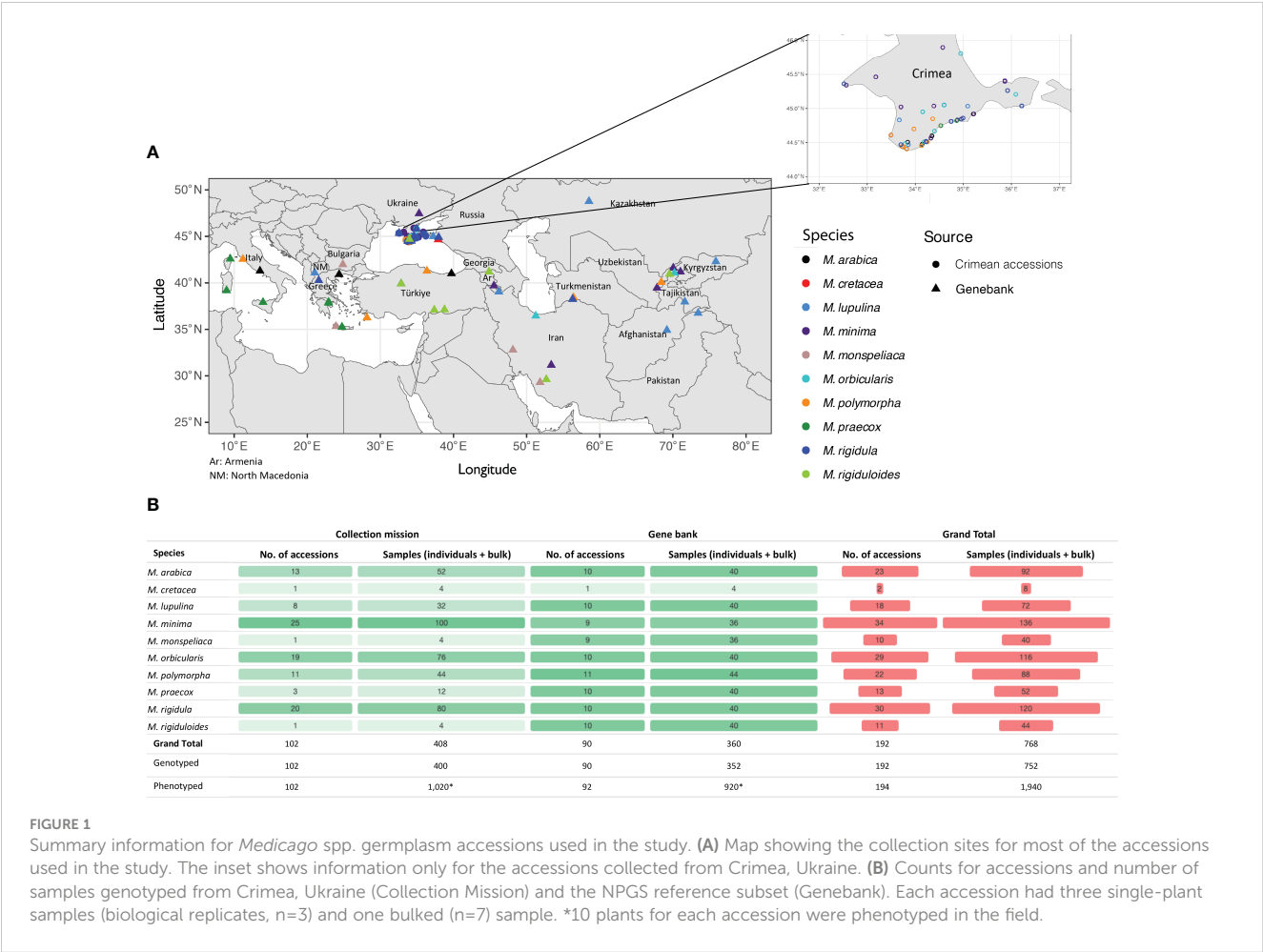
A total of 102 accessions across 10 *Medicago* species were collected along with their corresponding passport data from the Crimean Peninsula of Ukraine in 2008 (Figure 1A and Supplementary Table 1). The objective of the collecting trip was to fill gaps in the northern geographic range of several annual medic species, in particular, *M. truncatula*, which is used as an important genomic model species. The trip was a joint effort between the USDA NPGS, St. Petersburg State University, Russia, and the Ukrainian Genebank. A reference set of 92 accessions originating from geographically diverse regions of the world belonging to corresponding species available in the NPGS were also included in the study (Figure 1A and Supplementary Table 1).

Seeds for each accession were requested from the Pullman genebank, mechanically scarified, and 10 seeds for each accession were sown in bullet containers containing Sunshine Mix 3 soilless medium (Sun Gro, Agawam, MA) in early spring. Germinated seedlings were thinned to a single individual per container and were greenhouse-grown for approximately three months prior to field transplanting. Individual plants were established in non-replicated 10-plant field plots on the Washington State University (WSU),

Irrigated Agriculture Research and Extensions Center (IAREC), Roza Research farm in May 2019. Annual medic descriptor data was collected in the field and laboratory for 24 traits (Supplementary Table 2). That descriptors included phenological and phenotypic information on vegetative and reproductive stages following published protocols for annual medics (International Board for Plant Genetic Resources (IBPGR), 1991). Descriptor data was collected by assigning visual ratings and by using measuring equipment such as rules, calipers, and laboratory balances. A mean quantitative or categorical trait was assigned across the representative 10-plant sample for each accession.

2.2 Tissue collection and genotyping

Prior to field establishment, young leaf tissues were collected from greenhouse seedlings, lyophilized, and stored at -18°C for further processing. A total of ten individual plants were sampled from each accession for DNA extraction. The leaves/leaflets were collected into labeled 1.5 ml microcentrifuge tubes and came from three individual plants (single-plant sample; n=3) and from an additional seven individual plants where leaves/leaflets were pooled (bulked sample; n=7). Appropriate starting mass of the lyophilized leaf tissues were transferred to 96-well plates for DNA extraction,



which were performed inhouse using a DNeasy 96 Plant Kit (Qiagen, Valencia, CA) following the manufacturer's instructions. Samples were transferred to DArT-required Eppendorf twin.tec® microbiology Polymerase Chain Reaction (PCR) Plate 96 (Eppendorf, Enfield, CT). The DNA samples were then shipped to Diversity Array Technologies (DArT; www.diversityarrays.com) for genotyping.

Genotyping was performed using the recently developed alfalfa 3K DArTag panel (DArT product code: AlfaDArTagBICU; Zhao et al., 2023). This resulted in four genotyped samples (three single-plant samples and one bulked sample) for most accessions (Figure 1B). Altogether 768 samples across 192 accessions were genotyped using the DArTag panel. One *M. minima* and one *M. monspeliaca* accession from the reference set from NPGS were phenotyped but not genotyped. Briefly, the DArTag marker technology uses a targeted amplicon sequencing approach. Like other amplicon-based genotyping technologies, a short stretch of sequence (54–81 bp) around the target SNPs are captured, which allows for discovery of other off-target SNPs in addition to the reference and alternative microhaplotypes by assay design. Throughout the study, the genomic regions harboring the 3K SNPs were referred to as marker loci while the amplicons containing different combinations of the target and off-target SNPs from a marker locus were referred as microhaplotypes, which are provided in the DArTag report named missing allele discovery counts (MADC). For each marker locus, there could be varying numbers of microhaplotypes depending on the genetic diversity of the genotyped samples. It is worth noting that there may be some marker loci with paralogous sequences captured. However, these are usually with low read depths. Determining and eliminating paralogous microhaplotypes was out of scope for this study, therefore we expect some paralogous sequences to have been included in our final dataset.

2.3 Genotype processing

While processing the genotype in each of the sequenced samples, a read depth of one was considered a sequencing error for a single microhaplotype and was set to zero. To compare genetic

diversity captured in single-plant samples with bulked samples, the total number of microhaplotypes were counted for each sample. For each of the accessions that had three single-plant samples and one bulked sample ($n=181$ accessions), the median total number of microhaplotypes among the three single-plant samples was used to represent their genetic diversity and was compared with the bulked sample for that accession. To investigate the overall trend of microhaplotype differences between single-plant and bulked samples, all accessions were ordered based on their bulked sample microhaplotype number. A locally weighted scatterplot smoothing (LOWESS) regression (Cleveland, 1979) was performed using microhaplotype numbers from single-plant samples in R. The same methodology was applied to compare the combined single-plant sample (by pooling sequences of all three single-plant samples) with the bulked sample in each accession. Across the entire panel and within each of the 10 species, the average number of microhaplotypes per marker was calculated for all single-plant samples and all bulked samples separately. These microhaplotype counts were further corrected by the total number of microhaplotypes of each marker in the entire population to scale values between 0 and 1.

Marker loci with data (total read count > 10) in >5% of the samples were retained for genotype calling. Subsequent analyses were conducted using all SNPs, including both target and off-target SNPs, extracted from the microhaplotypes based on pairwise alignments with the reference microhaplotypes. SNP data were converted to variant call format (VCF) with the "FORMAT" field containing DP (total read depth), RA (reference allele read depth), and AD (read depth for each allele) (Figure 2). The polyRAD R package was used to estimate the ratio of individual heterozygosity to expected heterozygosity statistics (H_{ind}/H_E) at both the sample and SNP levels (Clark et al., 2019). The H_{ind}/H_E utilizes the likelihood that two sampled reads from a genotype will correspond to distinct alleles. A low H_{ind}/H_E value suggests the presence of alleles with low heterozygosity or sequencing errors, while a high H_{ind}/H_E value indicates that amplicons are likely derived from paralogous regions. In this study, only SNPs with a H_{ind}/H_E value between 0.1–1.0 were retained for subsequent analyses.

Furthermore, we acquired genotype information for 34 cultivated alfalfa accessions (Supplementary Table 1) utilized in the validation of

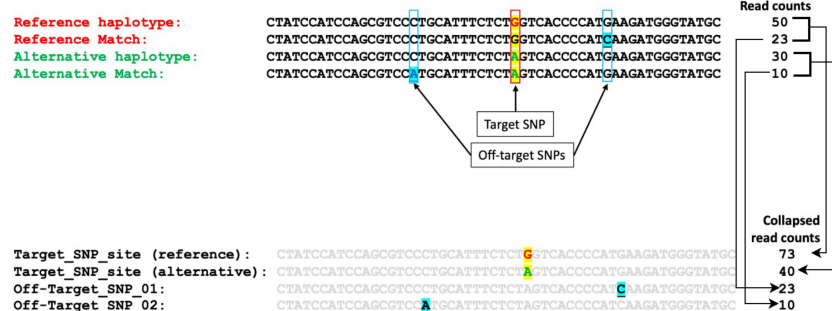


FIGURE 2

Illustration of the target and off-target SNPs extracted from microhaplotypes based on pairwise alignment with reference microhaplotypes and extraction of collapsed read counts from the individual read counts.

the 3K DArTag panel from Zhao et al. (2023). These accessions underwent the same genotype processing procedures, primarily for the purpose of comparing missing marker loci and calculating genetic distances (as detailed in Section 2.5).

2.4 Descriptive statistics and statistical analysis

All the statistical analyses were performed in R (R Core Team, 2013), unless otherwise mentioned. Pairwise Pearson correlation coefficients were calculated using read counts per marker for replicates (three single-plant samples and a bulked sample) within each of the 192 accessions using the R function “cor()”. Prior to the analysis of variance (ANOVA) for missing marker loci, Levene’s test was performed to determine if the variances between each group were equal. The test was done using “leveneTest()” function in R package “car” (Fox et al., 2007). Welch’s ANOVA (using “welch_anova_test()” function in R package “rstatix” (Kassambara, 2019) was then performed to address the violation of assumption of equal variances. Post-hoc tests for Welch’s ANOVA were performed using Games-Howell test. Games-Howell test is a multiple mean comparison for Welch’s ANOVA. Briefly, it is used to compare all possible combinations of group differences when the assumption of homogeneity of variances is violated. This post-hoc test provides confidence intervals for the differences between group means and shows whether the differences are statistically significant (Games and Howell, 1976). All the figures were generated using ggplot2 package in R (Wickham, 2011). The upset plot was generated using “UpSetR” package in R (Conway et al., 2017). The plant germplasm accessions used in the study were georeferenced using R package “rnatuarearth” (South, 2017; <https://CRAN.R-project.org/package=rnatuarearth>) and “ggspatial” (Dunnington, 2023; <https://CRAN.R-project.org/package=ggspatial>). The package “rnatuarearth” uses the public domain map dataset “Natural Earth” (<https://www.naturalearthdata.com/>) to generate the base layer of the map.

2.5 Population structure and phylogeny

Population structure based on phenotype data was performed using principal component analysis (PCA) with “PCA()” function in “FactoMineR” package in R (Lê et al., 2008). PCA was performed using only the accession and phenotypes without any missing data in any trait recorded, resulting in 175 accessions and 18 traits for the PCA. The character variables recorded for phenotypic traits were converted to numeric by one-hot encoding using the “model.matrix()” function in the “stat” package in R (Team R Core, 2013). Furthermore, population structure based on genotype was assessed using PCA by implementing the “AddPCA()” function in the “polyRAD” package in R (Clark et al., 2019). PCA was performed using dosage-based calls using both the entire set of markers and a subset of 448 marker loci that were evenly distributed across the genome and found in >95% of the genotyped accessions.

The clustering was based on the passport information of the germplasm. The phylogenetic tree was built using the Neighbor-Joining method (Saitou and Nei, 1987) and the Tamura-Nei genetic distance model (Tamura and Nei, 1993) with the entire genotype data and samples with the highest average read depth (altogether 192 samples) in Geneious Prime v2021.2.2 (Geneious Prime 2021; <https://www.geneious.com/>). Bootstrap support for the tree was obtained using 100 bootstrap replicates. The tree was then plotted using “ggtree” package in R (Yu et al., 2017). Weighted mean Fixation index (F_{ST}) values were calculated for all species pair combinations using genotype calls in vcftools with “- weier-fst-pop” parameter (Danecek et al., 2011). Additionally, a phylogenetic tree based on genetic distance (Fixation index: F_{ST}) was constructed using the R package “ape” (Paradis et al., 2004) and then visualized using the “ggtree” package (Yu et al., 2017).

3 Results

3.1 Phenotyping and population structure

The corresponding 10 field-established plants for all 194 Crimea and reference accessions were used to collect 25 phenological and highly heritable phenotypic descriptors following guidelines published for annual medics (International Board for Plant Genetic Resources (IBPGR), 1991). Descriptors included flowering and harvest dates, plant growth habit, leaf shape and markings, and pod (fruit), and seed characteristics and were recorded following instructions in guidelines by visual assessment or by using basic laboratory equipment such as calipers and balances (Figure 3; Supplementary Table 3). In a few situations, accessions did not flower or produce pods, but other traits were collected. Although slight variation might have been evident for plants within accessions, an average trait was collected for the entire accession. For example, there might have been some differences in the growth habit trait for some of the 10 plants within an accession. Some more upright than others, but on average a semi-erect descriptor trait value was assigned. In a few instances, and only within reference accessions, off-type plants were identified and were considered mistakes and excluded in the average for those traits. The off types in these situations were likely seed mixed in by mistake during the cleaning process of a species not corresponding to the accessions being evaluated. Notes were kept on occurrence of off types in accessions to try to remedy and/or warn future requestors. Also, differences in descriptor traits across accessions within species were evident with this variation being more prominent across reference accessions as they originated from a broader geographic area. Lastly, most descriptors for each of the ten species evaluated matched those in the published literature. An exception was accession W6 5374 which was received from the genebank labeled as a *M. polymorpha*, but likely was *M. rigidula* or *M. rigiduloides* based on comparisons to descriptors collected for those species. PCA based on the phenotypic data resulted in general species-wise clustering. Accessions associated with species such as *M. orbicularis*, *M. monspeliaca*, *M. lupulina*, and *M. minima* exhibited distinct clustering, indicating notable morphological variations among these

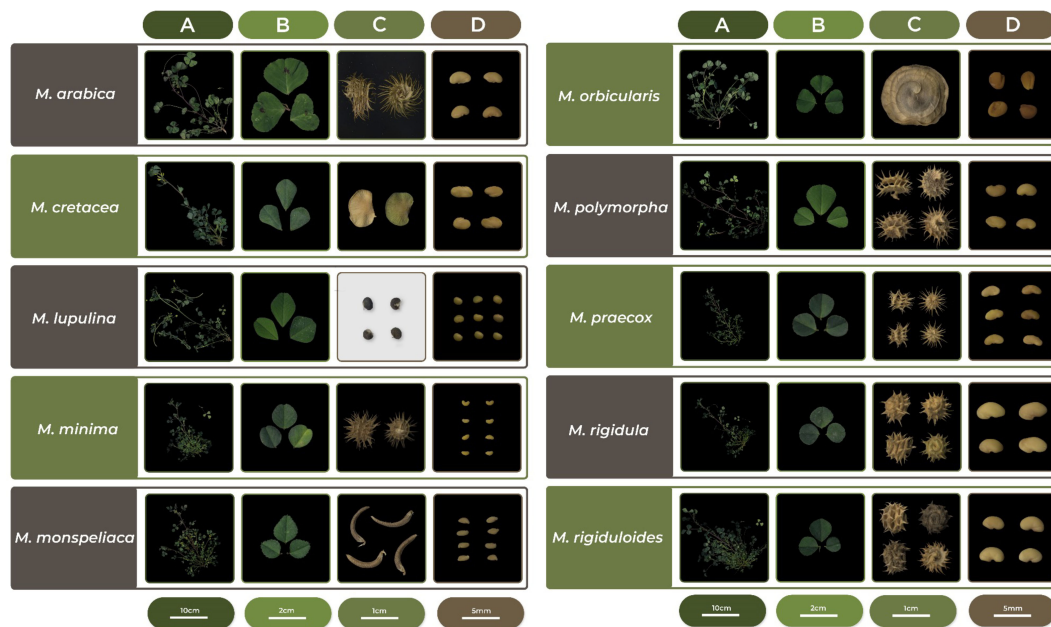


FIGURE 3

Representative phenotypic characteristics for the ten annual medic (*Medicago* spp.) species evaluated. (A) shoot architecture; (B) leaflets; (C) fruit/pods, and (D) seeds.

species (Figures 3, 4). Conversely, certain species like *M. rigidula*, *M. rigiduloides*, and *M. praecox* were found to cluster together, suggesting a high degree of phenotypic similarity (e.g., coiled pods with spines) among accessions even across different species (Figures 3, 4). Overall, these results underscore the presence of significant phenotypic diversity among the studied accessions.

3.2 Genotyping and filtering

Next, we aimed to investigate the genetic diversity underlying the observed phenotypic diversity in the studied accessions.

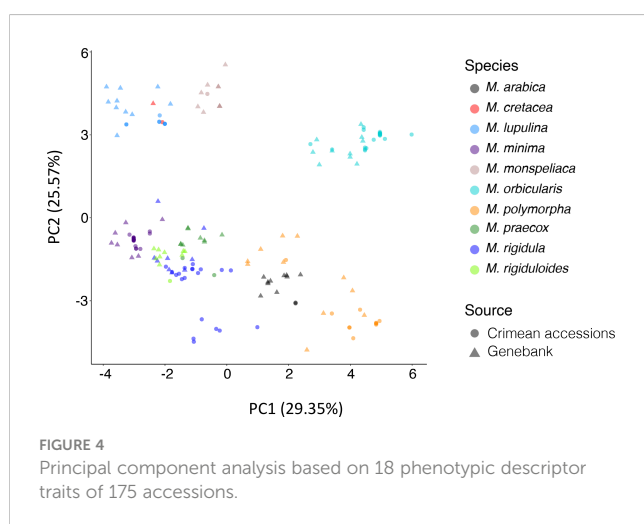


FIGURE 4

Principal component analysis based on 18 phenotypic descriptor traits of 175 accessions.

However, no molecular markers and genotyping platforms are currently available for research on annual medics; therefore, we tested the utility of an alfalfa 3K DArTag SNP panel (Zhao et al., 2023), developed specifically for alfalfa, a close relative of annual medics. A total of four samples (three single-plant and one bulked) for each accession were genotyped, with the exclusion of two reference accessions (PI 312480 – *M. monspeliaca* and W6 24404 – *M. minima*) to conform more readily to the 96-well genotyping platform format. It is worth noting that using this DArTag platform, for each marker locus, there could be varying numbers of microhaplotypes and SNP markers depending on the genetic diversity of the samples genotyped.

An initial filtering for missing data was conducted at the microhaplotype level by requiring at least a total of 10 reads per microhaplotype to be retained for subsequent analyses. Because the 3K panel was generated from alfalfa, we expected missing data rates would be higher for these annual medics (diploids). In total, 2,396 marker loci from the 3K panel showed signals in at least 5% of the 752 samples genotyped. On average, each sample exhibited signals from approximately 1,475 marker loci (Supplementary Table 4). To discover a shared set of SNP markers present in all *Medicago* species studied, we performed pairwise comparisons among all 10 species (Figure 5). However, most of the comparisons had very few common SNP markers (<20), indicating a distinct species-specific genetic variation among the accessions studied. Therefore, we elected to identify marker loci present in >95% of the samples, for which we successfully discovered 448 marker loci that were evenly distributed across the genome (Tables 1, 2). While the markers in these loci are shared among 95% of the samples, it is worth noting that some species-specific markers are valuable for the genetic

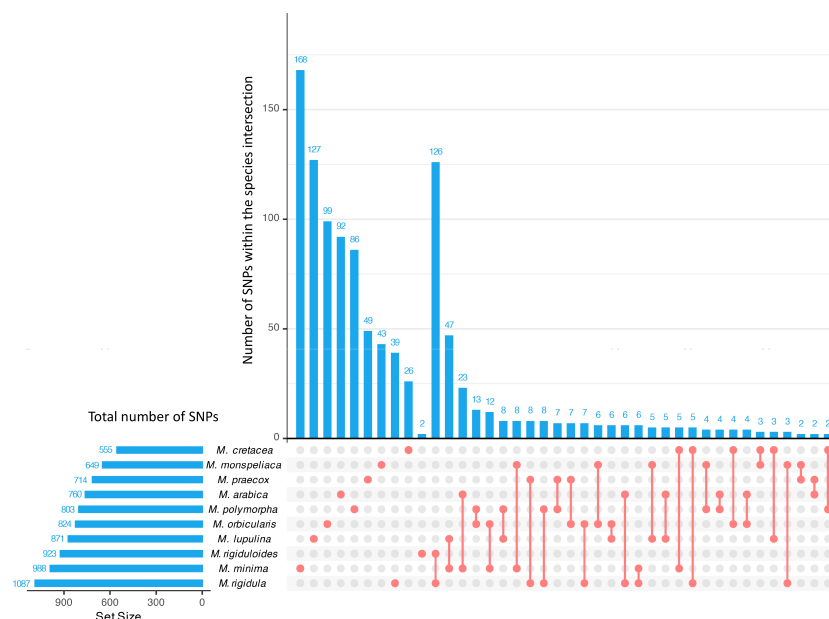


FIGURE 5

Upset plot of the intersection of polymorphic SNPs across different species. The horizontal bars represent total number of polymorphic SNPs found among the accessions within the corresponding species. The vertical bar plots represent the number of SNPs either unique to the specific species or common markers between two species indicated below the bars. The first ten vertical bars from the left to right represent SNPs unique to the species as indicated by the red dot whereas the remaining bars represent number of SNPs shared between two species indicated by red dot and the line segment.

characterization of different species (Figure 5). Hence, we utilized the larger marker dataset of 2,396 loci for further analysis.

3.3 Missing data rates

When evaluating the number of missing marker loci across all the samples, an interesting pattern regarding the variation in the number of missing marker loci was found. We explored missing marker loci across all the species and identified significant differences (p -value: 1.13×10^{-47}) (Figure 6). *M. monspeliaca* exhibited the highest average count ($2,037 \pm 288$) of missing marker loci, while *M. cretacea* displayed the lowest ($1,227 \pm 50$), indicating genomic variations among species (Figure 6B). Furthermore, a general trend of increasing missing marker loci across the 10 species was observed with their increasing genetic distances (Fixation index, F_{ST}) from *M. sativa* (Figure 6A).

3.4 Reproducibility and sensitivity of genotyping

The reproducibility and sensitivity (in terms of identifying rare alleles) of the genotyping results was also examined by calculating correlations among four samples (three single-plant samples and a bulked sample) for each accession. We observed a strong correlation among the four samples (average $R^2 = 0.965$), indicating consistent genotyping results (Figure 7 and

Supplementary Table 5). Likewise, the three single-plant samples and the bulked sample demonstrated strong correlations (average $R^2 = 0.967$) in read counts per marker locus, suggesting high reproducibility across the shared marker loci.

To compare genetic diversity that can be captured by the 3K DArTag panel, the total number of microhaplotypes were estimated for single-plant and bulked samples for each accession (Supplementary Table 6). Overall, the single-plant and bulked samples had an average of 2,027 and 2,126 microhaplotypes, respectively, from all 2,396 marker loci (a total of 12,945 microhaplotypes). Of the evaluated 181 accessions (with three single-plant samples and one bulked sample), 147 of the bulked samples showed a greater number (1 to 1,527; 0.01% to 11.80%) of microhaplotypes than the corresponding single-plant samples. Only 34 single-plant samples showed higher numbers (1 to 567; 0.01% to 4.38%) of microhaplotypes than their corresponding bulked samples (Figure 8; Supplementary Table 6). A LOWESS regression was employed to perform microhaplotype number smoothing across the 181 accessions. This was done by using the median haplotype number from the three single-plant samples for each accession. This analysis revealed a prevailing trend of slightly fewer microhaplotypes in single-plant samples compared to their bulked counterparts (Figure 8). Nevertheless, upon consolidating microhaplotypes from all three single-plant samples for each accession, we observed an increased count of microhaplotypes in 171 accessions and an average of 190 (14.68%) more microhaplotypes identified in the consolidated single-plant samples across all accessions (Supplementary Figure 1; Supplementary Table 6).

TABLE 1 Maker loci count found across accessions based on the percentage of accessions in which they were present.

% of genotyped samples	Maker loci #
5	2396
10	2291
15	2178
20	2081
25	1988
30	1877
35	1792
40	1709
45	1609
50	1510
55	1415
60	1320
65	1228
70	1132
75	1010
80	890
85	748
90	622
95	448
100	61

A set of 2,396 loci can be found in 5% of all samples. A set of 448 loci can be found in 95% of the samples.

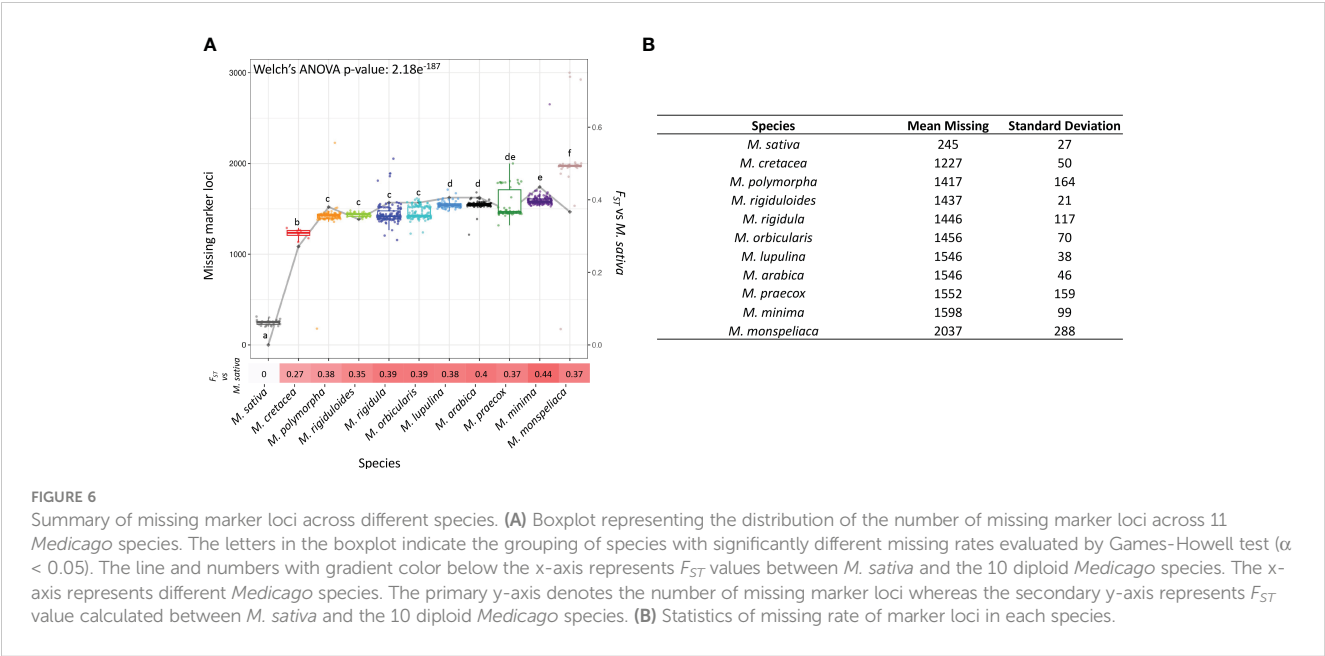
TABLE 2 Distribution of targeted loci across the eight chromosomes based on the 448-marker set found in 95% of samples in this study compared to the distribution of marker loci in the original 3K DArTag panel.

Chromosome	Loci # identified in 95% of samples	Loci # in 3K DArTag panel	Loci ratio identified in 95% of samples to 3K DArTag panel
chr1.1	53	414	0.128
chr2.1	54	364	0.148
chr3.1	52	419	0.124
chr4.1	67	426	0.157
chr5.1	64	390	0.164
chr6.1	31	210	0.148
chr7.1	56	367	0.153
chr8.1	71	410	0.173
Total	448	3000	0.149

To evaluate genetic diversity that can be captured at each marker locus, average proportion of microhaplotype numbers per marker locus were estimated across the entire panel and within each of the 10 species. Across the entire panel, 1,922 (80.22%) marker loci had more microhaplotypes identified in bulked samples than single-plant samples (Supplementary Figure 2A). Because significant differences in missing rate were observed across the 10 species, the same calculation was applied to each species. Overall, 28.83% to 73.77% of the marker loci had more microhaplotypes identified in bulked samples than single-plant samples (Supplementary Figure 2B). Across all 10 species, the proportion of marker loci that had elevated number of microhaplotypes in bulked samples was strongly correlated ($r = 0.726$) with the sample size of a species (Supplementary Figure 3).

3.5 Genotypic population structure

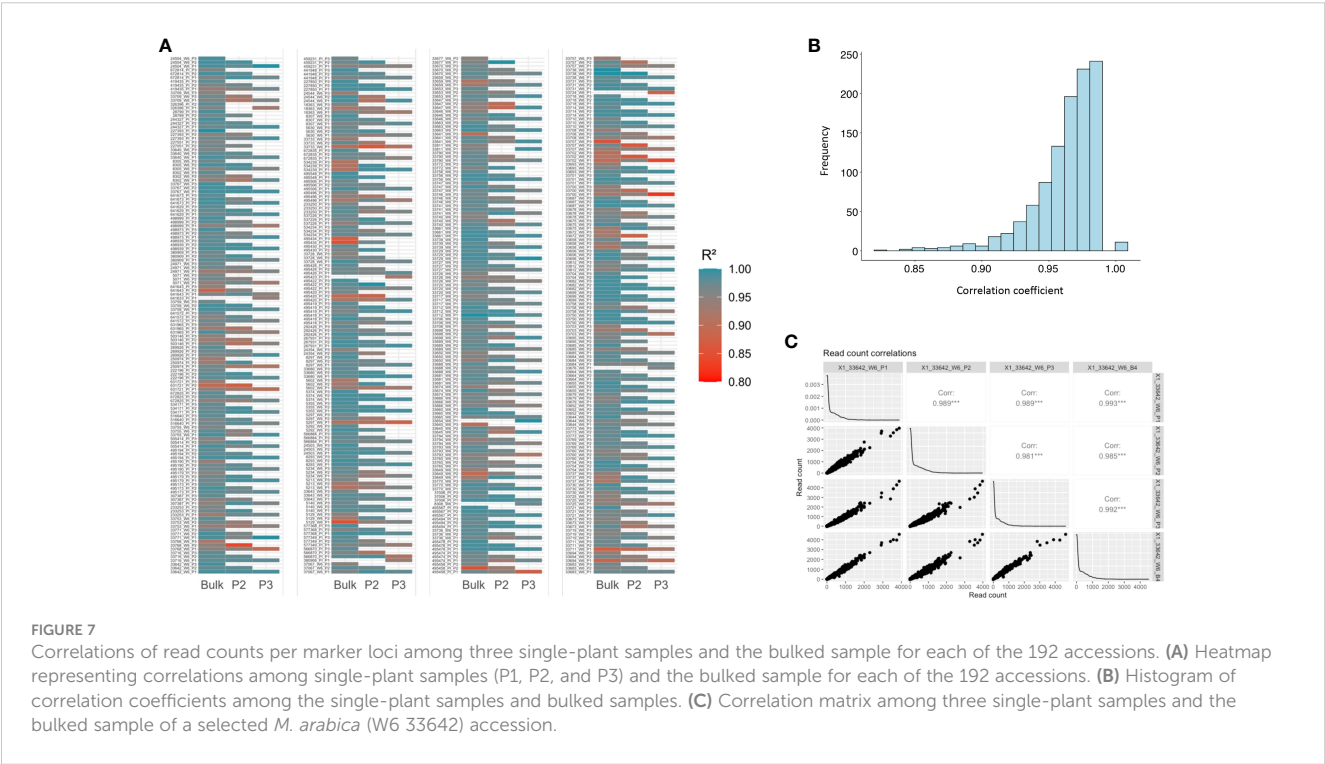
Population structure and relationship among the accessions underlying the observed genomic variations was investigated. To ensure a representative analysis, we selected a single plant replicate of each accession with the genome-wide highest average read depth (Supplementary Table 7) for further investigation. PCA was performed using dosage-based calls using both the entire set of marker loci and a subset of 448 marker loci that were evenly distributed across the genome and found in >95% of the genotyped accessions. PCA demonstrated distinct clustering patterns corresponding to different species (Figure 9A). Some accessions from different species exhibited close clustering, indicating higher genetic similarity among species, while others formed distinct clusters from the rest of the population, indicating greater genetic distinctiveness (Figures 9A, C). For instance, consistent with phenotypic PCA, *M. rigidula* accessions (30 accessions) and *M. rigiduloides* accessions (11 accessions) clustered nearby and had low F_{ST} value (0.10). Conversely, *M. polymorpha* accessions clustered distantly from *M. monspeliaca* (10 accessions), which exhibited high F_{ST} values (0.48) (Figure 9A). Generally, the accessions identified as the species clustered together regardless of the source of collection. Interestingly, we found five accessions grouped as outliers in unrelated species clusters. Two accessions labeled as *M. rigidula* (PI 672835 and W6 5630) were found to cluster with *M. rigiduloides*. Another labeled *M. rigidula* accession (PI 233250) was found to be closely associated with *M. praecox* accessions. Similarly, two *M. rigiduloides* accessions (W6 33753 and PI 37006) were found to cluster with *M. rigidula* accessions (Figure 9A). This indicates either a high similarity of the accession to other species, like an admixture, a possible mislabeling, or misidentification of the accession due to similar phenotypic characterization (Figure 3). Additionally, we successfully categorized most of the samples into distinct clusters representing various species by utilizing the 448 marker loci (Supplementary Figure 4). However, accessions belonging to species like *M. rigidula* and *M. rigiduloides* were not differentiated with the reduced set of markers, because these two species shared the most marker loci (126) and had very few unique markers (*M. rigidula*: 39 and *M. rigiduloides*: 2) (Figure 5). This indicates that the



identified common set of marker loci can be applied only for general purposes, such as determining population structure in general and assigning distinct *Medicago* species.

The phylogenetic tree constructed using the 2,396 SNP markers yielded results consistent with the PCA. We found that the accessions grouped into clades corresponding to their species, with high bootstrap values (>50) (Figure 9B). Furthermore, we observed related species grouping closer in the phylogenetic tree consistent with the PCA and F_{ST} values. For example, *M. minima* and *M. lupulina* were grouped relatively close in the PCA analyses

and found to arise from the same branch in the phylogenetic tree. They also had low F_{ST} values (0.35) (Figure 9C). Additionally, within the clade of the same species, *M. orbicularis*, accessions from the Crimean accessions formed separate branches, indicating genetic diversity within the clade (Figure 9B and Supplementary Figure 5). Notably, the *M. cretacea* accessions were not grouped with any other species, indicating their genetic distinctiveness from other species. It is worth noting that we only included two accessions of *M. cretacea*. Furthermore, consistent with PCA, we identified the same five accessions grouped in a different species



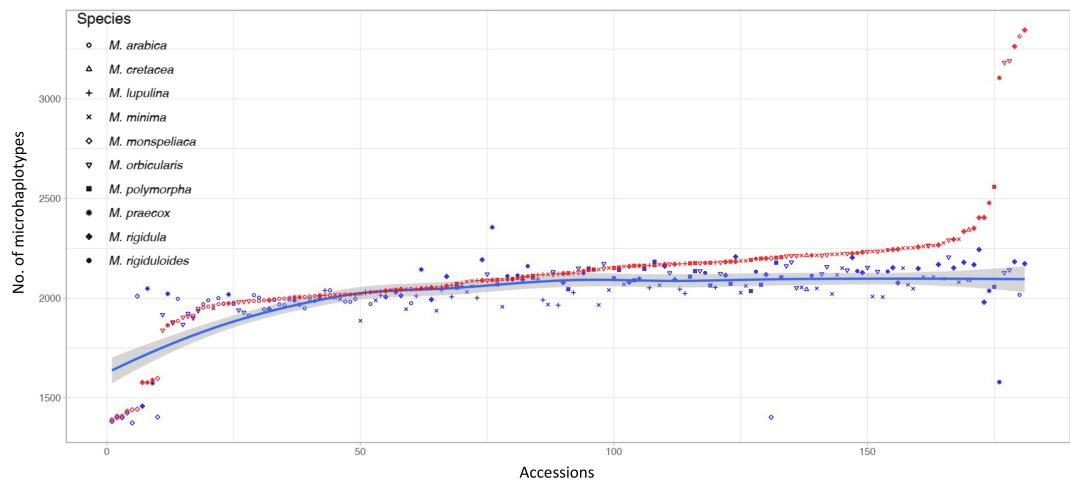


FIGURE 8 Scatter plot of sample-wise microhaplotype numbers in the single-plant (median of the three single-plant samples; blue dots) and bulked (red dots) samples for 181 accessions. The trend (blue curve) of microhaplotype numbers among single-plant samples was generated using a LOWESS regression.

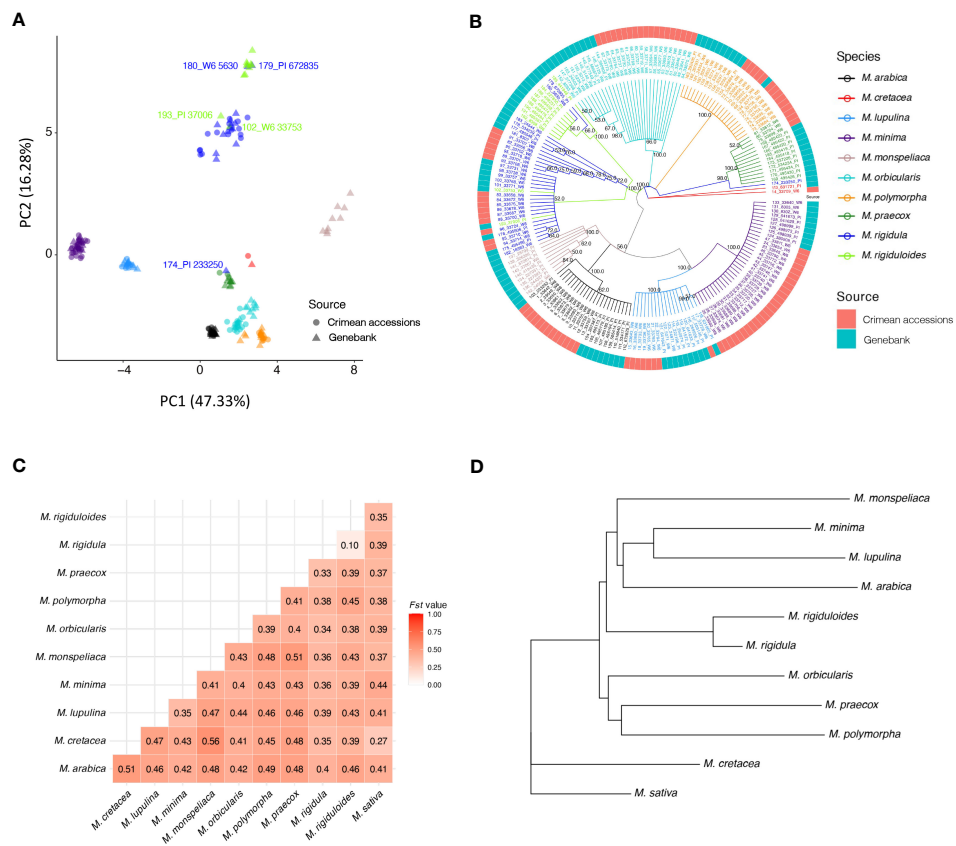


FIGURE 9 Population structure analysis of the accessions used in the study. (A) Principal component analysis using estimated allele dosages in 192 accessions. Possible mislabeled or misidentified accessions are listed in the figure. (B) Phylogenetic tree representing relationships among the 192 accessions. Different colors of accessions and branches represent unique species. The colors in the outer concentric circle represents the collection accession source. The numbers on the branches represent the bootstrap values (>50%). (C) Heatmap of pairwise F_{ST} values among all 10 species used in the study and *M. sativa* accessions from Zhao et al. (2023). (D) Simplified species tree representing genetic relatedness of the *Medicago* species under study based on the neighbor joining tree constructed using pairwise F_{ST} comparison with *M. sativa* as the root.

clade than their original labels. Moreover, the phylogenetic tree constructed from pairwise F_{ST} values (Figure 9D) also aligned with the other findings.

3.6 Crimea and Genebank accession comparison

To further evaluate the genetic diversity, we conducted a comparative analysis between accessions collected from Crimea and those reference accessions already in the NPGS. We observed a significant difference (p -value: 6.98×10^{-5}) in missing markers between the accessions collected in the Crimea and those collected from the NPGS collection (Supplementary Figure 6), indicating the presence of underlying genetic variation. Moreover, the PCA plot yielded a sub-clustering of accessions within species such as *M. orbicularis* based on their acquisition source (Figure 9A and Supplementary Figure 5). This grouping suggests the Crimea material added novel genetic diversity into the existing NPGS collection. Supporting this observation, we identified relatively high genetic dissimilarity for accessions collected for species like *M. orbicularis* ($F_{ST} = 0.130$) and *M. praecox* ($F_{ST} = 0.113$), which further underscores the impact of novelty in the collected germplasm (Table 3 and Supplementary Figure 5).

4 Discussion

In this study, we successfully characterized genetic diversity and population structure in a subset of annual medic accessions collected and integrated into the NPGS collections using both phenotypic and molecular markers. Annual medics are important forage legumes used extensively in regions of domestication, where they have naturalized, and where they have been introduced as crops like in Australia (Crawford et al., 1989; Nichols et al., 2012).

TABLE 3 Estimated F_{ST} between annual medic (*Medicago* spp.) accessions from Crimea and from the NPGS across different species.

Species	F_{ST} (Gene bank vs Collection)
<i>M. arabica</i>	0.003
<i>M. cretacea</i>	–
<i>M. lupulina</i>	0.013
<i>M. minima</i>	0.009
<i>M. monspeliaca</i>	–
<i>M. orbicularis</i>	0.130
<i>M. polymorpha</i>	0.044
<i>M. praecox</i>	0.113
<i>M. rigidula</i>	0.021
<i>M. rigiduloides</i>	–

“–” denotes absence of analysis either due to lack of samples in either of the group or very few number of accessions.

They are also related to alfalfa, and at least some have been used in breeding introgressing traits into this important crop (Irwin et al., 2010; Bingham, 2013; Humphries et al., 2021). Characterizing diversity in germplasm collections is an activity that helps inform curatorial management decisions while providing information to stakeholders. The genotyping research served as a proof of concept that DArTag markers developed in alfalfa could be used to genotype related annual medic species. The analyses conducted aided in the correct taxonomic identification and helped inform relationships between and among species when compared to a reference subset.

The findings of this study hold crucial significance for current agricultural practices and the production of annual medics. Through the comprehensive characterization of genetic diversity and population structure, we have provided valuable insights that can inform germplasm conservation strategies and breeding programs. The identification of unique genetic variants and their associations with specific phenotypic traits within the Crimean germplasm collection presents an opportunity to enhance the adaptability and productivity of annual medics and offer a foundation for targeted breeding programs. This knowledge is helpful for developing cultivars with improved traits, such as drought resistance, pest tolerance, and enhanced nutritional content, contributing to the sustainability and efficiency of forage legume production. The study also highlights the importance of preserving and exploring diverse germplasm collections to unlock hidden genetic potential, ensuring the continual improvement of annual medic crops.

Phenotypic traits describing phenology, plant, leaf/leaflet, flower, and fruit phenotypes were useful in showing differences among accessions within species, but clearly helped define the distinctions between species. Differences among accessions within species was greater in reference accessions when compared to those from Crimea and was likely due to their selection from a broad geographic area to be generally representative of each species. This was evident from PCA plots, too. In a few accessions, off-type plants were identified based on their phenotypes and eliminated from the average recorded descriptor trait. Also, in a few instances accessions appear to be mislabeled or misidentified with phenotypic descriptors matching other species being evaluated. In these situations where ‘errors’ were encountered, notes were taken and used for comparing to genotyping results. Also, when possible correct taxonomic assignment was assigned and amended in the NPGS Germplasm Resources Information Network (GRIN)-Global database.

In this research we utilized the recently published alfalfa 3K DArTag marker panel (Zhao et al., 2023) to genotype accessions to overcome the challenges often encountered when genotyping related species using marker panels developed for major crop species. Our findings demonstrated the applicability and effectiveness of the DArTag marker panel in genotyping the diverse annual medics, where targeted marker panels are unlikely to be available any time soon for these species. Despite the presence of a relatively high number of missing marker loci, the marker panel provided reliable genotyping results. The high correlation observed among single-plant samples and the bulked sample (Figure 7) for the same accession indicates the consistency and reproducibility of

the marker panel, suggesting that replicates of each accession may not be necessary due to the high reproducibility. Furthermore, we noted a high correlation between the genotyping results of the bulked and single-plant samples, indicating that the bulked sample can also be used to obtain reliable and consistent genetic information from the panel. However, if rare microhaplotypes are critical for the study, it is highly recommended to genotype multiple single-plant samples instead of bulking them into one sample, as indicated by our findings where we observed an increased count of microhaplotypes in the consolidated single-plant samples than the bulk (Supplementary Figure 1; Supplementary Table 6). If similar read coverage could be achieved, it is possible that rare alleles could be detected in single plant samples, resulting in the observation of more microhaplotypes when combining read counts from all three single-plant samples in comparison with their respective bulked samples. We observed slightly elevated genetic diversity in the bulked samples of seven plants compared with a representative sample (median of the three) of three individual plants (Figure 8). With increased read depths (~two times higher), the 3K DArTag panel was able to capture elevated genetic diversity in the combined single-plant samples (Supplementary Figure 1). Thus, genotyping several single-plant samples separately could likely allow for the discovery of rare alleles, but at a higher cost and with some potential “genotyping noise” from paralogous amplification. In summary, with similar read depths, the 3K DArTag panel was able to capture the increased genetic diversity in bulked samples compared with single-plant samples, and genotyping several single-plant samples and studying them together could increase the possibility of identifying rare alleles, thus, discovering more existing genetic diversity for a species.

This marker panel could allow the efficient assessment of genetic diversity, population structure, and taxonomic confirmation in other medic species. Larger collections exist in the NPGS for *M. polymorpha* (747), *M. orbicularis* (383), *M. minima* (382), and *M. lupulina* (289) [Source: GRIN-Global <https://npgsweb.ars-grin.gov/gringlobal>] where these types of evaluations would be valuable. Taxonomic misidentification or mislabeling between closely related taxa like *M. rigidula* and *M. rigiduloides* could also be addressed. Marker panel utility would increase when reducing the taxonomic complexity of the species being evaluated to increase the number of successfully amplified/sequence loci (reducing missing loci) in common within a species or closely related species. These types of evaluations could be structured in such a way that closely related taxa could be included as well as outgroup species like alfalfa.

The presence of missing marker loci can be attributed to various factors, including variations in primer sites and the potential presence of structural variations in the targeted regions. Structural variations, such as insertions, deletions, or duplications, can significantly impact gene expression and protein functions, leading to variations in phenotype (Alonge et al., 2020; Sapkota et al., 2023). Therefore, the observed missing marker loci are possibly associated with the phenotypic natural variation observed in the species being evaluated, providing valuable insights into the underlying genomic variation contributing to the diverse phenotypes within the collection. Moreover, a noteworthy

correlation was established between the frequency of missing marker loci and the degree of relatedness to *M. sativa*, for which the marker panel was originally developed. This correlation followed the anticipated pattern, with species that are evolutionarily closer to *M. sativa* exhibiting fewer missing marker loci, while those more distantly related displayed higher numbers (Figures 6, 9). This observation underscores the predictive value of the marker panel's efficiency based on a species' proximity to the panel's origin. Additionally, this also highlights the importance of considering species-specific genomic differences in marker development and genotyping analyses. In essence, the presence of missing marker loci not only sheds light on the potential genomic underpinnings of phenotypic diversity but also highlights the marker panel's utility as a tool for efficiently assessing genetic relatedness among species, offering valuable guidance for its application in diverse genetic studies. Future investigations are necessary to evaluate potential biological or technical causes behind species-specific variations in missing data rates among different *Medicago* species, aiming for a comprehensive understanding of this phenomenon.

Although previous studies have employed various marker platforms for assessing germplasm genetic diversity (Brummer et al., 1991; Kidwell et al., 1994; Musial et al., 2002; Maureira et al., 2004; Djedid et al., 2021; Emami-Tabatabaei et al., 2021), mapping (Brummer et al., 1993; Brouwer and Osborn, 1999; Kaló et al., 2000; Choi et al., 2004; Badri et al., 2011; Gorton et al., 2012) and establishing trait associations in breeding programs (Barcaccia et al., 2000; Brouwer et al., 2000; Tavoletti et al., 2000; Badri et al., 2011; Gorton et al., 2012) of *Medicago* spp., these approaches have inherent limitations. Marker techniques such as RAPD, RFLP and AFLP are challenging to execute and often lack reproducibility. Microsatellite or SSR markers incur high development costs and generally exhibit low throughput, surveying only a limited number of loci (Putman and Carbone, 2014). GBS, while advantageous in generating many SNP markers, encounters challenges, especially in bioinformatic analyses, particularly for polyploid organisms like alfalfa or those lacking reference genomes (e.g., annual medics), leading to resource-intensive analysis processes (Bhatia et al., 2013; He et al., 2014; Rajendran et al., 2022; Reyes et al., 2022). The alfalfa DArT panel, however, offers a promising solution, addressing many of the limitations associated with these methodologies. Its amplicon-based targeted genotyping approach, as opposed to the more complex and resource-intensive techniques, allows for efficient SNP and microhaplotype identification. Notably, it offers a balance between high-throughput marker generation and the practical considerations of cost and ease of execution. The amplicon-based targeted genotyping approach of the DArT panel, as demonstrated in this study, proves advantageous for diverse *Medicago* species, offering an effective solution for genetic diversity assessment and breeding program applications. This makes the DArT platform a valuable addition to the molecular toolkit for the breeders.

In addition to the successful genotyping of annual medics using the alfalfa 3K marker panel, our study also revealed valuable insights into the phenotypic and genotypic diversity within the Crimean germplasm collection. The collection, assembled from the

Crimean Peninsula of Ukraine, was specifically targeted to fill gaps in geographic coverage in the NPGS TFL collections. The diverse phenotypic traits observed in the collection highlight the richness of genetic resources that can be explored to use as crops, and for breeding and molecular biology applications. Our analysis revealed the presence of unique genetic diversity in accessions for certain species from the Crimean collection. These accessions exhibited distinct clustering patterns and genetic signatures, indicating their value in filling gaps in coverage in the existing collections, and as potential sources for breeding. The identification of such unique genetic variants within the collection also highlights the importance of preserving and studying diverse germplasm collections to unlock hidden genetic potential. Another important point to make was that a single accession of *M. cretacea*, a close alfalfa relative in the same taxonomic Section (*Medicago*), from Crimea and from the reference collections were included. That is because this is an underrepresented species in the collections that could be targeted for additional acquisitions.

Population structure analysis enables one to understand genetic diversity in each collection and identify the appropriate population for association mapping (Qiang et al., 2015). The population structure analysis conducted in our study played a crucial role in understanding the genetic diversity within species and identifying appropriate subpopulations for association mapping. The clustering of accessions based on both phenotype and genotype data generally agreed with the predicted evolutionary relationships among species. This consistency reinforces the reliability of the marker panel and its ability to capture the underlying genetic structure of the annual medics studied. Moreover, the clustering patterns also reflected the distinct genetic backgrounds of different species within the collection, providing important information for breeders and researchers interested in specific species or traits. Additionally, our study delved into the genetic relatedness and phylogeny among the various *Medicago* species and provided insights into the evolutionary relationships among the studied species. Notably, *M. cretacea* was found to be the closest, while *M. monspeliaca* was the furthest from *M. sativa*, a consistency observed in the respective number of missing marker loci within their accessions. Overall, both PCA and the phylogenetic tree analysis confirmed the species-level clustering and revealed additional insights into genetic relationships and diversity within and between species. While prior research has touched upon these aspects, the comprehensive inclusion of all these species in a single study is unique. Consequently, our study provides an overarching view of the genetic relatedness among the species in focus, addressing a pre-existing knowledge gap within the field.

The marker panel not only facilitated the characterization of genetic diversity but also helped in uncovering potentially misidentified, mislabeled or mixed-up samples within the collection. The similarity in phenotypes between many annual medic accessions and across different species can make it difficult to classify and assign taxonomy correctly. Conversely, a uniqueness in phenotypes of these few accessions allowed for individual accessions to be misclassified. By comparing the genotypic data with the known taxonomic identities, we were able to identify instances of misidentification or labeling errors, emphasizing the importance of reducing bias in sample management

within germplasm collections. Accurate sample identification is crucial for the integrity of germplasm collections, and the marker panel provided an efficient tool for detecting discrepancies despite being phenotypically very similar. This marker panel stands out among several tools available for identifying potential mislabeling or mix-ups in samples. This resource and information are invaluable for guiding breeders or germplasm curators to conduct further validation and confirmation of the labeling of these accessions. Correcting these identification errors is essential for ensuring the reliability and usability of the collection for future research and breeding efforts.

Moving forward, our findings set the stage for future studies that can delve deeper into the genetic architecture of annual medics. The marker panel could also lead to the possible identification of specific genomic regions associated with desirable traits that may be beneficial in these annual medic crops or as possible sources of desirable introgression in alfalfa (Irwin et al., 2010; Bingham, 2013; Humphries et al., 2021). This insight calls for extensive genome-wide association studies (GWAS) connecting phenotypic features with genetic markers. Such studies could unravel the intricate relationships between genotype and phenotype, identifying key genomic regions associated with desirable traits. Notably, the genomic regions or candidate genes underlying such associated loci could serve as prime targets for gene-editing research, with the advantage of having naturally occurring alleles as proof of concept. Integrating genomic information with both traditional breeding approaches and modern biotechnological tools can expedite the development of improved cultivars with tailored traits, reducing the time and resources required for breeding cycles and improving efficiency. Moreover, considering the existence of other crop-specific DArT panels, opportunities exist to replicate our marker panel evaluation approach in wild crop relatives across other vital crops. This exploration could shed light on the performance of marker panels, such as DArTag, in related species, providing insights into their adaptability and reliability. Given the slim probability of dedicated platforms for crop wild relatives, utilizing versatile tools like DArTag becomes crucial. For many germplasm curators, this becomes an invaluable starting point for their genotyping initiatives. Furthermore, this evaluation strategy aids in developing approaches to incorporate wild crop-related species into cultivated breeding programs by illustrating the proximity of wild species to their cultivated counterparts. In essence, our study not only has implications for annual medics but also sets a precedent for broader applications in crop genetics and breeding.

Overall, our study is the first to demonstrate the successful application of phenotypic and genotypic characterization (using the alfalfa 3K marker panel) efforts in assessing the genetic diversity and population structure of annual medics. Our study not only advances our understanding of the genetic diversity and population structure of annual medics but also provides practical implications for crop breeding, and germplasm curation. The panel's reliability, reproducibility, and ability to identify misidentified samples make it a valuable resource for future studies and breeding programs. The comprehensive understanding of phenotypic and genotypic diversity gained from this research lays the foundation for conservation strategies, possible targeted breeding efforts, and functional genomics studies in annual medics. By harnessing the genetic

resources present in the Crimean collection and leveraging the power of marker-assisted selection, researchers and breeders can accelerate the development of improved cultivars with enhanced traits, leading to advancements in forage legume agriculture.

Data availability statement

The data presented in the study are deposited in the figshare data repository, accessed through the link https://figshare.com/articles/dataset/Genotype_and_Phenotype_data/25458262. All phenotypic descriptor data and genotype data are provided in the **Supplementary Tables**. Phenotype data are available in **Supplementary Table S3** and genotype data are available as **Supplementary Table S8**. All phenotypic descriptor data are also uploaded to the GRIN-Global database and associated with accessions being evaluated.

Author contributions

DZ: Writing – original draft, Writing – review & editing, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization. MSa: Writing – original draft, Writing – review & editing, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization. ML: Writing – review & editing, Formal Analysis, Investigation, Software, Visualization. CB: Writing – review & editing, Conceptualization, Project administration, Supervision. MSh: Writing – review & editing, Conceptualization, Funding acquisition, Project administration, Resources, Supervision. SG: Writing – review & editing, Conceptualization, Resources. BI: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This

research was supported by the United States Department of Agriculture, Agricultural Research Service, appropriated research project 2090-21000-026-000-D and by Breeding Insight which is funded through Cooperative Agreements between USDA-ARS and Cornell (project numbers: 8062-21000-043-004-A, 8062-21000-052-002-A, and 8062-21000-052-003-A). The Crimean exploration was funded by the USDA ARS NPGS Plant Exchange Office.

Acknowledgments

We want to thank USDA ARS technical staff and in particular Ms. Estela Cervantes for hard work and dedication to this and many other efforts. We also want to thank Alexandre Afonin and Roman Roskov for managing the logistics of the collecting trip. Diversity Arrays Technology provided genotyping services.

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.2024.1339298/full#supplementary-material>

References

- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145–161. doi: 10.1016/j.cell.2020.05.021
- Badri, M., Chardon, F., Huguet, T., and Aouani, M. E. (2011). Quantitative trait loci associated with drought tolerance in the model legume *Medicago truncatula*. *Euphytica* 181, 415–428. doi: 10.1007/s10681-011-0473-3
- Barcaccia, G., Albertini, E., Rosellini, D., Tavoletti, S., and Veronesi, F. (2000). Inheritance and mapping of 2 n-egg production in diploid alfalfa. *Genome* 43, 528–537. doi: 10.1139/g00-017
- Barker, R. E., and Warnke, S. E. (2001). Application of molecular markers to genetic diversity and identity in forage crops in *Molecular breeding of forage crops: proceedings of the 2nd international symposium* (Springer Netherlands, Lorne and Hamilton, Victoria, Australia).
- Bhatia, D., Wing, R. A., and Singh, K. (2013). Genotyping by sequencing, its implications and benefits. *Crop improv* 40, 101–111.
- Bijlsma, R., and Loeschke, V. (2012). Genetic erosion impedes adaptive responses to stressful environments. *Evolution. Appl.* 5, 117–129. doi: 10.1111/j.1752-4571.2011.00214.x
- Bingham, E. (2013). Results using tetraploid *Medicago truncatula* cv. *Jemalong* in crosses with alfalfa. *Medicago Genet. Rep.* 13, 1–16. doi: 10.1002/csc.20274
- Blyton, M. D. J., Brice, K. L., Heller-Uszynska, K., Pascoe, J., Jaccoud, D., Leigh, K. A., et al. (2023). A new genetic method for diet determination from faeces that provides species level resolution in the koala. *bioRxiv*, 2023.2002.2012.528172. doi: 10.1101/2023.02.12.528172
- 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

- Brouwer, D. J., and Osborn, T. C. (1999). A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). *Theor. Appl. Genet.* 99, 1194–1200. doi: 10.1007/s001220051324
- Brummer, E. C., Bouton, J. H., and Kochert, G. (1993). Development of an RFLP map in diploid alfalfa. *Theor. Appl. Genet.* 86, 329–332. doi: 10.1007/BF00222097
- Brummer, E. C., Kochert, G., and Bouton, J. H. (1991). RFLP variation in diploid and tetraploid alfalfa. *Theor. Appl. Genet.* 83, 89–96. doi: 10.1007/BF00229230
- 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
- Choi, H. K., Kim, D., Uhm, T., Limpens, E., Lim, H., Mun, J. H., et al. (2004). A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. *Genetics* 166, 1463–1502. doi: 10.1534/genetics.166.3.1463
- Clark, L. V., Lipka, A. E., and Sacks, E. J. (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3: Genes Genomes Genet.* 9, 663–673. doi: 10.1534/g3.118.200913
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74, 829–836. doi: 10.1080/01621459.1979.10481038
- Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364
- Crawford, E. J., Lake, A. W. H., and Boyce, K. G. (1989). Breeding annual *Medicago* species for semiarid conditions in southern Australia. *Adv. Agron.* 42, 399–437. doi: 10.1016/S0065-2113(08)60530-1
- 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
- Djedid, I. K., Terzaghi, M., Brundu, G., Cicatelli, A., Laouar, M., Guarino, F., et al. (2021). Genetic diversity and differentiation of eleven *medicago* species from campania region revealed by nuclear and chloroplast microsatellites markers. *Genes* 13, 97. doi: 10.3390/genes13010097
- Dunnington, D. (2023). ggspatial: Spatial Data Framework for ggplot2. Available online at: <https://github.com/paleolimbot/ggspatial>.
- Emami-Tabatabaei, S. S., Small, E., Assadi, M., Dehshiri, M. M., and Mehregan, I. (2021). Genetic variation among Iranian *Medicago polymorpha* L. populations based on SSR markers. *Genet. Resour. Crop Evol.* 68, 1411–1424. doi: 10.1007/s10722-020-01071-7
- Fernandez, A. L., Sheaffer, C. C., Tautges, N. E., Putnam, D. H., and Hunter, M. C. (2019). *Alfalfa, wildlife, and the environment* (4630 Churchill St., #1 St. Paul, Minnesota 55126, USA: National Alfalfa and Forage Alliance).
- Feuerstein, U., Brown, A. H. D., and Burdon, J. J. (1990). Linkage of rust resistance genes from wild barley (*Hordeum spontaneum*) with isozyme markers. *Plant Breed.* 104, 318–324. doi: 10.1111/j.1439-0523.1990.tb00442.x
- Fisk, J. W., Hesterman, O. B., Shrestha, A., Kells, J. J., Harwood, R. R., Squire, J. M., et al. (2001). Weed suppression by annual legume cover crops in no-tillage corn. *Agron. J.* 93, 319–325. doi: 10.2134/agronj2001.932319x
- Fox, J., Friendly, G. G., Graves, S., Heiberger, R., Monette, G., Nilsson, H., et al. (2007). The car package Vol. 1109 (Vienna, Austria: R Foundation for Statistical Computing), p.1431. Available at: <http://www.r-project.org>, <http://socserv.socsci.mcmaster.ca/jfox/>.
- Gabriel, C. J. (1992). Managing global genetic resources: the US national plant germplasm system. *BioScience* 42, 201–203. doi: 10.2307/1311832
- Games, P. A., and Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. *J. Educ. Stat. I*, 113–125. doi: 10.3102/1076998600100211
- Gorton, A. J., Heath, K. D., Pilet-Nayel, M. L., Baranger, A., and Stinchcombe, J. R. (2012). Mapping the genetic basis of symbiotic variation in legume-rhizobium interactions in *Medicago truncatula*. *G3: Genes Genomes Genet.* 2, 1291–1303. doi: 10.1534/g3.112.003269
- 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, doi: 10.1155/2015/431487
- Greene, S., Hughes, S. J., Nair, R., Huguette, T., Aouani, M. E., Prosperi, J. M., et al. (2006). Wild accessions/populations in *The Medicago truncatula Handbook*. Available at: <http://www.noble.org/MedicagoHandbook>.
- Guzzon, F., Gianella, M., Giovannini, P., and Payne, T. S. (2022). Conserving wheat genetic resources in *Wheat improvement: food security in a changing climate* (Springer International Publishing, Cham), 299–318. doi: 10.1007/978-3-030-90673-3_17
- Hasan, N., Choudhary, S., Naaz, N., Sharma, N., and Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *J. Genet. Eng. Biotechnol.* 19, 1–26. doi: 10.1186/s43141-021-00231-1
- He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5, doi: 10.3389/fpls.2014.00484
- Helentjaris, T., King, G., Slocum, M., Siedenstrang, C., and Wegman, S. (1985). Restriction fragment polymorphisms as probes for plant diversity and their development as tools for applied plant breeding. *Plant Mol. Biol.* 5, 109–118. doi: 10.1007/BF00020093
- 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, e74612. doi: 10.1371/journal.pone.0074612
- Humphries, A. W., Ovalle, C., Hughes, S., del Pozo, A., Inostroza, L., Barahona, V., et al. (2021). Characterization and pre-breeding of diverse alfalfa wild relatives originating from drought-stressed environments. *Crop Sci.* 61, 69–88. doi: 10.1002/csc2.20274
- International Board for Plant Genetic Resources (IBPGR) (1991). *Descriptors for annual medics/Descripteurs pour Medicago annuelles* (Rome, Italy: International Board for Plant Genetic Resources), 33.
- Irish, B. M., and Greene, S. L. (2021). Germplasm collection, genetic resources, and gene pools in alfalfa in *In the alfalfa genome* (Springer International Publishing, Cham), 43–64. doi: 10.1007/978-3-030-74466-3_4
- Irwin, J. A. G., Armour, D. J., Pepper, P. M., and Lowe, K. F. (2010). Heterosis in lucerne testcrosses with *Medicago arborea* introgressions and Omani landraces and their performance in synthetics. *Crop Pasture Sci.* 61, 450–463. doi: 10.1071/CP10070
- Kaló, P., Endre, G., Zimanyi, L., Csanádi, G., and Kiss, G. B. (2000). Construction of an improved linkage map of diploid alfalfa (*Medicago sativa*). *Theor. Appl. Genet.* 100, 641–657. doi: 10.1007/s001220051335
- Kassambara, A. (2019). *Comparing groups: Numerical variables* Vol. Vol. 192 (Sydney, Australia: Datanovia).
- Kidwell, K. K., Austin, D. F., and Osborn, T. C. (1994). RFLP evaluation of nine *Medicago* accessions representing the original germplasm sources for North American alfalfa cultivars. *Crop Sci.* 34, 230–236. doi: 10.2135/cropsci1994.001183X003400010042x
- King, K. C., and Lively, C. M. (2012). Does genetic diversity limit disease spread in natural host populations? *Heredity* 109, 199–203. doi: 10.1038/hdy.2012.33
- Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Soft.* 25, 1–18. doi: 10.18637/jss.v025.i01
- Maureira, I. J., Ortega, F., Campos, H., and Osborn, T. C. (2004). Population structure and combining ability of diverse *Medicago sativa* germplasms. *Theor. Appl. Genet.* 109, 775–782. doi: 10.1007/s00122-004-1677-x
- Mizukami, Y., Kato, M., Takamizo, T., Kanbe, M., Inami, S., and Hattori, K. (2006). Interspecific hybrids between *Medicago sativa* L. and annual *Medicago* containing Alfalfa weevil resistance. *Plant Cell Tissue Organ Culture* 84, 80–89. doi: 10.1007/s11240-005-9008-8
- Muir, J. P., Ocumpaugh, W. R., and Butler, T. J. (2006). Winter harvests for annual forage medics in the southern great plains. *Forage Grazinglands* 4, 1–9. doi: 10.1094/FG-2006-0531-01-R5
- Mundt, C. C. (2002). Use of multiline cultivars and cultivar mixtures for disease management. *Annu. Rev. Phytopathol.* 40, 381–410. doi: 10.1146/annurev.phyto.40.011402.113723
- Musial, J. M., Basford, K. E., and Irwin, J. A. G. (2002). Analysis of genetic diversity within Australian lucerne cultivars and implications for future genetic improvement. *Aust. J. Agric. Res.* 53, 629–636. doi: 10.1071/AR01178
- Nichols, P. G. H., Revell, C. K., Humphries, A. W., Howie, J. H., Hall, E. J., Sandral, G. A., et al. (2012). Temperate pasture legumes in Australia—their history, current use, and future prospects. *Crop Pasture Sci.* 63, 691–725. doi: 10.1071/CP12194
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412
- Piano, E., and Francis, C. M. (1992). The annual species of *Medicago* in the Mediterranean region, ecogeography and related aspects of plant introduction and breeding 1992 in *Proceedings of the Xth International Conference of the EUCARPIA Medicago s Group*. (06466 Seeland, OT Gatersleben, Germany: EUCARPIA (European Association for Research on Plant Breeding)) 373–385.
- Postman, J., Hummer, K., Stover, E., Krueger, R., Forsline, P., Grauke, L. J., et al. (2006). Fruit and nut genebanks in the US National Plant Germplasm System. *HortScience* 41, 1188–1194. doi: 10.21273/HORTSCI.41.5.1188
- Putman, A. I., and Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol. Evol.* 4, 4399–4428. doi: 10.1002/ecs3.1305
- Putnam, D., and Meccage, E. (2022). Profitable alfalfa production sustains the environment in *Proceeding 2022 World alfalfa congress* (Davis, CA: University of California Davis), 14–17.
- 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, e0124592. doi: 10.1371/journal.pone.0124592
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. <https://www.R-project.org/>.
- Rajendran, N. R., Qureshi, N., and Pourkheirandish, M. (2022). Genotyping by sequencing advancements in barley. *Front. Plant Sci.* 13, doi: 10.3389/fpls.2022.931423
- Reyes, V. P., Kitony, J. K., Nishiuchi, S., Makihara, D., and Doi, K. (2022). Utilization of genotyping-by-sequencing (GBS) for rice pre-breeding and improvement: A review. *Life* 12, 1752. doi: 10.3390/life12111752

- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454
- Sapkota, M., Pereira, L., Wang, Y., Zhang, L., Topcu, Y., Tieman, D., et al. (2023). Structural variation underlies functional diversity at methyl salicylate loci in tomato. *PloS Genet.* 19, e1010751. doi: 10.1371/journal.pgen.1010751
- Small, E. (2011). *Alfalfa and relatives: evolution and classification of medicago* (Wallingford, UK: CABI), 727. doi: 10.1017/S0014479711001384
- South, A. (2017). rnatuarearth: world map data from natural earth. R package version 0.1.0 (Vienna, Austria: The R Foundation). Available at: <https://CRAN.R-project.org/package=rnatuarearth>.
- Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi: 10.1093/oxfordjournals.molbev.a040023
- Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Mol. Biol. Rep.* 1, 3–8. doi: 10.1007/BF02680255
- Tavoletti, S., Pesaresi, P., Barcaccia, G., Albertini, E., and Veronesi, F. (2000). Mapping the jp (jumbo pollen) gene and QTLs involved in multinucleate microspore formation in diploid alfalfa. *Theor. Appl. Genet.* 101, 372–378. doi: 10.1007/s001220051493
- Undersander, D., Cosgrove, D., Cullen, E., Grau, C., Rice, M. E., Renz, M., et al. (2021). *Alfalfa management guide* (5585 Guilford Road, Madison, WI 53711-5801 USA: John Wiley & Sons).
- Wickham, H. (2011). ggplot2 in *Wiley interdisciplinary reviews: computational statistics*, vol. 3, 180–185. doi: 10.1002/wics.147
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628
- Zhao, D., Mejia-Guerra, K. M., Mollinari, M., Samac, D., Irish, B., Heller-Uszynska, K., et al. (2023). A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.). *Genetic resources* 4, 55–63. doi: 10.46265/genresj.2023.8
- Zhu, Y., Sheaffer, C. C., and Barnes, D. K. (1996). Forage yield and quality of six annual *Medicago* species in the north-central USA. *Agron. J.* 88, 955–960. doi: 10.2134/agronj1996.00021962003600060019x



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied Sciences,
Norway

REVIEWED BY

Flemming Yndgaard,
Nordic Genetic Resource Centre (NordGen),
Sweden
Hesam Mousavi,
Inland Norway University of Applied Sciences,
Norway

*CORRESPONDENCE

Jenyne Loarca

✉ jloarca@wisc.edu

[†]These authors share senior authorship

RECEIVED 22 November 2023

ACCEPTED 26 February 2024

PUBLISHED 19 April 2024

CITATION

Loarca J, Liou M, Dawson JC and Simon PW
(2024) Advancing utilization of diverse
global carrot (*Daucus carota* L.) germplasm
with flowering habit trait ontology.
Front. Plant Sci. 15:1342513.
doi: 10.3389/fpls.2024.1342513

COPYRIGHT

© 2024 Loarca, Liou, Dawson and Simon. 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.

Advancing utilization of diverse global carrot (*Daucus carota* L.) germplasm with flowering habit trait ontology

Jenyne Loarca^{1,2*}, Michael Liou³, Julie C. Dawson^{2†}
and Philipp W. Simon^{1,2†}

¹Vegetable Crops Research Unit, United States Department of Agriculture, Madison, WI, United States,

²Department of Plant and Agroecosystem Sciences, University of Wisconsin-Madison, Madison,

WI, United States, ³Department of Statistics, University of Wisconsin-Madison, Madison, WI, United States

Biennial vegetable crops are challenging to breed due to long breeding cycle times. At the same time, it is important to preserve a strong biennial growth habit, avoiding premature flowering that renders the crop unmarketable. Gene banks carry important genetic variation which may be essential to improve crop resilience, but these collections are underutilized due to lack of characterization for key traits like bolting tendency for biennial vegetable crops. Due to concerns about introducing undesirable traits such as premature flowering into elite germplasm, many accessions may not be considered for other key traits that benefit growers, leaving crops more vulnerable to pests, diseases, and abiotic stresses. In this study, we develop a method for characterizing flowering to identify accessions that are predominantly biennial, which could be incorporated into biennial breeding programs without substantially increasing the risk of annual growth habits. This should increase the use of these accessions if they are also sources of other important traits such as disease resistance. We developed the CarrotOmics flowering habit trait ontology and evaluated flowering habit in the largest (N=695), and most diverse collection of cultivated carrots studied to date. Over 80% of accessions were collected from the Eurasian supercontinent, which includes the primary and secondary centers of carrot diversity. We successfully identified untapped genetic diversity in biennial carrot germplasm (n=197 with 0% plants flowering) and predominantly-biennial germplasm (n=357 with <15% plants flowering). High broad-sense heritability for flowering habit ($0.81 < H^2 < 0.93$) indicates a strong genetic component of this trait, suggesting that these carrot accessions should be consistently biennial. Breeders can select biennial plants and eliminate annual plants from a predominantly biennial population. The establishment of the predominantly biennial subcategory nearly doubles the availability of germplasm with commercial potential and accounts for 54% of the germplasm collection we evaluated. This subcollection is a useful source of genetic diversity for breeders. This method could also be applied to other biennial vegetable genetic resources and to introduce higher levels of genetic diversity into commercial cultivars, to reduce crop genetic vulnerability. We encourage breeders and researchers of biennial crops to optimize this strategy for their particular crop.

KEYWORDS

plant genetic resources, plant breeding, diverse germplasm, biennial crops, flowering habit, annual crops, crop resilience, crop wild relatives

Introduction

In sexually reproducing crop plants, flowering is essential for breeding and seed production. Flowering phenology has important implications for other agronomic traits, such as plant biomass, disease onset, fruit ripening, seed yield, marketability, and harvest window - as such, there has been great interest in characterizing crop germplasm collections for phenological flowering data, as has been done in lentil (*Lens culinaris* Medik.) (Tullu et al., 2008), turnip and rutabaga (*Brassica napus* L.) (Cruz et al., 2007), safflower (*Carthamus tinctorius* L.) (Elfadl et al., 2010), honeysuckle (*Lonicera caerulea* L.) (Gerbrandt et al., 2017), grape (*Vitis vinifera*) (Wolkovich et al., 2017), blueberry (*Vaccinium* species) (Campa and Ferreira, 2018), cassava (*Manihot esculenta* Crantz) (Silva Souza et al., 2020), olive (*Olea europaea* L.) (Belaj et al., 2020), peach (*Prunus persica* L.) (Atagul et al., 2022), and common bean (*Phaseolus vulgaris* L.) (Basavaraja et al., 2023).

In many commercial vegetable crops, premature initiation of flowering stem (bolting) or flower primordia (early-flowering), has a well-documented severe adverse impact on yield and quality, especially for vegetable crops with both annual and biennial lifeforms. Many plants that provide global human sustenance originated as wild annual plants that have been selected for delayed annual flowering or biennial flowering habit. As such, many commercial vegetables can have annual or biennial lifeforms, with annual lifeforms being adapted to warmer climates and requiring less time in cold exposure to flower, and biennials requiring more cold exposure (and a second season of growth) to induce flowering. Before bolting, these vegetables have edible tender shoots and/or nutritious storage organs. These vegetables have significant culinary and nutritional properties, as well as cultural, social, and economic value. They are encompassed by a few taxonomic families, including *Apiaceae* (root vegetables such as carrot and parsnip, as well as culinary herbs such as leaf parsley, cilantro, fennel, and caraway); *Alliaceae* (culinary vegetables such as onion, shallot, and leek); *Amaranthaceae* (with edible roots such as beet and leafy greens such as chard and spinach); and *Brassicaceae*, a family with many species used as vegetables due to highly diversified edible vegetative storage organs. With the exception of genus *Raphanus* (radish and daikon), most are from the genus *Brassica* and are dispersed across six species, including *B. oleracea*: leaves (cabbage, collard greens, kale), stems (kohlrabi), and buds (Brussels sprouts); *B. rapa*: root (turnip), seeds (field mustard), and leaves (Napa cabbage (Chinese cabbage), bok choy, and rapini); *B. napus*: root (rutabaga); and the culinary spice, mustard seed, from *B. nigra*, *B. juncea*, and *B. carinata*.

Premature bolting in many vegetable crops results in significant economic losses, due to coincidence with root lignification, production of bitter secondary metabolites, and limited yield potential as plant reserves are shuttled toward reproductive rather than vegetative growth. When these plants transition from their vegetative growth phase to their reproductive growth phase, they channel resources to the growth and development of the seed stalk; this in turn causes rapid deterioration of the vegetative tissues, which senesce and become unmarketable (Quiros, 1993). In vegetable crops where the economic product is the vegetative shoot tissue, bolting initiates biochemical changes that cause

edible vegetative shoot tissue to become unpalatable due to damage and hardening from senescence in lettuce (Rosental et al., 2021), Chinese cabbage (Yui and Yoshikawa, 1991; Wang et al., 2014; Jiang et al., 2023), celery (Quiros et al., 1987; Quiros, 1993), and spinach (Ribera et al., 2020), as well as to secretion of latex and bitter secondary metabolites in lettuce (Ciriaci et al., 2013) and spinach (Abe et al., 2014). In vegetable crops where the fleshy storage root is the economic product, bolting gives way to root lignification, preventing tap root thickening (Villeneuve, 2020), and rendering an inedible, woody, unmarketable product, thus causing serious economic losses to growers of carrot (Dowker and Jackson, 1975; Prohens and Nuez, 2008; Simon and Grzebelus, 2020), table beet (Holland and Dowker, 1969; Dowker et al., 1971; Goldman, 2004), onion (Khokhar et al., 2007; Hyun et al., 2009; Baldwin et al., 2014; Havey, 2018), and turnip (Nishioka et al., 2005).

Breeding for bolting resistance or toward delayed flowering has long been recognized as a solution to premature flowering in vegetables crops with annual and biennial lifeforms, and has been cited as a priority breeding objective in all of the aforementioned vegetables, as well as celeriac (Bruznican et al., 2020), semi-tropical beet (McGrath and Panella, 2018), and chard (Colley, 2017). Evaluation of flowering time in diverse germplasm collections of vegetable crops with annual and biennial lifeforms has resulted in the identification of accessions with delayed bolting or non-bolting genotypes in carrot (Tabor et al., 2016), arugula (Morales et al., 2006), coriander (Bashtanova and Flowers, 2011), spinach (Chitwood et al., 2016), lettuce (Jang et al., 2019; Lebeda et al., 2019), and caraway (Von Maydell et al., 2024), suggesting that this is a viable strategy to successfully identify germplasm with commercial potential for use in breeding programs. The longstanding recognition that biennial flowering habit has a demonstrated strong genetic underpinning means that breeding delayed-bolting or bolting-resistant cultivars is a powerful, economical, and achievable strategy to improve this critical trait.

Despite these success stories and the knowledge that landrace varieties harbor great genetic potential for beneficial traits that promote crop resilience, commercial crop breeders are reluctant to utilize genebank germplasm due to linkage drag, or the unintended coinheritance of undesirable alleles alongside a trait of interest (Zamir, 2001; Bohra et al., 2022). The perception is realistic that genebank accessions could be difficult to work with due to challenges related to desirable traits being in linkage with poor agronomic performance traits such as premature flowering; thus, lack of phenological flowering data remains a significant deterrent to utilization of plant genetic resources in many crop improvement programs (Dempewolf et al., 2017; Zamir, 2001; Bohra et al., 2022). This is an especially daunting prospect in breeding biennial crops, given that genetic gain is limited by cycle time, and biennial crops achieve at most one cycle per year. We are thus at risk of not utilizing important genetic diversity for other key traits that growers need, rendering crops more vulnerable to diseases, pests, and other environmental stresses. The impetus to characterize germplasm based on critical phenological flowering traits is a logical starting point to advance utilization and prioritization of biennial crop genetic resources.

Carrot (*Daucus carota* ssp. *sativus*) is a vegetable crop with a relatively recent domestication history (~900 years ago, to date), known for being a significant source of dietary fiber and provitamin A

carotenoids (Simon et al., 2019). Major primary domestication syndrome traits in root crops like carrot include biennial growth habit, the ability to form a fleshy storage root from secondary growth, and reduced lateral root branching (Macko-Podgórní et al., 2017; Ellison, 2019). Recent molecular studies have confirmed that domesticated carrots were derived from wild populations of Central Asian *D. carota* ssp. *carota*, also known as Queen Anne's Lace (Iorizzo et al., 2013). The emergence of biennial carrot plants from annual types was a consequence of human-mediated selection for maximal vegetative growth prior to reproduction (Goldman, 2004). Selection under the process of domestication after carrots arrived in Europe modified the life cycle of carrots from annuals to biennials, thereby ensuring a full summer season of vegetative growth without floral initiation. Whether consciously or unconsciously, European farmers and breeders leveraged this natural genetic adaptation, using carrot's large vegetative reserves as sustenance in colder climates. Selection for a larger taproot and short growing season in colder climates necessitated a biennial lifecycle for carrot to achieve maximum vegetative growth without reduction of consumer quality. Consequently, the biennial carrot was derived from its wild annual progenitor by prioritizing vegetative traits and eliminating annual reproductive growth. Biennial growth habit is critical for non-woody, succulent storage root development that can be used as a food crop, and was undoubtedly one of the first selected traits in the lineage that become domesticated carrot (Macko-Podgórní et al., 2017; Ellison et al., 2018; Ellison, 2019). As such, biennial growth habit is biologically linked to fleshy root storage, the hallmark domestication trait in carrots. However, as a group, annual cultivated carrots developed for subtropical and semi-arid regions are at least as fleshy as biennial carrots.

Broad variation in bolting and flowering initiation reflects the ecological adaptations of plants to their local climatic conditions (Lebeda et al., 2019). In carrot, vernalization time requirement is genotype-dependent (Wohlfeiler et al., 2021), and variation for time requirement has been reported within annuals and biennials, with annuals needing shorter periods of cold exposure (5°C or 41°F from 5 to 30 days) and biennial cultivars requiring longer period of cold exposure (11–12 weeks) to initiate floral stem elongation and flower morphogenesis (Linke et al., 2019; Wohlfeiler et al., 2022). 'Annual' refers to plants that flower without a vernalization requirement and in the first growing season, or first season in which the seed is planted. In nature, vernalization is a natural genetic adaptation to environments in which it is advantageous to delay flowering in favor of a period of vegetative growth. Energy reserves that accumulate in root tissues of biennial root crops during the first season of growth fuel reproductive structure development during the second season of growth. Without vernalization, an obligate biennial carrot may never flower.

The genetic control of carrot flowering is under extensive study. A region on the distal arm of chromosome 2 has been implicated by several independent studies as a likely target region during the course of carrot domestication. In this region, two overlapping selective sweeps (Grzebelus et al., 2014; Ellison et al., 2018) are in close proximity to the vernalization gene *Vrn1* (Alessandro et al., 2013) and a candidate domestication gene (DcAHLc1) involved in root tissue thickening (Macko-Podgórní et al., 2017). Furthermore, the candidate domestication syndrome gene (DcAHLc1) systematically differentiates wild and cultivated accessions, and it is hypothesized

that this gene is involved in the development of the carrot storage root, as the localization of the gene overlapped with one of the QTL for root thickening. In the most extensive investigation of carrot flowering-time regulation genes, 45 unigenes were identified (Ou et al., 2017), including three putative FLOWERING LOCUS (FLC) genes, which are known to delay or repress flowering in *Arabidopsis* (Michaels and Amasino, 2000). These putative FLC genes were also differentially expressed between wild carrots and domesticated carrots (Ou et al., 2017). Taken together, these studies confirm biennial growth habit is a genetically controlled trait that was a primary target during domestication (Alessandro et al., 2013; Ellison, 2019). More recently, it was found that post-vernalization day length does not influence carrot flowering (Wohlfeiler et al., 2022).

Over 13,400 *Daucus* accessions are conserved globally by 62 institutions (Allender, 2019), yet essential phenological data, such as flowering habit, is not available for many accessions. Phenological flowering data is critical to selecting locally adapted and commercially relevant germplasm to screen for a breeding program. For example, in semi-arid and subtropical climates, temperatures rarely achieve the sustained lows required for vernalization, and large-volume refrigerated coolers for vernalization are unavailable (Simon & Grzebelus, 2020). As such, carrots cultivated in this region require late-flowering annual habit; some carrot plants are used to produce the root crop, and some generate the seed stock later in the same season. Biennial plants are unsuited to globally warm regions as they cannot contribute to the seed crop in subtropical/semi-arid markets. In contrast, cultivated carrots grown commercially in temperate climates, such as Europe, North America, and Australia, are of obligate-biennial stock (Goldman, 2004). Flowering at any time in the first season of growth (annual-flowering habit, whether early-flowering or late-flowering) is problematic and intolerable in temperate climates of commercial carrot root production (Rubatzky et al., 1999; Prohens & Nuez, 2008). As with other biennial vegetable crops, the transition from the vegetative to reproductive phase in carrot coincides with rapid lignification of the xylem, even before the floral stalk/bolting stem elongates, rendering the roots fibrous and inedible, and resulting in complete loss of consumer quality and commercial value (Peterson, 1986; Amasino, 2005; Alessandro & Galmarini, 2007; Ou et al., 2017; Linke et al., 2019; Simon et al., 2019). As such, biennial carrots are required in commercial carrot root production in temperate climates, as annual flowering habit results in complete loss of commercial root crop value and significant economic loss to the grower. For this reason, breeders of temperate carrot routinely select against annual flowering habit (Goldman, 2004).

Few global cultivated carrot germplasm collections have been evaluated for flowering habit or bolting tendency. High broad sense heritability was estimated for bolting tendency among 48 open-pollinated carrot varieties of European and Asiatic origin, studied in India (Manikanta et al., 2018), suggesting genetic potential for improvement. In an evaluation of a carrot germplasm collection (101 accessions) in China, purple rooted accessions demonstrated 48.4% premature bolting tendency, compared with 2.7% - 7% in orange rooted accessions (Bao et al., 2010). This likely reflects breeding efforts toward biennial flowering habit in orange-fleshed roots rather than true genetic linkage of anthocyanin with annual flowering habit. An assessment of 140 U.S. commercial carrot cultivars noted negligible amounts of bolting in this panel (Luby

et al., 2016). Similarly, few wild carrot germplasm collections have been evaluated for flowering habit or bolting tendency. A recent study of 14 wild Nordic carrots (*Daucus carota* subsp. *carota*) found that sowing time had a strong influence on flowering time, with earlier sowings resulting in increased annual behavior (Solberg and Yndgaard, 2015). A similar study of 10 wild carrot accessions found high inter- and intra-accession variation for flowering time in multi-environmental trials, with percent-flowering having a significant location effect and an insignificant genotype effect (Geoffriau et al., 2019), suggesting low genetic diversity for flowering in this population, high environmental influence, or both. The range of results reflects the variation in flowering habit across carrot germplasm collections.

Whereas premature bolting is an irredeemable trait in commercial crop production, it is possible and necessary to disentangle the undesirable early-bolting accessions from desirable late-flowering annuals and biennials to optimize use of genetic resources. Given the critical role of plant genetic resources in crop resilience, it is important to have the ability to use a wider range of genetic resources for resistance to emerging diseases, pests and environmental stresses. There may be many genetic resources suitable to various global production environments and market needs which are not being used due to the perception that they will bring in undesirable flowering habits. As such, characterization of flowering phenology has great potential to increase engagement with plant genetic resources, increase levels of genetic diversity in commercial crop cultivars, and reduce genetic vulnerability to shifts in production conditions for these nutritionally, economically, culinarily, and culturally important biennial vegetables.

Materials and method

Population under study

Daucus accessions (N=1381) are maintained through the U.S. National Plant Germplasm System (NPGS) at the North Central Regional Plant Introduction Station (NCRPIS) in Ames, IA, with information on the accessions (also known as genotypes or plant introductions) in the Germplasm Resources Information Network (GRIN) database of the NPGS (GRIN-Global, 2023). Each carrot accession is a genetically unique, heterogeneous, heterozygous population. Accessions were selected from the GRIN system for our diversity panel if passport information suggested the presence of domestication traits (N=695). These cultivated carrots represent global carrot germplasm, collected over multiple plant exploration trips between 1947 and 2015 from 60 countries, with over 80% of accessions originating from the Eurasian supercontinent: 53% from Asia, 34% from Europe and the Caucasus, and 13% (in descending order) collected from the Americas, Africa, Australia, and New Zealand. This collection includes 148 total accessions from the primary center of diversity in central Asia (modern-day Afghanistan and surrounding countries) and secondary center of diversity in western Asia (modern-day Turkey) (Vavilov, 1951; Banga, 1957). This collection includes landraces and heirloom cultivars with annual, biennial, or mixed flowering habits. Although biennial flowering habit

is a known domestication trait in carrot, reliable accession flowering habit data was not available prior to this study. GRIN-Global maintains flowering habit data for each accession ('lifecycle'), but this data is not reliable due to being recorded in many different environments and on variable numbers of plants. As such, this data was not a criterion for identifying domesticated germplasm in this evaluation. This study is the first and largest (N = 695 accessions) multi-year field evaluation of flowering habit in a diverse carrot germplasm that includes landraces. The carrot accessions in this study are maintained by the United States Department of Agriculture National Plant Germplasm System (USDA-NPGS). All or parts of this global USDA germplasm collection have previously been evaluated in studies on canopy vigor (Loarca et al., 2024b), core collection curation (Corak et al., 2019), demographic history of carrot domestication and breeding (Coe et al., 2023), genetic structure, phylogeny, and carotenoid presence (Ellison et al., 2018), taproot shape (Brainard et al., 2021), plant growth traits (Acosta-Motos et al., 2021), antioxidant capacity (Pérez et al., 2023), resistance to the necrotrophic fungal pathogen *Alternaria dauci* (Tas, 2016), and several studies on seed germination under abiotic stress (Bolton et al., 2019; Bolton and Simon 2019; Simon, 2019; Simon et al., 2021).

Experimental design

In 2016-2018, one plot of appx. 50 seeds from each accession (N=695) were hand-planted in each of two blocks of a randomized complete block design (RCBD) at the Hancock Agricultural Research Station (ARS), located in the central sands region of Wisconsin. Bed preparation and planting methods are described in detail in the companion paper of the present study (Loarca et al., 2024a). In each year, we collected flowering data on 679 - 695 accessions. With data over multiple years, we have characterized flowering habit on 668 accessions.

Trait phenotyping

Shoot-growth phenotyping methodology is provided in greater detail in the companion paper of the present paper (Loarca et al., 2024a). Flowering habit is a trait that can be assigned to a single plant. Because carrot growth, including the initiation of flowering, can vary widely in carrot, and gene bank accessions are often highly heterozygous, plants from the same accession may express different flowering habits. As such, we phenotyped flowering on a plot-level basis by estimating the percentage of plants showing signs of flowering (Figure 1). We scored plots on a 5-point scale (0%, 25%, 50%, 75%, 100%) based on the percentage of plants within each plot with signs of flowering visible to the naked eye, such as stem length greater than 8mm (Villeneuve, 2020) and/or presence of flower primordia, which vary morphologically by genotype and maturity and will require some practice and training to visually identify (Figure 1).

Given the undesirability of annual plants for biennial breeding programs, there has been historically no systemic evaluation of flowering habit before end of season (100 DAS in our study). However, it is known that among annual flowering carrots, wild germplasm tends to flower much earlier than cultivated annual germplasm (Simon & Grzebelus, 2020). For this reason, we also scored flowering at 60 DAS, which was the earliest time point at which we observed signs of flowering (Figure 1) in 5% - 10% of plots



FIGURE 1
Various expressions of carrot flower primordia at 60 DAS.

in two of our three studies, and also at 100 DAS (harvest day or end of season). Phenotypic data are stored at the CarrotOmics database (www.carrotomics.org/) (Rolling et al., 2022).

Data management

We used RStudio Version 2023.6.1.524 (Posit team, 2023) and R Version 4.3.1 (R Core Team, 2023) to perform all statistical analyses. Rosner's Test in the EnvStats identified multiple simultaneous potential outliers for each trait in each year (Millard, 2013). Various utility packages were crucial to our analysis, such as ggthemes (Arnold, 2021), beepr (Bååth, 2018), flextable (Gohel and Skintzos, 2023), and the tidyverse suite of packages (Wickham et al., 2019).

Two-way analysis of variance & broad-sense heritability estimation

F-tests of significance were performed to identify significant sources of variation for each trait in each year using fixed effects models in a two-way Analysis of Variance (ANOVA) with Type III sums of squares with the *car* package (Fox et al., 2012/). For each year, a fixed effects model was structured to calculate the proportion of variance in each trait (percentage of flowering plants in plot at 60 DAS or 100 DAS) attributable to genotype and block: $T_{ik} = u + g_i + b_k + e_{ik}$, where T_{ik} = phenotype measured on the trait of interest, u = intercept, g_i = genotype, b_k = block, and e_{ik} = error with $e_{ik} \sim \text{i.i.d. } N(0, \sigma^2)$.

The multi-year fixed effects model includes accessions with trait data across all years and accounts for variation across years: $T_{ijk} = u + g_i + y_j + (gy)_{ij} + b_{k(j)} + e_{ijk}$, where T = phenotype of the trait of interest, g_i = genotype, y_j = year, $(gy)_{ij}$ = genotype*year interaction,

$b_{k(j)}$ = block within year, and e_{ijk} = error with each component assumed to follow a respective, independent normal distribution [$e_{ijk} \sim \text{i.i.d. } N(0, \sigma^2)$]. Due to unbalanced data from abnormal weather events (destructive hail), we ran two multi-year analyses: one that included the 2017 flowering data and one that excluded the 2017 flowering data.

Variance components (V) for each trait were estimated using within-year (single-year) and across-years (multi-year) random effects models with the *lme4* package (Bates et al., 2015). These used the same model as above with all effects random. Broad-sense heritability (H^2) for each trait, within years (single-year model) and across years (multi-year model), was estimated from variance components, including genotypic variance (V_g) and phenotypic variance (V_p). Single-year broad-sense heritability (for each year 2016–2018) was calculated for each trait:

$$H^2 = \frac{V_g}{V_p} = \frac{V_g}{V_g + \frac{V_{error}}{\#reps}}$$

Multi-year broad-sense heritability was estimated for each trait:

$$H^2 = \frac{V_g}{V_p} = \frac{V_g}{V_g + \frac{V_{gy}}{\# \text{ years}} + \frac{V_{error}}{\# \text{ years} * \#reps}}$$

Mixed models and estimated marginal means

Accessions were categorized into flowering habit based on the estimated marginal means of their percentage flowering at 60 DAS and 100 DAS. We used the same model terms above in a mixed model using the *lme4* package (Bates et al., 2015), with genotype as fixed effect, to extract estimated marginal means for flowering percentage for each accession within and across years with the *emmeans* package (Lenth, 2023). Estimated marginal means on flowering-percentage, within and between years, were used to assign

accessions to flowering habit categories based on the 2016 & 2018 data sets.

Flowering habit ontology

As of 2023, GRIN-Global describes three categories for carrot flowering habit (or 'life form' in the GRIN system): annual, biennial, or a mixture (of annual and biennial plants), however, it was not clear what criteria or thresholds were used to categorize accessions in to one of the three flowering categories. In absence of this information, we created a theoretical construct ('2023 GRIN-Global Flowering Habit' in Figure 2, right panels) with these three traditional flowering habit categories: annuals had 100% flowering at 100 DAS, biennials had 0% flowering at 100 DAS, and mixtures had between 1% and 99% flowering at 100 DAS.

Starting with these assumptions, we made minor threshold adjustments for the CarrotOmics Flowering Habit Ontology presented in this paper. Given that annual flowering habit is intolerable in temperate regions of carrot root production, we maintained the 0% flowering at 100 DAS threshold for the biennial category. We characterized annual-flowering plants as plots where the vast majority (85% - 100%) of plants were flowering at 100 DAS – despite this slight allowance for non-flowering plants, these are not considered mixed populations. Mixtures had between 1% and 85% of plants flowering at 100 DAS (Figure 2, right panels).

Using this three-category ontology as a baseline, we then subdivide it into different types of annuals and mixtures. We

added an additional time-point (60 DAS) for evaluating flowering habit to capture precocity and uniformity of flowering among annuals (Figure 2, left panels). We designated annual accessions with 85% - 100% at 60 DAS as *uniformly-early annuals*. Similarly, we designated *uniformly-late annuals* as those with uniform-flowering by the end of the season, with no or low flowering at 60 DAS ($\leq 15\%$) and high amounts of flowering (85% - 100%) at 100 DAS. The remaining annual accessions flower non-uniformly and sporadically between 60 DAS and 100 DAS – these accessions may contain various proportions of uniformly-early and uniformly-late plants, all of which ultimately flower (85% - 100%) by end of season (100 DAS). As described above, mixtures have both biennial and annual plants (1% - 85% flowering at 60 DAS and 100 DAS). Recognizing the need for biennial germplasm for breeding and root production in temperate climates, we partitioned some low-flowering mixtures into a predominantly biennial subcategory, which we characterize as having between 1% and 15% flowering at 60 DAS and 100 DAS (Figure 2, left panels). This flowering trait ontology (Table 1 and Figure 3) is stored in the CarrotOmics database (www.carrotomics.org/) (Rolling et al., 2022).

Trait correlations

Grouped by flowering habit, we evaluated correlations between vegetative growth traits, including seed viability, seed weight, stand count, and canopy height. Pearson correlations were calculated and

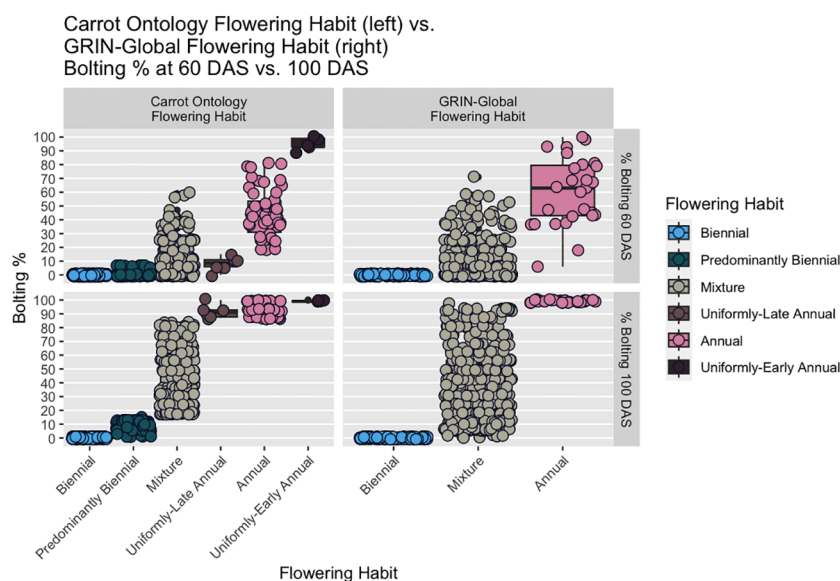


FIGURE 2

Boxplots of proposed CarrotOmics flowering habit trait ontology (left) vs. 2023 GRIN-Global flowering traits (right) compared with at 60 DAS and 100 DAS. In the new CarrotOmics trait ontology (left), uniformly-early annuals are defined as plots with $>85\%$ flowering at 60 DAS and 100 DAS; uniformly-late annuals are plots with $\leq 15\%$ flowering at 60 DAS and $>85\%$ flowering at 100 DAS; and the remaining non-uniformly flowering annuals are plots with $<15\% < \% \text{ flowering at } 60 \text{ DAS} < 85\%$ and $85\% \leq \% \text{ flowering at } 100 \text{ DAS}$. Mixtures (annual and biennial plants) are characterized as $15\% < \% \text{ flowering at } 60 \text{ DAS} < 85\%$ and $15\% < \% \text{ flowering at } 100 \text{ DAS} < 85\%$. In the GRIN-Global panels (right), annual is defined as 100% flowering at 100 DAS.

TABLE 1 CarrotOmics flowering habit trait ontology developed in this paper, compared with current trait descriptions in GRIN-Global. This paper elaborates on traits that were previously recognized as important in CarrotOmics and provides standard methodologies that carrot researchers can follow, enabling collaboration across programs.

CarrotOmics Trait Ontology for Flowering Habit	2023 GRIN-Global Flowering Trait Descriptors
Percent flowering (60, 100)	Percent bolt 1st Year
Percentage of flowering plants within a plot in the first planting season (AKA “first year”) Measured directly from field plots on a 5-point scale (0%-100% in increments of 25%).	Percent bolt in the 1st Year
Annual flowering habit (60, 100)	life cycle. AN=Annual.
During the first growth season, estimated greater than 85% of plants flowering within the plot.	Annual flowering habit. Plant will flower without a vernalization requirement
i. Uniformly-early annuals (60)	
Greater than 85% of plants within plot with signs of flowering at mid-season.	Early flowering field
ii. Uniformly-late annuals (60, 100)	-
Fewer than 15% of plants within plot with signs of flowering at mid-season and greater than 85% of plants within plot flowering end-of-season. Note: Potentially useful commercial germplasm for subtropical markets.	-
iii. Annuals (non-uniform flowering) (60, 100)	-
Greater than 15% flowering plants within the plot at mid-season and greater than 85% flowering plants within the plot at end-of-season. Note: This is not considered a mixture.	-
Biennial flowering habit (60, 100)	life cycle. BI = Biennial
0% plants within plot flowering during the first growth season. Roots require vernalization to induce flowering in the second growth season. Note: Potentially useful commercial germplasm for temperate markets.	-
Mixed flowering habit population (mixture) (60, 100)	life cycle. MX = Mixed
Both annual (uniformly-early and uniformly-late) and biennial plants in various proportions (1% ≤ % flowering< 85%) in the first season of growth. Annuals flower in the first growth season, while biennials’ roots require vernalization in order to flower in the second season of growth.	Mixed population of annual and biennial plants
Predominantly biennial (60, 100)	-
Type of mixture. Greater than 0% and less than 15% flowering plants in the plot at end of season. Annuals flower in the first growth season, while biennials’ roots	-

(Continued)

TABLE 1 Continued

CarrotOmics Trait Ontology for Flowering Habit	2023 GRIN-Global Flowering Trait Descriptors
require vernalization in order to flower in the second season of growth. Note: Potentially useful commercial germplasm for temperate markets.	

Data collection times may vary by location, cultivar, market type, and length of growing season. Refer to Figure 3 for logical flowchart of flowering habit ontology.
Data Collection Time (DAS).
“-” indicates no trait descriptor available.

smoothed trend lines were visualized in a correlation matrix using GGally, along with boxplots and scatterplots.

Results

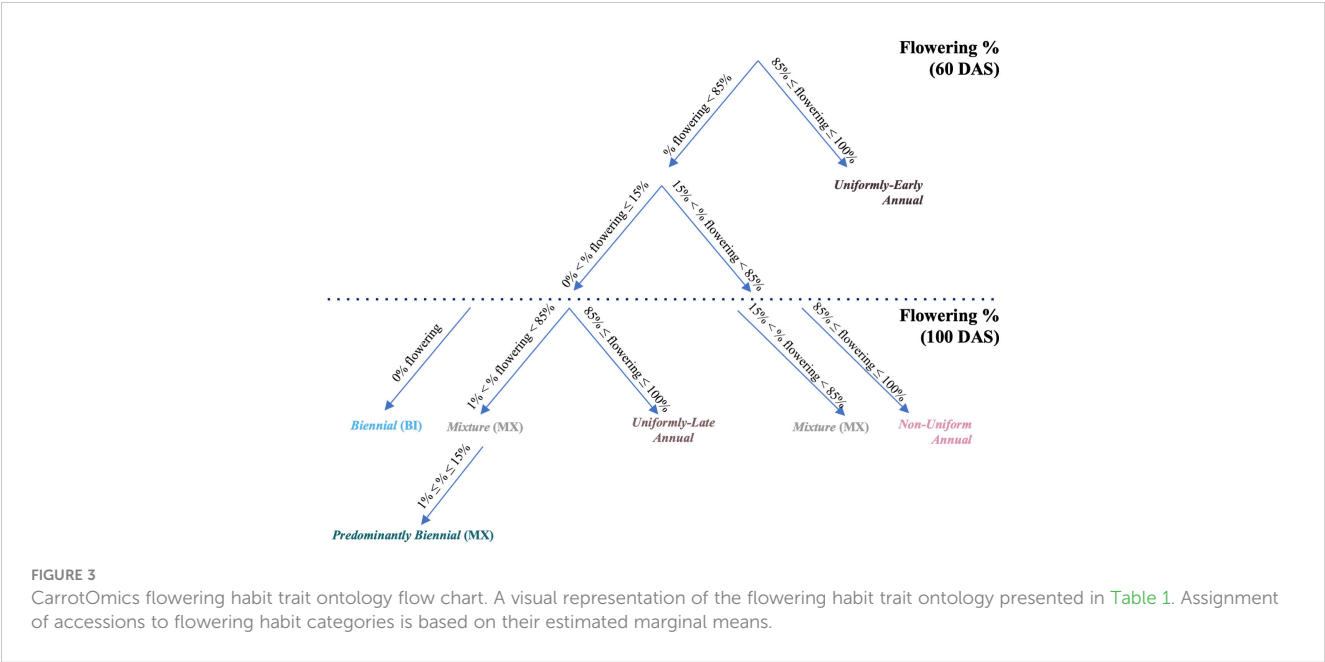
Summary statistics of the collection using GRIN-Global flowering habit

The GRIN-Global three-category model characterized biennials (29.5%; n=197) and annuals (9.1%; n=61), with mixtures (61.4%; n=410) representing the vast majority of the collection (Table 2A). By 100 DAS, 70.5% of the collection expressed some proportion of flowering. These mixtures are not well-characterized except that accessions in this category contain both annuals and biennials in various proportions.

Summary statistics of the collection using CarrotOmics flowering habit ontology

Summary statistics are presented using our proposed new ontology to characterize the collection at both the 60 DAS and 100 DAS data (Table 2B). Flowering at 60 DAS was the earliest time point at which we observed signs of flowering (Figure 1) in 4%-10% of plots in two of our three studies. Flowering thresholds and subcategories did not change the number of biennials identified (n=197), but did increase the number of annuals (10% of accessions; n=61) with fewer than 2% being uniformly-early annuals (n=5) and uniformly-late annuals (n=6). In every year of our evaluation, uniformly-early annuals are a distinct group from uniformly-late annuals at 60 DAS, and both are distinct from the non-uniform flowering annuals (n=50) (Figure 4). These remaining annuals flowered non-uniformly with 15% - 85% flowering at 60 DAS and 100 DAS.

While flowering is typically measured at 100 DAS, the addition of the 60 DAS time-point enabled differentiation between uniformly-early annuals and uniformly-late annuals, a distinction that would otherwise be lost by 100 DAS, as indicated in every year of our study (Figure 5). The predominantly biennial subcategory accounts for 24% of accessions (n=160) in the germplasm collection. Flowering percentage among low-flowering mixtures is indiscernible from predominantly biennial populations at 60 DAS, but distinct at 100 DAS. After partitioning the predominantly



biennial population, there are 250 mixed-flowering accessions, which have between 15% and 85% annual plants at 100 DAS.

Analysis of variance & broad-sense heritability

ANOVA results indicate that genotype was a highly significant factor influencing flowering percentage at 60 DAS (Table 3A), with broad-sense heritability (H^2) estimates also consistently very high in each year (Table 2A: $0.87 < H^2 < 0.93$). The block effect was not significant in 2016 or 2018 significant at the $p < 0.05$ level and in 2017. Similarly, F-tests of significance at 100 DAS flowering had a highly significant genotype effect and high broad-sense heritability (Table 3B: $0.81 < H^2 < 0.84$). Genotype and genotype x year interaction are highly significant factors across all three years in the multi-year ANOVA (Table 4A). Multi-year broad-sense heritability (H^2) is moderate at 60 DAS ($0.57 < H^2 < 0.64$) and high at 100 DAS. ($0.88 < H^2 < 0.89$). Excluding 2017 data did not substantially change heritability estimates or which factors were considered significant (Table 4B). P-values for flowering-percentage ANOVA results are available in Supplementary Tables 1-4.

Trait correlations

Using the strictest definitions for annual (100% flowering) and biennial (0% flowering), we explored vegetative trait correlations among the three traditional flowering habit categories (Figure 6A). Correlations among shoot-growth traits vary by flowering habit category. Correlation between seed viability and emergence was high for biennials ($r = 0.68$) and mixtures ($r = 0.66$) and very low and not significant for annuals ($r = 0.26$). Seed viability and seed weight had low negative correlation in biennials and mixtures, and was uncorrelated in annuals. Correlation between seed viability and late-season canopy height is moderate and positive for annuals ($r = 0.56$) and very low for biennials ($r = 0.17$) and mixtures ($r = 0.16$). Across all three flowering habits, emergence has low correlation with canopy height (80 DAS) and no correlation with seed weight. There also appeared to be a difference in mean and variation between biennial and annual accessions for canopy height. In the simple three-category model, mixtures tend to behave more similarly to biennials than to annuals, sharing very similar correlations among trait pairs, with one exception: the correlation between seed weight and canopy height was slightly higher for mixtures and annuals ($r = 0.41$) than for biennials ($r = 0.25$). There

TABLE 2A Summary statistics based on 2023 GRIN-Global flowering habit categories (biennial, annual, mixture) at 60 DAS and 100 DAS for all accessions planted 2016-2018.

Flowering Habit	2016		2017		2018		2016 & 2018 ⁱ	
	N=679	%	N=695	%	N=681	%	N=668	%
Biennial	246	36.23	477	68.63	297	43.61	197	29.49
Mixture	373	54.93	172	24.75	307	45.08	410	61.5
Annual	60	8.84	46	6.62	77	11.31	61	9.1

N represents the raw number of accessions in each flowering habit category; percentage (%) indicates the proportion of accessions in the germplasm collection in each flowering habit category. These categories define biennial as 0% flowering, annual as 100% flowering, and mixed as between 1%-99% flowering.

TABLE 2B Summary statistics based on CarrotOmics flowering habit trait ontology.

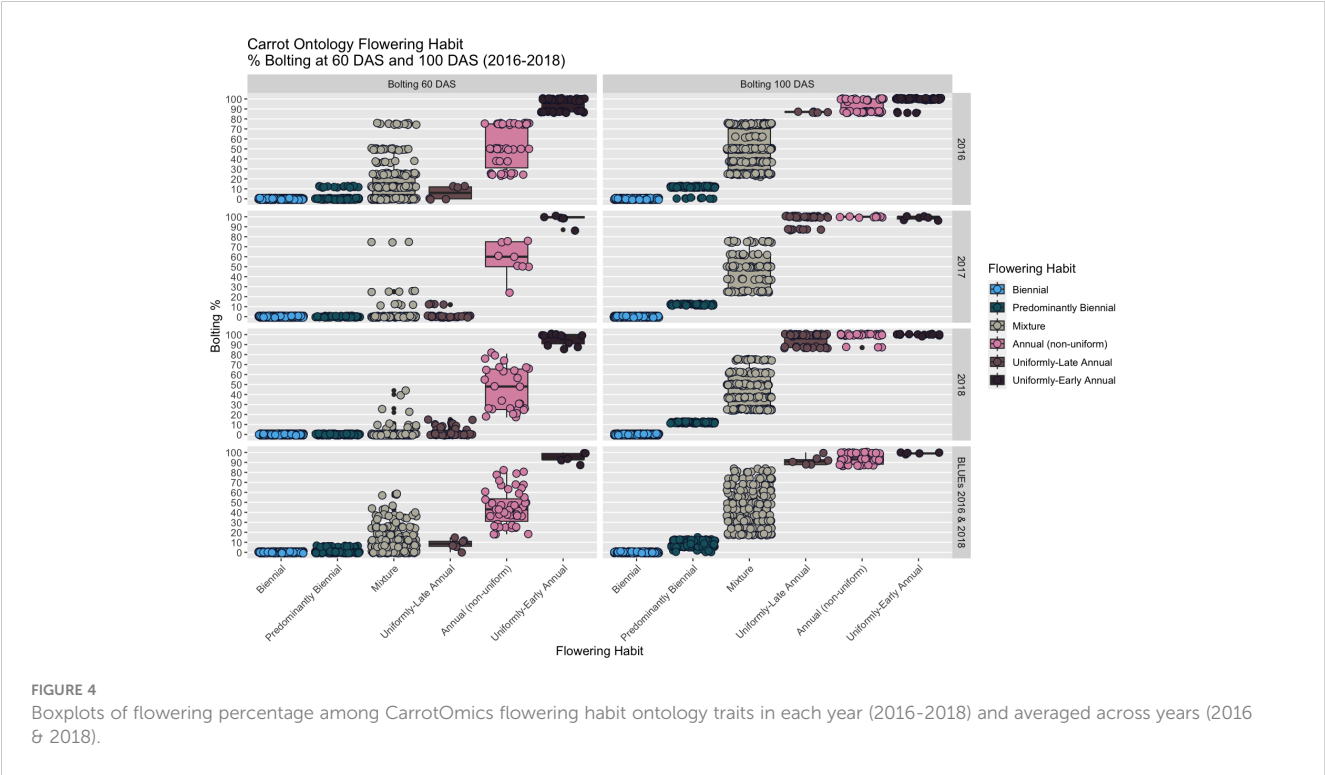
Flowering Habit	2016		2017		2018		2016 & 2018	
	N	%	N	%	N	%	N	%
Biennial	246	36.23	477	68.63	297	43.61	197	29.49
Mixture	332	48.90	161	23.17	288	42.29	410	61.38
Predominantly biennial 0%< % flowering plants ≤ 15%	114	48.90	54	7.77	111	16.30	160	23.95
All other mixtures	218	32.11	107	15.40	177	25.99	250	37.43
Annual	101	14.87	57	8.19	96	14.09	61	9.14
Annual (non-uniform)	47	6.92	9	1.29	27	3.96	50	7.49
Uniformly-early annual	48	7.07	6	0.86	12	1.76	5	0.75
Uniformly-late annual	6	0.88	42	6.04	57	8.37	6	0.90
% Entries Flowering								
Flowering 60 DAS	256	37.70	32	4.60	67	9.84	247	36.98
Flowering 100 DAS	433	63.77	218	31.37	384	56.39	471	70.51

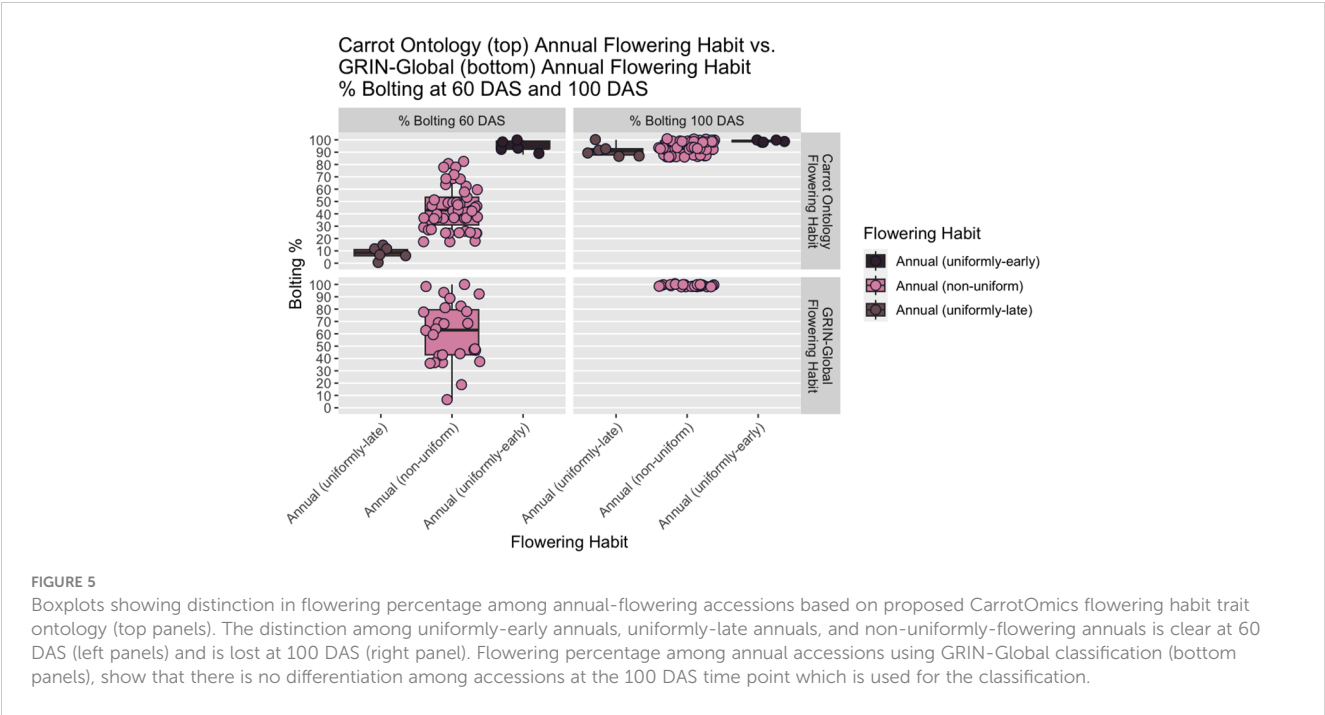
Flowering data for annual, biennial, mixed, and subcategories at 60 DAS and 100 DAS for all accessions planted in 2016-2018. N=number of accessions in flowering habit category; % = percentage of accessions in collection in each flowering habit category. Bold values describe the main three flowering categories (biennial, mixture, and annual), while plain text values (indented) beneath describe the subcategories of each main category.

also appears to be a difference in mean and variation between biennial and annual accessions for canopy height.

We evaluated relationships among our trait ontology’s categories and subcategories (Figure 6B).

Correlations among shoot-growth traits vary by flowering habit category. As with Figure 6A, correlation between seed viability and emergence is high for biennials ($r = 0.68$), predominantly biennials ($r = 0.67$) and mixtures ($r = 0.66$) and low to non-existent for annual populations ($r = 0.0$ for late annuals, $=0.35$ for early annuals and 0.39 for non-uniform annuals). Correlation between emergence and late-season canopy height is very high ($r = 0.84$) for uniformly-early annuals, as is the relationship between seed viability and seed





weight ($r = 0.94$), and low for all other flowering categories. Data presented are on estimated marginal means across years. Uniformly-early annuals have the highest mean seed weight, emergence, and canopy height. We observed that 37% of accessions in this collection expressed some proportion of flowering by 60 DAS and that this increased to 70.5% by 100 DAS, suggesting that the 60 DAS measurement is insufficient to predict whether plants will flower, as only half of accessions that will flower by 100 DAS are flowering at 60 DAS (Figure 7). Many of the accessions that are flowering at 100 DAS may be late-flowering and could be of use to breeders, but would likely have been discarded using the prior classification system.

Discussion

This study is the largest and most diverse global carrot germplasm collection yet evaluated for flowering at mid-season and end-of-season in multi-year trials. This diverse carrot collection

demonstrates maximal breadth of variation for phenological flowering characteristics, with high broad-sense heritability estimates within and across years. This carrot flowering ontology attempts to capture this diverse range of flowering variation with additional categories based on perceived usefulness. High broad sense heritability for flowering reported in our study is similar to estimations in other carrot germplasm collections (Manikanta et al., 2018), and consistent with documented simple inheritance of flowering habit in carrot, with two recessive loci conditioning biennial habit (Alessandro & Galmarini, 2007; Alessandro et al., 2013; Wohlfeiler et al., 2019), and in other plants such as *Arabidopsis* (Michaels & Amasino, 2000), sugar beet (Abe et al., 1997), celery (Quiros et al., 1987), brassicas (Pelofske & Baggett, 1979; Baggett & Kean, 1989), and lettuce (Whitaker, 1944). Our study also agrees with Solberg and Yndgaard (2015), that the diversity within accessions for flowering habit is not well captured by the GRIN-global classification the genebanks' information system, where accessions are categorized as biennial, annual, or mixture. The flowering habit assigned to accessions in this study

TABLE 3A Single year ANOVA (2016-2018) of % flowering (60 DAS).

Source of Variation	Flowering % (60 DAS)								
	2016			2017			2018		
	df	F	p	df	F	p	df	F	p
Accession	678	9.575	***	694	7.295	***	680	13.519	***
Block	1	NS	NS	1	4.788	*	1	0.482	NS
Residuals	631			689			640		
H ²	0.9			0.87			0.93		

Broad-sense heritability (H^2) is very high for all three years studied. P-values in Supplementary Table 1. Statistically significant at * $p \leq 0.05$; *** $p \leq 0.001$; NS, otherwise. ANOVA results indicate that genotype is a highly significant factor for mid-season flowering habit.

TABLE 3B Single year ANOVA (2016–2018) of % flowering (100 DAS).

Source of Variation	Flowering % (100 DAS)								
	2016			2017			2018		
	df	F	p	df	F	p	df	F	p
Accession	678	5.635	***	694	5.142	***	680	6.239	***
Block	1	7.315	***	1	24.493	***	1	0.64	NS
Residuals	631			689			640		
H ²	0.83			0.81			0.84		

Broad-sense heritability (H^2) is high for all years studied. P-values in [Supplementary Table 2](#).
Statistically significant at *** $p \leq 0.001$; NS, otherwise.
ANOVA results indicate that genotype is a highly significant factor for mid-season flowering habit in every year studied.

largely disagreed with flowering designation in the GRIN-Global passport data, even when using the three original categories. Carrot accession flowering habit data in GRIN-Global was often collected on a limited number of carrot plants, and frequently on a single plant. This is further confounded by carrot’s outcrossing nature and that many accessions in this collection originated from landraces and open-pollinated varieties that were increased in the open field, giving way to the possibility of pollen contamination. Taken together, it cannot be assumed that the flowering habit of the individual plant will be reflected in the progeny. Additionally, flowering in carrot is mediated by a network of genes that are differentially influenced by photoperiod, temperature, and illumination intensity (Ou et al., 2017); consequently, the accession’s flowering phenotype at its collection origin could differ when trialed in other climatic conditions. This underscores the need for breeders to use more detailed phenotyping for germplasm they plan to introduce into their program in their respective target environments.

Descriptive statistics and ANOVA results are consistent: flowering percentage varied somewhat by year, with the starkest differences in 2017, a year with poor stand establishment that overestimated biennials and underestimated mixtures compared with 2016 and 2018, which had more comparable values for flowering-category counts (Table 4A). Our data suggests that a year with poor stand establishment, such as 2017, may overestimate the number of biennials. One possible reason for this in our trial is that uniformly-early annuals, with higher and

perhaps earlier emergence, were damaged by hail, as well as early-emerging seedlings from mixed populations, but the later emerging seedlings survived. Mixtures accounted for 36% of the collection, and may contain some amount of early-flowering annual seedstock. Consequently, weather events that eliminate early-emerging plants may bias the resulting stand against annuals and in favor of biennials. Senescence of early annuals can also occur between 60 DAS and 100 DAS, which can in turn increase the perception of biennials. Though this happened infrequently (four of 695 accessions in our study), this can be detected and handled in the data when there is a decline in flowering percentage from 60 DAS to 100 DAS. Population parameter data for 2016 indicates that far more plants (37.7%) had signs of flowering at 60 DAS than in 2017 and 2018 (4.6% - 9.4%). This could have been due to early-season weather events that accelerated flowering in 2016. In large plant breeding trials, there is a need to balance efficiency and precision of measurements. Often large trials must sacrifice some precision for efficiency. Our five-point scoring system increases efficiency, with each plot taking fewer than five seconds to evaluate. High broad-sense heritability estimates for all flowering traits in all years ($0.81 < H^2 < 0.93$) suggests that our methodology successfully detects genetic signals for flowering habit at both time points, at least in a diverse collection. Directly measuring flowering percentage on a continuous scale may result in more precise estimates, but gains may be trivial and unnecessary compared to the time spent on such a large collection. Extreme values in scatterplots of Figure 6B boxplots appeared to have little leverage or influence on

TABLE 4A Multi-Year (2016–2018) ANOVA for flowering (%) (60 DAS) and bolting (%) (100 DAS) results indicate that genotype and genotype x year interaction are highly statistically significant factors in both traits across all three years.

Source of Variation	Flowering % 60 DAS			Flowering % 100 DAS		
	df	F	p	df	F	p
Accession	657	17.493	***	657	5.688	***
Year	2	5.753	**	2	8.572	***
Accession x Year	1310	4.821	***	1310	1.428	***
Block within Year	3	0.780	NS	3	9.627	***
Residuals	1889			1889		
H ²	0.64			0.88		

Multi-year broad-sense heritability (H^2) is moderately high for both traits. P-values in [Supplementary Table 3](#).
Statistically significant at ** $p \leq 0.01$; *** $p \leq 0.001$; NS, otherwise.

TABLE 4B Multi-year ANOVA (2016 & 2018) for flowering (%) (60 DAS) and bolting (%) (100 DAS) results indicate that genotype and genotype x year interaction are highly statistically significant factors across all three years.

Source of Variation	Flowering % 60 DAS			Flowering % 100 DAS		
	df	F	p	df	F	p
Accession	657	12.637	***	657	5.340	***
Year	1	6.165	**	1	6.776	**
Accession x Year	643	4.659	***	643	1.119	*
Block within Year	2	0.785	NS	2	4.239	**
Residuals	1234			1234		
H ²	0.57			0.89		

Multi-year broad-sense heritability (H^2) is moderately high for both traits. P-values in [Supplementary Table 4](#). Statistically significant at * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; NS, otherwise.

slope. The three annual subcategories we created in [Figure 6B](#) enabled us to observe that the correlation we see in annuals in [Figure 6A](#) is driven by the uniformly-flowering early annuals. However, this correlation could be driven by genetic drift and small sample size ($n=5$). The inclusion of wild germplasm, which tends to flower early, in a future study could clarify this interpretation.

Flowering trait ontology

A fundamental purpose of crop ontology is to create a descriptive and consistent vocabulary for crop traits, facilitating communication across collaborators and comparison of data between trial years, locations, and breeding programs ([Shrestha et al., 2012](#); [Walls et al., 2012](#)). We have provided improved descriptions for flowering habit characteristics in carrot, including standard methodologies and time-frames for trait evaluation. Previous studies in carrot have used “early”/“late” as modifiers to flowering habit in the first growing season and second growing season, respectively. The term “annual” has conventionally been used synonymously with “early-flowering”, while the term “biennial” has been used synonymously with “late-flowering” ([Alessandro & Galmarini, 2007](#); [Alessandro et al., 2013](#); [Wohlfeiler et al., 2019](#); [Simon & Grzebelus, 2020](#); [Wohlfeiler et al., 2021](#)). However, confusion arises when early-flowering has also been used to describe wild carrot plants that flower earlier in the first growth season than cultivated annuals, which tend to flower later in the same season ([Simon & Grzebelus, 2020](#)). Furthermore, recent research suggests that a gradient of vernalization requirements exists within annuals and biennials ([Wohlfeiler et al., 2021](#)) – we propose a standardized vocabulary to refer to this germplasm.

While flowering has been typically measured at 100 DAS, the 60 DAS measurement enabled differentiation between annuals that flower uniformly early in the season (uniformly-early annuals) from annuals that flower uniformly at the end of the season (uniformly-late flowering annuals) ([Figure 5](#)). As posited by [Wohlfeiler et al. \(2019\)](#), these accessions could represent a range of vernalization requirements for flowering, with later flowering annuals needing more cold exposure and uniformly-early annuals needing far less, if any, cold exposure. Mixtures, which did not have a clear threshold defined previously, are entries with both biennial and annual plants

in intermediate quantities, which we defined as between 1% and 85% flowering. We created a useful subcategory of mixture – predominantly biennial ($n=160$) – that constitutes a sizable 24% of accessions in this germplasm collection. Predominantly biennial accessions contain fewer than 15% flowering plants at the end of the season. Given the simple inheritance of biennial habit in carrot ([Wohlfeiler et al., 2019](#)), it is practical to select biennial plants and eliminate annual plants from a predominantly biennial population with other favorable traits for breeding. The establishment of the predominantly biennial subcategory nearly doubles the availability of germplasm with commercial potential, from 197 biennials to 357 biennials and predominantly biennials, accounting for 54% of the germplasm collection. This subcollection is a useful source of genetic diversity for breeders.

In this crop ontology, we propose categories that more fully describe the flowering habit of cultivated carrot, and we encourage carrot researchers to utilize and expand upon the descriptive terminology we provide in [Table 1](#) and [Figure 3](#). In this ontology, “uniformly-early flowering” describes both precocity and uniformity of annual flowering, characterized by a high proportion of plants flowering at 60 DAS. “Late” describes only uniformity of end-of-season flowering, characterized by a low proportion of plants flowering at 60 DAS and a high proportion of plants flowering on harvest day or end-of-season. Following this logic, researchers can extend their evaluation of biennial carrot germplasm into the second season of growth. Collecting data on precocity of biennial flowering could identify uniformly-early biennials, which is important for carrot seed producers. This study provides a framework and an opportunity to study carrots in the second season of growth, and lays the groundwork for performing seed-to-seed phenotyping over the crop’s lifecycle.

Logically, the threshold of 85% for annual flowering means that these groups could also include mixtures. It would be more correct biologically to say that annuals are plants with 100% flowering at end of season, and any entries less than 100% flowering, and greater than 0% flowering, is a mixture. However, given that annuals are undesirable in temperate carrot production systems, we found little practical use in defining a “predominantly annual” population. It is more likely that non-flowering plants in an annual population at the end of the season are late-flowering annuals rather than biennials.



FIGURE 6

(A) Correlation matrix of vegetative growth traits factored by GRIN-Global flowering descriptors (biennial, annual, mixture) as defined by 100 DAS evaluation: 0% flowering (biennial), 100% flowering (annual), and 1% - 99% flowering (mixture) (2016 and 2018 marginal means). (B) Correlation matrix of vegetative growth traits factored by proposed CarrotOmics flowering habit trait ontology (2016 and 2018 marginal means). Pearson's correlations are statistically significant at * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; NS, otherwise.

Furthermore, breeding programs typically avoid intercrossing temperate (biennial) and subtropical (annual) carrots, and rarely use wild germplasm, given that substantial new challenges outstrip the benefits of such a cross (Simon & Grzebelus, 2020). The accessions identified, as well as the methodology provided for the identification of uniformly-late annuals, could prove useful for carrot breeders in subtropical climates or semi-arid climates, where uniformly-late annuals are essential for successful cultivar development; carrots adapted to subtropical regions often flower

with reduced exposure to cold and tend to flower prolifically in temperate regions. In subtropical climates, carrots are managed as late-flowering annuals, with some plants used to produce the root crop and the remainder generate the seed stock in the same season. Biennial plants are unsuited to subtropical and semi-arid market, as they do not contribute to the seed crop.

The ability to distinguish between uniformly-early annuals and uniformly-late annuals is present at 60 DAS and lost by 100 DAS (Figures 2, 4, 5), illustrating that the additional subcategories are

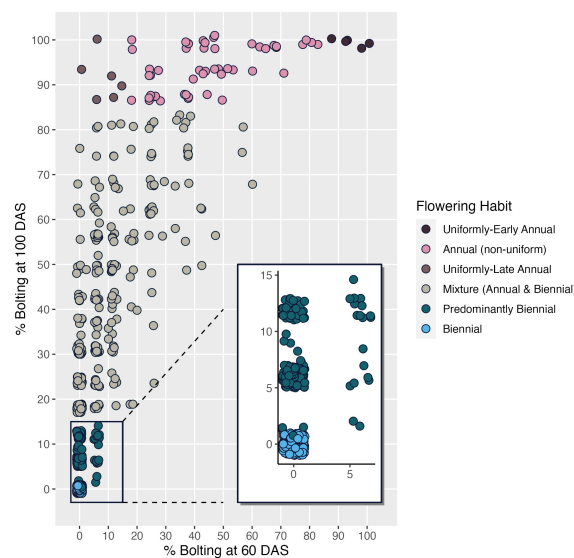


FIGURE 7

Scatterplot visualization of CarrotOmics flowering habit trait ontology based on estimated marginal means from 2016 and 2018.

distinct from one another in all three years evaluated. This demonstrates the need to evaluate flowering at two timepoints to capture the phenological variation in this population. The 60 DAS measurement alone is too early to identify the true biennials, but crucial when combined with the 100 DAS measurement for distinguishing uniformly-early annuals from uniformly-late annuals. The 100 DAS measurement alone is too late to differentiate the uniformly-early annuals from uniformly-late annuals, but critical to differentiate uniformly-late annuals from biennials, as well as identifying predominantly biennial populations. Our 100 DAS flowering evaluation allows us to further distinguish between accessions with uniformly-early annual habit (0.75% of accessions), uniformly-late annuals (0.9% of accessions), and the remaining annuals with non-uniform flowering (7.5% of accessions).

Identification of predominantly biennial accessions provides improved characterization of mixed population accessions and nearly doubles the availability of germplasm with commercial potential ($n=357$) for temperate areas of root production, compared with strictly biennial germplasm ($n=197$). Custom core collections could be curated from this data for carrot breeders and researchers interested in new sources of genetic diversity for specific traits, climatic conditions, production uses, or market types (Brown, 1989) by filtering on data such as plant morphology, ecogeographical origin, molecular marker data, and genetic relatedness (Berger et al., 2013; Byrne et al., 2018; Corak et al., 2019; Corak, 2021). In the biennial and predominantly biennial custom core collection, a minicore can be curated from our shoot-growth phenotypic data for traits such as plant height, canopy coverage, emergence, or available data for any trait of interest. Cores that maintain diversity while also maximizing desirable traits have great utility to breeders. The upper threshold for emergence is just as high in the biennial and predominantly biennial groups, which demonstrates the advantage of evaluating this custom core collection more closely for agronomically important traits.

Agromorphological data has been leveraged to create core collections in sweet potato (Huamán et al., 1999), potato (Huamán et al., 2000), groundnut (Upadhyaya et al., 2003), pigeonpea (Reddy et al., 2005), maize (Malosetti & Abadie, 2001; Li et al., 2005; Risliawati et al., 2023), safflower (Dwivedi et al., 2005), yam (Girma et al., 2018), walnut (Mahmoodi et al., 2019), pomegranate (Razi et al., 2021), lentil (Tripathi et al., 2022), and Indian mustard (Nanjundan et al., 2022). Corak compared methods for creating custom core collections in a subset of 433 accessions from our study's carrot diversity panel, and found that custom methods combined with representative methods built cores balanced for genetic representation and enriched for desirable phenotypes, though it is important to note that carrot has low population structure (Corak et al., 2019; Corak, 2021). Similarly, the mixtures and annuals we identified in this collection can be used as sources of genetic diversity by carrot breeders and researchers targeting carrots for subtropical/semi-arid climates.

Limitations of our study

There is a potential bias in every cultivated germplasm collection, as traits that are considered useful are relative to culture, production system, environment, technological access, local economy, and myriad unmeasurable factors. One bias inherent in the USDA cultivated carrot germplasm collection is preference for biennial germplasm, as that is what is grown in the U.S. Other gene banks may have higher diversity and larger samples for annual habits. We expect to see varying proportions of flowering in carrot germplasm, with the least in biennial cultivated carrot bred for cool temperate climates, and increasing amount of flowering in annual cultivated carrot bred for subtropical/semi-arid climates and wild carrot, indicating that the proportion of flowering in a population is relative to the germplasm under evaluation. Previous studies on

U.S. commercial carrot cultivars, which have been selected for bolting resistance, had insignificant amounts of bolting, such that they were noted but not analyzed (Luby et al., 2016). In wild germplasm, most were bolting in early planting but all were non-bolting in later plantings (Solberg and Yndgaard, 2015), which could be due to cooler spring with sufficiently low temperatures to induce vernalization, and a warm, temperate summer resulting in no bolting. These varying results likely reflect the genotypic base of the carrot germplasm and the environment under study. Similarly, biennial genotypes in warm climates will never flower, while warm-acclimated annuals in heat-stressed environments may flower readily and prolifically (Simon et al., 2019). This study is limited to three years in one temperate environment. With this said, this environment is a commercially relevant region, with Wisconsin ranking in the top 3 U.S. states for carrot root crop production. Studies are underway to characterize this USDA germplasm collection in multiple other commercially relevant temperate carrot production environments. Accessions characterized in Wisconsin may not be stable in other growing regions. Climate warming could stimulate early-flowering or increased total flowering in germplasm we have already characterized.

Conclusions and recommendations

Motivations for this flowering ontology were twofold: to attempt to understand the essential nature of carrot flowering phenology and to promote utilization of diverse germplasm by way of its characterization for agronomically critical traits. The former goal was satisfied by combining 60 DAS and end-of-season flowering data and identifying subtle but significant distinctions among annuals that open questions into carrot's life history and domestication. The ontology provides trait definitions and methods for measuring flowering habit in diverse germplasm. Users of this ontology can set their own threshold based on what they see as tolerable for their own program. Future evaluations can be improved by overseeding accessions with low germination or low emergence to achieve the sample size required for accurate characterization. To better elucidate the relationship between trait stability in other temperate and economically relevant carrot production regions, genomic data and multi-environmental data will be integrated with this study's phenotype data in future carrot diversity panel studies on QTL x E. Multi-environmental GWAS studies in temperate, subtropical, and semi-arid climates will facilitate molecular characterization of flowering habit in other *Daucus* germplasm collections.

We were also motivated by utility, which is the primary concern of breeders evaluating diverse germplasm. Identification of predominantly biennial germplasm fulfilled this goal, expanding the availability of genetic backgrounds that can be leveraged in temperate breeding programs. This evaluation has improved access to useful plant introductions in mixtures by identifying predominantly biennial accessions, doubling the size of the commercially promising accession gene pool for temperate carrot production regions. Given the relatively simple inheritance of biennial flowering habit, breeders for temperate root production can select biennial individuals out of predominantly biennial

germplasm or other mixed flowering populations. However, selections should be evaluated for other agronomic traits and validated in multi-year, multi-environment trials in target locations. Data from canopy studies can be used in combination with flowering studies to identify accessions with high emergence and vigorous shoot growth for temperate climates. Similarly, the mixtures and annuals we identified can be leveraged as sources of genetic diversity by carrot breeders and researchers targeting carrots for subtropical and semi-arid climates. As such, flowering habit characterization has increased access to genetic resources with baseline commercial potential and provided useful data and methods to global users of carrot germplasm.

While within-accession diversity can be a challenge for *ex situ* conservation systems (Solberg and Yndgaard, 2015), we propose leveraging it as an opportunity to perform selection for desirable ecotypes, enabling identification of accessions with flowering traits required by local markets. We have provided a roadmap for evaluating and characterizing flowering habit in vegetable crops with mixed lifeforms, and this methodology is immediately useful to breeders and users of carrot PGR. Evaluating flowering habit as a gradient, rather than a binary trait, expands availability of commercially viable germplasm, further lowering the barrier to utilization of carrot PGR.

Data availability statement

Data on this study's accessions are hosted on the CarrotOmics database. <https://www.CarrotOmics.org/file/409952>. Researchers may request these accessions through the USDA Germplasm Resources Information Network (GRIN) database of the U.S. National Plant Germplasm System (NPGS). <https://npgsweb.ars-grin.gov/gringlobal/search>.

Author contributions

JL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. ML: Formal analysis, Software, Validation, Writing – review & editing. JD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing. PS: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2016-51181-25400.

Acknowledgments

The authors would like to thank Kathleen Reitsma for her capable assistance in securing carrot germplasm; Lucia Gutiérrez, Irwin Goldman, and Edgar Spalding for their insightful comments on thesis chapters that led up to this manuscript; Keo Corak, Ken Owens, David Spooner, and Bill Rolling for meaningful discussions on crop genetic diversity; Doug Senalik for managing the CarrotOmics database; Tom Horesji, Shelby Ellison, Kevser Özel, Sarah Turner, Gunay Yildiz, Sarah Acosta, Edie Africano, Annelise Atwood, Alexis Lightner, and Hayley Stoneman for data collection and assistance in the field.

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.2024.1342513/full#supplementary-material>

SUPPLEMENTARY TABLE 1

P-values of single-year ANOVA (2016–2018) for flowering (%) (60 DAS) results indicate that genotype is a highly significant factor in all three years studied. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 2

P-values of single-year ANOVA (2016–2018) of flowering (%) (100 DAS). Results indicate that genotype is a highly significant factor in every year studied. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 3

P-values of multi-year ANOVA for flowering (%) (60 DAS) and flowering (%) (100 DAS) results indicate that genotype and genotype x year interaction are highly statistically significant factors across all three years. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 4

P-values of multi-year ANOVA (2016 & 2018) for flowering (%) (60 DAS) and flowering (%) (100 DAS) indicate that genotype and genotype x year interaction are highly statistically significant factors across all three years. Statistically significant p-values in bold.

References

- Abe, E., Fujino, K., Masuda, K., and Yamaguchi, Y. (2014). Isolation and expression profiling of a CONSTANS-like gene and two FLOWERING LOCUS T-like genes from *spinacia oleracea* L. *AJPS* 05, 4018–4028. doi: 10.4236/ajps.2014.526420
- Abe, J., Guan, G.-P., and Shimamoto, Y. (1997). A gene complex for annual habit in sugar beet (*Beta vulgaris* L.). *Euphytica* 94, 129–135. doi: 10.1023/A:1002963506818
- Alessandro, M. S., and Galmarini, C. R. (2007). Inheritance of vernalization requirement in carrot. *J. Am. Soc. Hortic. Sci.* 132, 525–529. doi: 10.21273/JASHS.132.4.525
- Alessandro, M. S., Galmarini, C. R., Iorizzo, M., and Simon, P. W. (2013). Molecular mapping of vernalization requirement and fertility restoration genes in carrot. *Theor. Appl. Genet.* 126, 415–423. doi: 10.1007/s00122-012-1989-1
- Allender, C. (2019). "Genetic resources for carrot improvement," in *The carrot genome, compendium of plant genomes*. Eds. P. Simon, M. Iorizzo, D. Grzebelus and R. Baranski (Cham: Springer International Publishing, Cham), 93–100. doi: 10.1007/978-3-030-03389-7_6
- Amasino, R. M. (2005). Vernalization and flowering time. *Curr. Opin. Biotechnol.* 16, 154–158. doi: 10.1016/j.copbio.2005.02.004
- Arnold, J. B. (2021). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.4. Available online at: <https://CRAN.R-project.org/package=ggthemes>.
- Atagul, O., Calle, A., Demirel, G., Lawton, J. M., Bridges, W. C., and Gasic, K. (2022). Estimating heat requirement for flowering in peach germplasm. *Agronomy* 12, 1002. doi: 10.3390/agronomy12051002
- Bååth, R. (2018). beeper: Easily Play Notification Sounds on any Platform. R package version 1.3. Available online at: <https://CRAN.R-project.org/package=beeper>.
- Baggett, J. R., and Kean, D. (1989). Inheritance of annual flowering in brassica oleracea. *HortScience* 24, 662–664. doi: 10.21273/HORTSCI.24.4.662
- Baldwin, S., Revanna, R., Pither-Joyce, M., Shaw, M., Wright, K., Thomson, S., et al. (2014). Genetic analyses of bolting in bulb onion (*Allium cepa* L.). *Theor. Appl. Genet.* 127, 535–547. doi: 10.1007/s00122-013-2232-4
- Banga, O. (1957). Origin of the European cultivated carrot. *Euphytica* 6 (1), 54–63. doi: 10.1007/BF00179518
- Bao, S., Ou, C., Zhuang, F., Chen, J., and Zhao, Z. (2010). Study of premature bolting of carrot in spring cultivation. *China Vegetables* 6, 38–42.
- Basavaraja, T., Tripathi, A., Chandora, R., Pratap, A., Manjunaatha, L., Gurumurthy, S., et al. (2023). Evaluation of phenological development and agronomic traits in exotic common bean germplasm across multiple environments. *Plant Genet. Resour.* 21, 195–203. doi: 10.1017/S1479262123000618
- Bashtanova, U. B., and Flowers, T. J. (2011). Diversity and physiological plasticity of vegetable genotypes of coriander improves herb yield, habit and harvesting window in any season. *Euphytica* 180, 369–384. doi: 10.1007/s10681-011-0396-z
- Bates, D., Maechler, 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
- Belaj, A., de la Rosa, R., León, L., Gabaldón-Leal, C., Santos, C., Porras, R., et al. (2020). Phenological diversity in a World Olive Germplasm Bank: Potential use for breeding programs and climate change studies. *Span J. Agric. Res.* 18, e0701. doi: 10.5424/sjar/2020181-15017
- Berger, J. D., Hughes, S., Snowball, R., Redden, B., Bennett, S. J., Clements, J. C., et al. (2013). Strengthening the impact of plant genetic resources through collaborative collection, conservation, characterisation, and evaluation: A tribute to the legacy of Dr Clive Francis. *Crop Pasture Sci.* 64, 300. doi: 10.1071/CP13023
- Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S. R., et al. (2022). Reap the crop wild relatives for breeding future crops. *Trends Biotechnol.* 40, 412–431. doi: 10.1016/j.tibtech.2021.08.009
- Bolton, A., Nijabat, A., Mahmood-ur-Rehman, M., Naveed, N. H., Mannan, A. M., Ali, A., et al. (2019). Variation for heat tolerance during seed germination in diverse carrot [*Daucus carota* (L.)] germplasm. *HortScience* 54, 1470–1476. doi: 10.21273/HORTSCI13333-18
- Bolton, A., and Simon, P. (2019). Variation for salinity tolerance during seed germination in diverse carrot [*Daucus carota* (L.)] germplasm. *HortScience* 54, 38–44. doi: 10.21273/HORTSCI13333-18
- Brainard, S. H., Bustamante, J. A., Dawson, J. C., Spalding, E. P., and Goldman, I. L. (2021). A digital image-based phenotyping platform for analyzing root shape attributes in carrot. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.690031
- Brown, A. H. D. (1989). Core collections: A practical approach to genetic resources management. *Genome* 31, 818–824. doi: 10.1139/g89-144
- Bruznican, S., De Clercq, H., Eeckhaut, T., Van Huylenbroeck, J., and Geelen, D. (2020). Celery and celeriac: A critical view on present and future breeding. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01699

- 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, 18. doi: 10.2135/cropsci2017.05.0303
- Campa, A., and Ferreira, J. J. (2018). Genetic diversity assessed by genotyping by sequencing (GBS) and for phenological traits in blueberry cultivars. *PLoS One* 13, e0206361. doi: 10.1371/journal.pone.0206361
- Chitwood, J., Shi, A., Mou, B., Evans, M., Clark, J., Motes, D., et al. (2016). Population structure and association analysis of bolting, plant height, and leaf erectness in spinach. *Horts* 51, 481–486. doi: 10.12173/HORTSCI.51.5.481
- Ciriaci, T., Beretta, M., Donati, F., and Sabatini, E. (2013). “G×E Interaction in lettuce bolting phenomena,” In *III International Symposium on Molecular Markers in Horticulture* 1100, 141–144.
- Coe, K., Bostan, H., Rolling, W., Turner-Hissong, S., Macko-Podgórni, A., Senalik, D., et al. (2023). Population genomics identifies genetic signatures of carrot domestication and improvement and uncovers the origin of high-carotenoid orange carrots. *Nat. Plants* 9 (10), 1643–1658.
- Colley, M. (2017). What's red, pink, and not bolting right now? Organic seed alliance newsletter, July 22. Available online at: <https://seedalliance.org/2017/whats-red-pink-and-notbolting-right-now/>.
- Corak, K. E. (2021). *Strategies to identify and introgress production and quality traits from genetic resources to elite carrot cultivars* (Proquest, LLC: The University of Wisconsin-Madison). doi: 10.2135/cropsci2006.03.0169gag
- Corak, K. E., Ellison, S. L., Simon, P. W., Spooner, D. M., and Dawson, J. C. (2019). Comparison of representative and custom methods of generating core subsets of a carrot germplasm collection. *Crop Sci.* 59, 1107–1121. doi: 10.2135/cropsci2018.09.0602
- Cruz, V. M. V., Luhman, R., Marek, L. F., Rife, C. L., Shoemaker, R. C., Brummer, E. C., et al. (2007). Characterization of flowering time and SSR marker analysis of spring and winter type Brassica napus L. germplasm. *Euphytica* 153, 43–57. doi: 10.1007/s10681-006-9233-1
- Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., and Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Sci.* 57, 1070–1082. doi: 10.2135/cropsci2016.10.0885
- Dowker, B. D., Bowman, A. R. A., and Faulkner, G. J. (1971). The effect of selection during multiplication on the bolting resistance and internal quality of Avoncreed red beet. *J. Hort. Sci.* 46, 307–311. doi: 10.1080/00221589.1971.11514411
- Dowker, B. D., and Jackson, J. C. (1975). Bolting in some carrot populations. *Ann. Appl. Biol.* 79, 361–365. doi: 10.1111/j.1744-7348.1975.tb01592.x
- Dwivedi, S. L., Upadhyaya, H. D., and Hegde, D. M. (2005). Development of Core Collection using Geographic Information and Morphological Descriptors in Safflower (*Carthamus tinctorius* L.) Germplasm. *Genet. Resour. Crop Evol.* 52, 821–830. doi: 10.1007/s10722-003-6111-8
- Elfadl, E., Reinbrecht, C., and Claupen, W. (2010). Evaluation of phenotypic variation in a worldwide germplasm collection of safflower (*Carthamus tinctorius* L.) grown under organic farming conditions in Germany. *Genet. Resour. Crop Evol.* 57, 155–170. doi: 10.1007/s10722-009-9458-7
- Ellison, S. (2019). “Carrot domestication,” In *The carrot genome*. Eds. P. W. Simon, M. Iorizzo, D. Grzebelus and R. Baranski (Cham: Springer International Publishing), 77–91. doi: 10.1007/978-3-030-03389-7_5
- Ellison, S. L., Luby, C. H., Corak, K. E., Coe, K. M., Senalik, D., Iorizzo, M., et al. (2018). Carotenoid presence is associated with the *or* gene in domesticated carrot. *Genetics* 210, 1497–1508. doi: 10.1534/genetics.118.301299
- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., et al. (2012). *Package 'car'* Vol. 16 (Vienna: R Foundation for Statistical Computing), 333.
- Geoffriau, E., Charpentier, T., Huet, S., Hägnfeldt, A., Lopes, V., Nothnagel, T., et al. (2019). “CarrotDiverse: understanding variation in a wild relative of carrot,” In *II International Symposium on Carrot and Other Apiaceae* 1264, 151–156. doi: 10.17660/ActaHortic.2019.1264.18
- Gerbrandt, E. M., Bors, R. H., Chibbar, R. N., and Baumann, T. E. (2017). Spring phenological adaptation of improved blue honeysuckle (*Lonicera caerulea* L.) germplasm to a temperate climate. *Euphytica* 213, 172. doi: 10.1007/s10681-017-1958-5
- Girma, G., Bhattacharjee, R., Lopez-Montes, A., Gueye, B., Ofodile, S., Franco, J., et al. (2018). Re-defining the yam (*Dioscorea* spp.) core collection using morphological traits. *Plant Genet. Resour.* 16, 193–200. doi: 10.1017/S1479262117000144
- Gohel, D., and Skintzos, P. (2023). *flextable: Functions for Tabular Reporting*. R package version 0.9.1. Available online at: <https://CRAN.R-project.org/package=flextable>.
- Goldman, I. L. (2004). “Breeding biennial crops,” in *Encyclopedia of plant and crop science* (Marcel Dekker, Inc., Manhattan), 1–3.
- GRIN-Global (2023). Germplasm resources information network - global, USDA-ARS (National Plant Germplasm System). Available online at: <https://www.grin-global.org/> (Accessed April 26, 2023).
- Grzebelus, D., Iorizzo, M., Senalik, D., Ellison, S., Cavagnaro, P., Macko-Podgórni, A., et al. (2014). Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. *Mol. Breed.* 33, 625–637. doi: 10.1007/s11032-013-9979-9
- Havey, M. J. (2018). Onion breeding. *Plant Breeding Reviews*. 42, 39–85. doi: 10.1002/9781119521358.ch2
- Holland, H., and Dowker, B. D. (1969). The breeding of Avoncreed, a red beet variety resistant to bolting. *J. Hort. Sci.* 44, 257–264. doi: 10.1080/00221589.1969.11514307
- Huamán, Z., Aguilar, C., and Ortiz, R. (1999). Selecting a Peruvian sweetpotato core collection on the basis of morphological, eco-geographical, and disease and pest reaction data. *Theor. Appl. Genet.* 98, 840–844. doi: 10.1007/s001220051142
- Huamán, Z., Ortiz, R., and Gómez, R. (2000). Selecting a *Solanum tuberosum* subsp. *andigena* core collection using morphological, geographical, disease and pest descriptors. *Am. J. Potato Res.* 77, 183–190. doi: 10.1007/BF02853943
- Hyun, D. Y., Kim, O.-T., Bang, K.-H., Kim, Y.-C., Yoo, N. H., Kim, C. W., et al. (2009). Genetic and molecular studies for regulation of bolting time of onion (*Allium cepa* L.). *J. Plant Biol.* 52, 602–608. doi: 10.1007/s12374-009-9078-y
- Iorizzo, M., Senalik, D. A., Ellison, S. L., Grzebelus, D., Cavagnaro, P. F., Allender, C., et al. (2013). Genetic structure and domestication of carrot (*Daucus carota* subsp. *Sativus*) (Apiaceae). *Am. J. Bot.* 100, 930–938. doi: 10.3732/ajb.1300055
- Jang, S.-W., Kwak, J.-H., Choi, S.-K., Park, S., Lee, J.-N., Cho, C.-H., et al. (2019). Breeding of late bolting and high yield lettuce 'Jahokmaschima.' *Korean J. Breed. Sci.* 51, 146–150. doi: 10.9787/KJBS.2019.51.2.146
- Jiang, M., Zhang, Y., Yang, X., Li, X., and Lang, H. (2023). Brassica rapa orphan gene BR1 delays flowering time in Arabidopsis. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1135684
- Khokhar, K. M., Hadley, P., and Pearson, S. (2007). Effect of cold temperature durations of onion sets in store on the incidence of bolting, bulbing and seed yield. *Scientia Hort.* 112, 16–22. doi: 10.1016/j.scienta.2006.12.038
- Lebeda, A., Křístková, E., Kitner, M., Majesky, J., Doležalová, I., Khoury, C. K., et al. (2019). Research gaps and challenges in the conservation and use of north american wild lettuce germplasm. *Crop Sci.* 59, 2337–2356. doi: 10.2135/cropsci2019.05.0350
- Lenth, R. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.8. Available online at: <https://CRAN.R-project.org/package=emmeans>.
- Li, Y., Shi, Y., Cao, Y., and Wang, T. (2005). Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. *Genet. Resour. Crop Evol.* 51, 845–852. doi: 10.1007/s10722-005-8313-8
- Linke, B., Alessandro, M. S., Galmarini, C. R., and Nothnagel, T. (2019). “Carrot floral development and reproductive biology,” in *The carrot genome*. Eds. P. W. Simon, M. Iorizzo, D. Grzebelus and R. Baranski (Cham: Springer International Publishing), 27–57. doi: 10.1007/978-3-030-03389-7_3
- Loarca, J., Liou, M., Dawson, J. C., and Simon, P. W. (2024a). Carrot trait ontology enables evaluation of shoot-growth variation in diverse carrot (*Daucus carota* L.) germplasm resources for crop improvement. *Front. Plant Sci.* 15, 1342512. doi: 10.3389/fpls.2024.1342512
- Luby, C. H., Dawson, J. C., and Goldman, I. L. (2016). Assessment and Accessibility of Phenotypic and Genotypic Diversity of Carrot (*Daucus carota* L. var. *sativus*) Cultivars Commercially Available in the United States. *PLoS One* 11, e0167865. doi: 10.1371/journal.pone.0167865
- Macko-Podgórni, A., Machaj, G., Stelmach, K., Senalik, D., Grzebelus, E., Iorizzo, M., et al. (2017). Characterization of a Genomic Region under Selection in Cultivated Carrot (*Daucus carota* subsp. *Sativus*) Reveals a Candidate Domestication Gene. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00012
- Mahmoodi, R., Dadpour, M. R., Hassani, D., Zeinalabedini, M., Vendramin, E., Micali, S., et al. (2019). Development of a core collection in Iranian walnut (*Juglans regia* L.) germplasm using the phenotypic diversity. *Scientia Hort.* 249, 439–448. doi: 10.1016/j.scienta.2019.02.017
- Malosetti, M., and Abadie, T. (2001). Sampling strategy to develop a core collection of Uruguayan maize landraces based on morphological traits. *Genet. Resour. Crop Evol.* 48, 381–390. doi: 10.1023/A:1012003611371
- Manikanta, D. S., Poleshi Chaitra, A., and Cholin, S. (2018). Understanding the genetic variability, heritability and association pattern for the characters related to reproductive phase of carrots (*Daucus carota* L.) in tropical region. *JAHS* 20, 225–232. doi: 10.37855/jah.2018.v20i03.39
- McGrath, J. M., and Panella, L. (2018). Sugar beet breeding. *Plant Breeding Reviews*. 42, 167–218. doi: 10.1002/9781119521358.ch5
- Michaels, S. D., and Amasino, R. M. (2000). Memories of winter: Vernalization and the competence to flower. *Plant Cell Environ.* 23, 1145–1153. doi: 10.1046/j.1365-3040.2000.00643.x
- Millard, S. P. (2013). *EnvStats: an R package for environmental statistics* (New York: Springer). Available at: <https://www.springer.com>. ISBN 978-1-4614-8455-4. doi: 10.17660/ActaHortic.2002.588.10
- Morales, M. R., Maynard, E., and Janick, J. (2006). Adagio: A slow-bolting arugula. *HortScience* 41 (6), 1506–1507. doi: 10.21273/HORTSCI.41.6.1506
- Nanjundan, J., Aravind, J., Radhamani, J., Singh, K. H., Kumar, A., Thakur, A. K., et al. (2022). Development of Indian mustard [*Brassica juncea* (L.) Czern.] core collection based on agro-morphological traits. *Genet. Resour. Crop Evol.* 69, 145–162. doi: 10.1007/s10722-021-01211-7
- Nishioka, M., Tamura, K., Hayashi, M., Fujimori, Y., Ohkawa, Y., Kuginuki, Y., et al. (2005). Mapping of QTLs for Bolting Time in Brassica rapa (syn. campestris) under Different Environmental Conditions. *Breed. Sci.* 55, 127–133. doi: 10.1270/jsbbs.55.127
- Ou, C.-G., Mao, J.-H., Liu, L.-J., Li, C.-J., Ren, H.-F., Zhao, Z.-W., et al. (2017). Characterising genes associated with flowering time in carrot (*Daucus carota* L.) using transcriptome analysis. *Plant Biol.* 19, 286–297. doi: 10.1111/plb.12519

- Pelofske, P. J., and Baggett, J. R. (1979). Inheritance of internode length, plant form, and annual habit in a cross of cabbage and broccoli (*Brassica oleracea* var. *Capitata* L. and var. *Italica* Plenck.). *Euphytica* 28, 189–197. doi: 10.1007/BF00029191
- Pérez, M. B., Carvajal, S., Beretta, V., Bannoud, F., Fangio, M. F., Berli, F., et al. (2023). Characterization of purple carrot germplasm for antioxidant capacity and root concentration of anthocyanins, phenolics, and carotenoids. *Plants* 12, 1796. doi: 10.3390/plants12091796
- Peterson, C. E. (1986). Carrot breeding. Breeding Vegetable Crops. Available online at: <https://ci.nii.ac.jp/naid/10006567255/>.
- Posit team (2023). *RStudio: integrated development environment for R* (Boston, MA: Posit Software, PBC). Available at: <http://www.posit.co/>.
- Prohens, J., and Nuez, F. (2008). *Vegetables II: fabaceae, liliaceae, solanaceae, and umbelliferae* (New York: Springer). doi: 10.1007/978-0-387-74110-9
- Quiros, C. F. (1993). “Celery: *apium graveolens* L,” in *Genetic improvement of vegetable crops* (Pergamon), 523–534. doi: 10.1016/B978-0-08-040826-2.50041-2
- Quiros, C. F., Douches, D., and D’Antonio, V. (1987). Inheritance of annual habit in celery: Cosegregation with isozyme and anthocyanin markers. *Theor. Appl. Genet.* 74, 203–208. doi: 10.1007/BF00289969
- Razi, S., Soleimani, A., Zeinalabedini, M., Vazifeshenas, M. R., Martínez-Gómez, P., Mohsenzade Kermani, A., et al. (2021). Development of a multipurpose core collection of new promising Iranian pomegranate (*Punica granatum* L.) genotypes based on morphological and pomological traits. *Horticulturae* 7 (10), 350. doi: 10.3390/horticulturae7100350
- R Core Team (2023). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: <https://www.R-project.org/>.
- Reddy, L. J., Upadhyaya, H. D., Gowda, C. L. L., and Singh, S. (2005). Development of Core Collection in Pigeonpea [*Cajanus cajan* (L.) Mills] using Geographic and Qualitative Morphological Descriptors. *Genet. Resour. Crop Evol.* 52, 1049–1056. doi: 10.1007/s10722-004-6152-7
- Ribera, A., Bai, Y., Wolters, A.-M. A., Van Treuren, R., and Kik, C. (2020). A review on the genetic resources, domestication and breeding history of spinach (*Spinacia oleracea* L.). *Euphytica* 216, 48. doi: 10.1007/s10681-020-02585-y
- Risliawati, A., Suwarno, W. B., Lestari, P., Trikoesoemaningtyas, and Sobir, (2023). A strategy to identify representative maize core collections based on kernel properties. *Genet. Resour. Crop Evol.* 70, 857–868. doi: 10.1007/s10722-022-01469-5
- Rolling, W. R., Senalik, D., Iorizzo, M., Ellison, S., Van Deynze, A., and Simon, P. W. (2022). CarrotOmics: A genetics and comparative genomics database for carrot (*Daucus carota*). *Database* 2022, baac079. doi: 10.1093/database/baac079
- Rosental, L., Still, D. W., You, Y., Hayes, R. J., and Simko, I. (2021). Mapping and identification of genetic loci affecting earliness of bolting and flowering in lettuce. *Theor. Appl. Genet.* 134, 3319–3337. doi: 10.1007/s00122-021-03898-9
- Rubatzky, V. E., Quiros, C. F., and Simon, P. W. (1999). Carrots and related vegetable Umbelliferae. CABI Publishing.
- Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., et al. (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front. Physiol.* 3. doi: 10.3389/fphys.2012.00326
- Silva Souza, L., Cunha Alves, A. A., and De Oliveira, E. J. (2020). Phenological diversity of flowering and fruiting in cassava germplasm. *Scientia Hort.* 265, 109253. doi: 10.1016/j.scienta.2020.109253
- Simon, P. W., and Grzebelus, D. (2020). “Carrot genetics and breeding,” in *Carrots and related Apiaceae crops, 2nd ed.* Eds. G. Emmanuel and W. S. Philipp (CABI), 61–75. doi: 10.1079/9781789240955.0061
- Simon, P. W. (2019). Beyond the genome: carrot production trends, research advances, and future crop improvement. *Acta Hort.* 1264, 1–8. doi: 10.17660/ActaHortic.2019.1264.1
- Simon, P., Iorizzo, M., Grzebelus, D., and Baranski, R. (2019). *The carrot genome* (Cham: Springer International Publishing). doi: 10.1007/978-3-030-03389-7
- Simon, P. W., Rolling, W. R., Senalik, D., Bolton, A. L., Rahim, M. A., Mannan, A. M., et al. (2021). Wild carrot diversity for new sources of abiotic stress tolerance to strengthen vegetable breed.
- Solberg, S. Ø., and Yndgaard, F. (2015). Morphological and phenological diversity in Scandinavian wild carrot. *Gene Conserve* 14, 29–51.
- Tabor, G., Yesuf, M., Haile, M., Kebede, G., and Tilahun, S. (2016). Performance of some Asian carrot (*Daucus carota* L. ssp. *sativa* Hoffm.) cultivars under Ethiopian conditions: Carrot and seed yields. *Scientia Hort.* 207, 176–182. doi: 10.1016/j.scienta.2016.05.011
- Tas, P. M. (2016). *Evaluating resistance to alternaria dauci and related traits among diverse germplasm of daucus carota* (United States: The University of Wisconsin-Madison).
- Tripathi, K., Kumari, J., Gore, P. G., Mishra, D. C., Singh, A. K., Mishra, G. P., et al. (2022). Agro-morphological characterization of lentil germplasm of Indian national genebank and development of a core set for efficient utilization in lentil improvement programs. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.751429
- Tullu, A., Tar’an, B., Warkentin, T., and Vandenberg, A. (2008). Construction of an intraspecific linkage map and QTL analysis for earliness and plant height in lentil. *Crop Sci.* 48, 2254–2264. doi: 10.2135/cropsci2007.11.0628
- Upadhyaya, H. D., Ortiz, R., Bramel, P. J., and Singh, S. (2003). Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genet. Resour. Crop Evol.* 50, 139–148. doi: 10.1023/A:1022945715628
- Vavilov, N. I. (1951). Centres of origin, variation, immunity and breeding of cultivated plants. *Chronica Botanica* 13, 1–366.
- Villeneuve, F. (2020). “Carrot growth and development,” in *Carrots and related apiaceae crops*. Eds. E. Geoffriau and P. W. Simon (CABI, Wallingford), 76–91. doi: 10.1079/9781789240955.0076
- Von Maydell, D., Beleites, C., Stache, A.-M., Riewe, D., Krämer, A., and Marthe, F. (2024). Genetic variation of annual and biennial caraway (*Carum carvi*) germplasm offers diverse opportunities for breeding. *Ind. Crops Products* 208, 117798. doi: 10.1016/j.indcrop.2023.117798
- Walls, R. L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M. A., Jaiswal, P., et al. (2012). Ontologies as integrative tools for plant science. *Am. J. Bot.* 99, 1263–1275. doi: 10.3732/ajb.1200222
- Wang, Y. G., Zhang, L., Ji, X. H., Yan, J. F., Liu, Y. T., Lv, X. X., et al. (2014). Mapping of quantitative trait loci for the bolting trait in *Brassica rapa* under vernalizing conditions. *Genet. Mol. Res.* 13, 3927–3939. doi: 10.4238/2014.May.23.3
- Whitaker, T. W. (1944). The inheritance of certain characters in a cross of two american species of *lactuca*. *Bull. Torrey Botanical Club* 71, 347–355. doi: 10.2307/2481308
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Software* 4 (43), 1686. doi: 10.21105/joss.01686
- Wohlfeiler, J., Alessandro, M. S., Cavagnaro, P. F., and Galmarini, C. R. (2019). Multiallelic digenic control of vernalization requirement in carrot (*Daucus carota* L.). *Euphytica* 215, 37. doi: 10.1007/s10681-019-2360-2
- Wohlfeiler, J., Alessandro, M. S., Cavagnaro, P. F., and Galmarini, C. R. (2021). Gradient of vernalization requirement in carrot cultivars from diverse geographical origins. *Crop Sci.* 61, 3373–3381. doi: 10.1002/csc2.20526
- Wohlfeiler, J., Alessandro, M. S., Morales, A., Cavagnaro, P. F., and Galmarini, C. R. (2022). Vernalization Requirement, but Not Post-Vernalization Day Length, Conditions Flowering in Carrot (*Daucus carota* L.). *Plants* 11, 1075. doi: 10.3390/plants11081075
- Wolkovich, E. M., Burge, D. O., Walker, M. A., and Nicholas, K. A. (2017). Phenological diversity provides opportunities for climate change adaptation in winegrapes. *J. Ecol.* 105, 905–912. doi: 10.1111/1365-2745.12786
- Yui, S., and Yoshikawa, H. (1991). Bolting resistant breeding of Chinese cabbage. 1. Flower induction of late bolting variety without chilling treatment. *Euphytica* 52, 171–176. doi: 10.1007/BF00029393
- Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* 2 (12), 983–989. doi: 10.1038/35103590



OPEN ACCESS

EDITED BY

Svein Øivind Solberg,
Inland Norway University of Applied
Sciences, Norway

REVIEWED BY

Javaid Akhter Bhat,
Nanjing Agricultural University, China
Flemming Yndgaard,
Nordic Genetic Resource Centre
(NordGen), Sweden

*CORRESPONDENCE

Jenyne Loarca

✉ jloarca@wisc.edu

[†]These authors share senior authorship

RECEIVED 22 November 2023

ACCEPTED 26 February 2024

PUBLISHED 19 April 2024

CITATION

Loarca J, Liou M, Dawson JC and Simon PW
(2024) Evaluation of shoot-growth variation
in diverse carrot (*Daucus carota* L.)
germplasm for genetic improvement of
stand establishment.
Front. Plant Sci. 15:1342512.
doi: 10.3389/fpls.2024.1342512

COPYRIGHT

© 2024 Loarca, Liou, Dawson and Simon. 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.

Evaluation of shoot-growth variation in diverse carrot (*Daucus carota* L.) germplasm for genetic improvement of stand establishment

Jenyne Loarca^{1,2*}, Michael Liou³, Julie C. Dawson^{2†}
and Philipp W. Simon^{1,2†}

¹Vegetable Crops Research Unit, United States Department of Agriculture, Madison, WI, United States,

²Department of Plant and Agroecosystem Sciences, University of Wisconsin–Madison, Madison, WI, United States, ³Department of Statistics, University of Wisconsin–Madison, Madison, WI, United States

Carrot (*Daucus carota* L.) is a high value, nutritious, and colorful crop, but delivering carrots from seed to table can be a struggle for carrot growers. Weed competitive ability is a critical trait for crop success that carrot and its apiaceous relatives often lack owing to their characteristic slow shoot growth and erratic seedling emergence, even among genetically uniform lines. This study is the first field-based, multi-year experiment to evaluate shoot-growth trait variation over a 100-day growing season in a carrot diversity panel (N=695) that includes genetically diverse carrot accessions from the United States Department of Agriculture National Plant Germplasm System. We report phenotypic variability for shoot-growth characteristics, the first broad-sense heritability estimates for seedling emergence ($0.68 < H^2 < 0.80$) and early-season canopy coverage ($0.61 < H^2 < 0.65$), and consistent broad-sense heritability for late-season canopy height ($0.76 < H^2 < 0.82$), indicating quantitative inheritance and potential for improvement through plant breeding. Strong correlation between emergence and canopy coverage ($0.62 < r < 0.72$) suggests that improvement of seedling emergence has great potential to increase yield and weed competitive ability. Accessions with high emergence and vigorous canopy growth are of immediate use to breeders targeting stand establishment, weed-tolerance, or weed-suppressant carrots, which is of particular advantage to the organic carrot production sector, reducing the costs and labor associated with herbicide application and weeding. We developed a standardized vocabulary and protocol to describe shoot-growth and facilitate collaboration and communication across carrot research groups. Our study facilitates identification and utilization of carrot genetic resources, conservation of agrobiodiversity, and development of breeding stocks for weed-competitive ability, with the long-term goal of delivering improved carrot

cultivars to breeders, growers, and consumers. Accession selection can be further optimized for efficient breeding by combining shoot growth data with phenological data in this study's companion paper to identify ideotypes based on global market needs.

KEYWORDS

plant genetic resources, weed competition, seedling emergence, diverse germplasm, crop resilience, diversity panel, crop wild relatives, crop ontology

Introduction

Carrot (*Daucus carota* L.) is a widely grown vegetable crop that provides consumers with an affordable rich supply of nutrients (Drewnowski, 2013). Carrot is in the top nine most nutrient rich and cost-accessible vegetables, it is an exceptional source of beta-carotene, which functions as a provitamin A carotenoid, and offers appreciable quantities of B vitamins (thiamin, riboflavin, and niacin) compared with other commonly consumed vegetables, and is a good source of fiber (Arscott and Tanumihardjo, 2010; Drewnowski, 2010; Suchánková et al., 2015). Carrot breeding programs have prioritized breeding for traits that improve taproot quality and yield, such as flavor, color, texture, disease resistance, and pest resistance (Rubatzky et al., 1999). Excellent root quality has been a driver of carrot's wide commercial acceptance and high frequency of use (Drewnowski, 2013). However, while foliar and root diseases have received much attention by carrot researchers, little attention has been paid to the above-ground vegetative growth. Shoot vigor traits have recently become a priority for carrot growers, particularly those growers in the organic sector, who grow crops with limited weed control options. Apiaceous crops like carrot are characterized by erratic seed germination, slow emergence, slow growth, and delayed canopy closure (Rubatzky et al., 1999; Colquhoun et al., 2017). Shoot architecture and shoot biomass affect light acquisition, and consequently, slow growth limits season-long weed-competitive ability, plant growth, development, and crop productivity. Moreover, weed interference causes misshapen roots, thereby decreasing root quality and reducing root yield.

Few published studies have evaluated genetic improvement of crop stand establishment, as it has historically been managed with horticultural practices (Maynard et al., 2006). While these practices are often costly, laborious, and time-consuming, they have been the preferred management method because of their potential to provide more immediate relief than plant breeding, which is a long term strategy that requires significant investment in resources and time. However, once achieved, plant breeding solutions have the potential to save significant costs in time, and labor. To date, most studies on carrot stand establishment have attempted to address the problem

with weed management strategies or by developing treatments that increase carrot germination rate and uniformity.

In general, practices that improve crop stand establishment include irrigation methods in maize (El-Sanatawy et al., 2021), planting date by variety interaction in alfalfa (Beveridge and Wilsie, 1959), and planting depth in cereals (Hadjichristodoulou et al., 1977). Other relevant practices include intercropping, weed removal, and overplanting to compensate for poor germination or emergence. Environmental factors such as temperature, moisture, pathogens, pests, and soil health are examples of important factors that affect crop establishment (Grassbaugh and Bennett, 1998). Soil composition, soil crusting, soil water-holding capacity, and soil heterogeneity can create impedances to carrot seedling emergence if not properly managed (Hegarty, 1979). As such, bed preparation and moisture availability are critical to stand establishment. Soils with high water-holding capacity, such as heavy clay soils, are prone to crusting, which creates a hard impenetrable physical barrier through which successfully growing seedlings are unable to emerge. Sandy soils can still form a crust if not sufficiently tilled. Although sandy soil is well-draining, it has a low water-holding capacity; however, this can be mitigated with sufficient watering.

Moisture availability is also a critical factor in post-germination and pre-emergence carrot growth in the field (Finch-Savage and Pill, 1990; Finch-Savage et al., 1998), which is why seed priming is a prevalent area of research in carrot. In carrot, immature embryo dormancy partially contributes to asynchronous germination, with mature seed germinating 3.7 days earlier than immature seed (Brocklehurst and Dearman, 1980). Seed priming functions by imbuing the seed with water, which promotes the activation stage of germination, then drying the seed for later planting. This method allows less-developed seed, which may take more time to reach activation, to achieve this stage; the process synchronizes the differently matured seed. Seed priming and other seed treatments are often used to improve the rate and percentage of germination, which leads to increased synchrony and speed of seedling emergence and crop establishment (Rubatzky et al., 1999; Prohens and Nuez, 2008). Other seed treatments include seed pelleting, film coating, and fungicide treatments (Maynard et al., 2006). Even with added cost, growers see enough benefit to justify seed treatment in many cases. The positive impact of seed

treatments on crop germination have been researched in many crops, such as barley (Abdulrahmani et al., 2007), corn (El-Sanatawy et al., 2021), soybean (Arif et al., 2008), rice (Farooq et al., 2006; Ella et al., 2011), cowpea (Eskandari and Kazemi, 2011), table beet (Khan et al., 1992), and lettuce (Seale and Cantliffe, 1986). Osmotic priming has been used effectively in Apiaceous crops such as celery (Salter and Darby, 1976), parsnip (Gray et al., 1984), and carrot (Finch-Savage and McQuistan, 1988; Sanders et al., 1990; Bennett et al., 1992). Seed treatments that improve emergence and yield in carrot field studies include osmotic priming (Szafirowska et al., 1981), hormonal priming (Lada et al., 2004), and seed priming with salicylic acid (Mahmood-ur-Rehman et al., 2020). Recent research has focused on carrot seed quality screening (Marchi and Cicero, 2017) and seed treatment to improve germination (Aazami and Zahedi, 2018; Sowmeya et al., 2018; Mahmood-ur-Rehman et al., 2020; Guragain et al., 2021; Muhie et al., 2021; Muhie et al., 2024). These treatments make significant, but fractional, improvements to total carrot germination and uniform carrot germination. And while high germination rate is a necessary condition of high emergence, it does not assure high emergence.

Studies on horticultural control of weeds among carrots also outnumber studies on genetic control of carrot weed competitive ability. Weed management is a critical and proactive approach to mitigate carrot crop losses to competition, as carrot is the most sensitive (among 26 crops studied) to weed interference (Van Heemst, 1985). The first six weeks of the carrot growth cycle is known as the critical weed-free period, when slow-growing carrots are most susceptible to competition from weeds (Swanton et al., 2010). Chemical control of weeds is a common strategy in conventional carrot growing operations, but few herbicides are designated specifically for carrot. Among many herbicides (Colquhoun et al., 2019), one such example, linuron, is used for broadleaf weed control for carrot 3 - 6 weeks post-emergence. However, linuron requires that carrot shoots to achieve a threshold height of 7.6 cm before application, which is typically achieved around the fifth or sixth week after planting, which is fairly late into the critical weed-free period (Bellinder et al., 1997). In addition, lack of herbicide rotation has had an ecological impact due to the evolution of linuron-resistant pigweed populations (Colquhoun et al., 2017). Linuron is also not a management option in organic carrot growing operations (Colquhoun et al., 2017), which represent 14% of U.S. carrot production (USDA Economic Research Service (ERS), 2023), so machine- and/or hand-weeding are management strategies for organic growers. Without weeding, carrot yield losses can range from 38% - 87% (Colquhoun et al., 2017). Within-row weeding also removes late-emerging carrot seedlings that act as weed-like competitors to earlier emerged seedlings. However, hand-weeding is laborious, costly, time-intensive, and disruptive to established seedlings. Moreover, despite continuous hand-weeding, growers may still suffer an average yield loss of 15% (Colquhoun et al., 2017). Recent studies maintain focus on weed management strategies to reduce competition with carrot shoots (De Boer et al., 2019; Miao et al., 2019; Colquhoun, 2020; Ying et al., 2021; Mou et al., 2023).

Identifying genetically-controlled carrot traits that increase weed competitive ability would enable breeders to deliver carrot cultivars

that help solve the issue of weed-related yield losses for growers. For a plant breeding approach to be effective, the attributes must be heritable, and so understanding the heritability of carrot growth attributes that confer weed tolerance/suppression, such as vigorous growth, uniform emergence, and early canopy closure is essential (Colquhoun et al., 2017). There is further interest in dissecting correlation among weed-competitive traits and carrot yield, which are poorly understood given the limited number of studies on genetic architecture of carrot shoot growth. One of the earliest studies on carrot shoot and root characteristics found that early emerging seedlings tended to have less root size variation at harvest (Mann & MacGillivray, 1949) and that seed weight in turn correlates positively with improved emergence and high early root yield (Austin and Longden, 1967). Consistently, other studies found that variation in embryo size, spread of emergence, seedling size and weight at emergence all influenced taproot weight variability at harvest, but seed weight and size did not (Gray and Steckel, 1983a; Gray et al., 1986, 1983b; Salter et al., 1981). Variability in embryo length is also a reliable early indicator of root crop uniformity (Dowker, 1978). Dowker (1978) acknowledged that embryo length heritability of seedling traits had not been calculated. In a study of one cultivar, Gray (1984) misinterpreting Dowker, claimed that genetic factors of embryo length were not important and that variation in embryo length is not influenced by the genetic constitution of the seed. It is correct that non-genetic factors influence embryo length variation, and include umbel order and seed harvest date, both of which are controlled through seed-parent planting density and seed harvesting methods that eliminate underdeveloped seed (Gray and Steckel, 1985). Non-genetic sources of embryo length variation include umbel order and seed harvest date, both of which are controlled through seed-parent planting density and seed harvesting methods that eliminate underdeveloped seed (Gray and Steckel, 1985). A recent study found that 0.6% of variation in emergence is explained by varietal identity, and 70% by environmental factors, concluding that environmental conditions are more important than genetic background or intrinsic seed quality (Hundertmark-Bertaud et al., 2019). However, because Hundertmark-Bertaud only used five, genetically similar, F1 varieties of the same market class, there was likely insufficient genetic variation present to substantiate their claim that genetic factors are universally less important and environmental factors, or whether their conclusions were generalizable or not to other carrot populations, particularly those with higher genetic diversity.

There has been little research dedicated to unraveling the genetics of stand establishment traits in carrot, but a few recent studies provide strong supporting evidence that these traits have heritable variation. Carrot growers have long recognized large differences in canopy size among commercial cultivars (Simon et al., 2017). That study reported wide ranges for canopy size and canopy height in a diverse carrot germplasm collection that included purple, yellow, and red colored roots, open-pollinated varieties, segregating filial generations (F₂-F₅ populations), and inbred lines from the United States Department of Agriculture - Agricultural Research Service (USDA-ARS) Vegetable Crops Research Unit (VCRU) (Colquhoun et al., 2017), whose study of

nine commercial carrot varieties were evaluated for weed-suppressive traits in the early-mid season (Colquhoun et al., 2017). Their study observed significant differences in emergence rate, canopy development (ground cover), and weed tolerance. Genetic variation for late-season carrot shoot growth and shoot architecture has also been documented by Turner et al. (2018a), who developed an imaging pipeline that extracts size and shape traits for carrot shoots, including shoot morphology, petiole number, petiole length, petiole width, and biomass. This pipeline was then used to phenotype a F2 mapping population (N = 316) that segregated for various shoot traits, leading to the identification of quantitative trait loci (QTL) for shoot characteristics on chromosomes 1, 2, and 7, suggesting genetic control of shoot growth (Turner et al., 2018b). Canopy height variation at harvest was documented by Luby et al. (2016), who studied a panel of commercially available U.S. carrot varieties (N = 140) and found broad-sense heritability to be 0.82, suggesting that genetic factors contribute to end-of-season top height variation (Luby et al., 2016). An in-depth literature review of crop stand establishment studies can be accessed for further information (Loarca, 2021).

An extensive review of global *Daucus* germplasm collections is provided by Allender (2019). Globally, *ex situ* *Daucus* germplasm collections are extensive, with more than 13,400 accessions conserved (not yet accounting for duplicated material) across 62 institutions. Global *Daucus* genetic resources have been collected from over 75 countries, though sampling depth is low from Africa and South America (Allender, 2019; Mezghani et al., 2019). According to the Genesys database, the USDA carrot germplasm collection (1381 accessions) is among the largest *Daucus* collections, with about 695 accessions classified as cultivated and the rest wild relatives. Other international carrot germplasm collections such as the German genebank at the Leibniz Institute of Plant Genetics and Crop Plant Research (493 *Daucus* accessions), the Plant Breeding and Acclimatization Institute in Poland (629 *Daucus* accessions), and the UK Vegetable Genebank (1457 *Daucus* accessions), which was designated at the world base for carrot germplasm by the International Board for Plant Genetic Resources (IBPGR, now Bioversity). Other notable *Daucus* collections that are not cataloged by Genesys are maintained by the Vavilov Institute in Russia (3102 accessions), the Institute of Vegetables and Flowers at Chinese Academy of Agricultural Sciences (~400 accessions), the National Bureau of Plant Genetic Resources in India, and a national network 'Carrot and other *Daucus* genetic resources' (3131 accessions) in France. It was not clear from Allender's review what proportions of these collections are cultivated carrot or carrot crop wild relatives (CWR).

This study is the first and largest (N = 695 accessions) multi-year field evaluation of agronomically important shoot growth traits spanning an entire field season, from germination to emergence to harvest, in diverse carrot germplasm that includes landraces. The carrot accessions (also known as plant introductions) in this study are maintained by the United States Department of Agriculture National Plant Germplasm System (USDA-NPGS), which is a major source of useful plant genetic resources (PGR) for breeding programs, and yet one of the major barriers to using PGR is accession evaluation (Byrne et al., 2018). All or parts of this global

USDA germplasm collection have previously been evaluated in studies on canopy vigor (Loarca et al., 2024), core collection curation (Corak et al., 2019), demographic history of carrot domestication and breeding (Coe et al., 2023), genetic structure, phylogeny, and carotenoid presence (Ellison et al., 2018), taproot shape (Brainard et al., 2021), plant growth traits (Acosta-Motos et al., 2021), antioxidant capacity (Pérez et al., 2023), resistance to the necrophytic fungal pathogen *Alternaria dauci* (Tas, 2016), and several studies on seed germination under abiotic stress (Bolton et al., 2019; Bolton and Simon, 2019; Simon, 2019; Simon et al., 2021). In addition, the collection has been used for ecogeographic variation analysis (Mezghani et al., 2019), genomic core collection curation (Corak et al., 2019), and evaluation of genomic prediction strategies (Corak et al., 2023), as well as in studies of carrot CWR on subspecies identification (Spoonner et al., 2014).

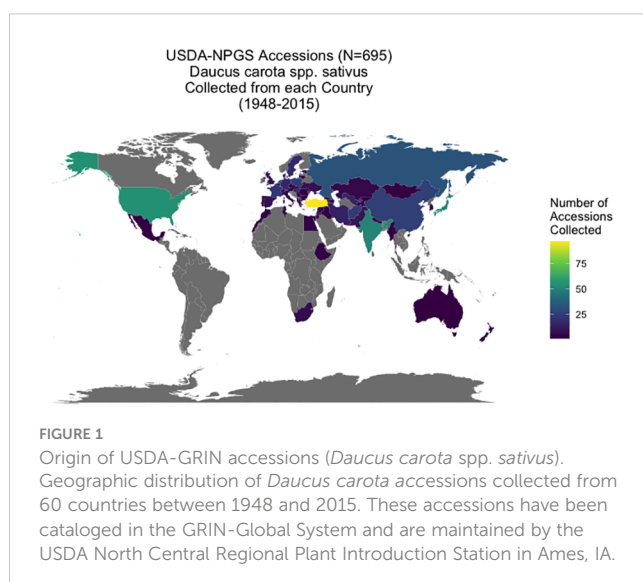
Genetic improvement of stand establishment in carrot is an achievable and desirable alternative strategy to weed mitigation and seed treatments, as these horticultural strategies have significant drawbacks with regards to expense, labor, and time for seed producers and growers. Evaluation of genetic resources is an essential activity that promotes their utilization and adaptation (Gepts, 2006). Germplasm repositories do not systematically evaluate agronomic traits in germplasm collections (Byrne et al., 2018) or morphological traits such as root color root and shape (Allender, 2019). Trait ontologies are essential to germplasm evaluation, as they provide consistent and shared vocabulary with agreed-upon definitions to accurately and consistently document and describe plant phenotypes (Shrestha et al., 2010; Walls et al., 2012). Previous trait ontologies for carrot, such as the now-defunct RoBuST, included traits such as carotene content, flavor, pungency, lutein, pathogen resistance, xylem/phloem color, and root shape (Bhasi et al., 2010). Above-ground vigor and biomass were not part of the traits included in that system. The recently established CarrotOmics database has a far more extensive suite of traits, including 280 traits defined, most of which pertain to the carrot root and very few characterize pertaining to the carrot shoot (Rolling et al., 2022). Given that recent studies found moderate heritability of shoot growth traits among cultivars (Luby et al., 2016) and mapping populations (Turner et al., 2018b), we hypothesized that heritable variation can be found in diverse carrot germplasm. Characterization of the germplasm collection broadens the genetic base that breeders can leverage in carrot breeding programs, and provides useful recommendations of accessions to include in breeding programs targeting improvement of seedling vigor, stand establishment, and weed competitive ability.

Materials and methods

Population under study

Daucus accessions (N = 1381) are maintained through the USDA-NPGS at the North Central Regional Plant Introduction Station (NCRPIS) in Ames, IA, with information on the accessions in the Germplasm Resources Information Network (GRIN) database of the NPGS (GRIN-Global, 2023). From a genetic perspective, a

carrot accession is a heterogeneous, heterozygous population increased by open-pollination and is genetically distinct from other accessions in the collection. This carrot collection represents global carrot germplasm, collected over multiple plant exploration trips between 1947 and 2015 from 60 countries. Over 80% of accessions originate from the Eurasian supercontinent: 53% from Asia, 34% from Europe and the Caucasus, and 13% (in descending order) were collected from the Americas, Africa, Australia, and New Zealand. At least 148 accessions originated in the primary center of diversity in central Asia (modern-day Afghanistan and surrounding countries) and secondary center of diversity in western Asia (modern-day Turkey) (Vavilov, 1951; Banga, 1963) (Figure 1). Passport data from GRIN-Global also provided seed viability (average germination percentage based on four independent replicates of 50 seeds per accession at standard germination conditions) and weight of 100 seeds in grams (average of two replicates of 100 seeds) for each accession (personal communication: Kathleen Reistma). Accessions were selected for this study's diversity panel if they had passport information suggesting they were domesticated or exhibited domestication traits in preliminary screening, as evidenced by taproot traits such as increased pigmentation (carotenoid or anthocyanin), reduced lateral root branching, and increased taproot size (Ellison et al., 2018), resulting in 695 cultivated carrot accessions for this study. Many of these accessions are considered landraces or heirloom cultivars with annual, biennial, or mixed flowering habit. Although biennial flowering habit is a known domestication trait in carrot (Alessandro et al., 2013; Ellison, 2019), accession flowering habit data from the gene bank (annual, biennial, or mixed population) was of limited use due to having been phenotyped in various environments. Consequently, flowering habit was not a criterion used in curating this cultivated diversity panel. However, flowering habit data was collected during the course of this study and is the subject of this article's companion paper (Loarca et al., 2024). A list of accessions used in this study from GRIN-Global is available (Supplementary Table 6) and raw data for each accession is available on the CarrotOmics database (Rolling et al., 2022).



Experimental design

In mid-May in 2016, 2017, and 2018, we hand-planted one replicate of each accession in each of two blocks of a randomized complete block design (RCBD) at the Hancock Agricultural Research Station (ARS). Seeds were hand-planted in meter-length plots, in two adjacent hand-created furrows, 2 cm apart, at an approximate planting depth of 0.5 cm, with approximately 50 seeds per plot. Planting beds were prepared with a bed shaper. Each planting bed was 6 meters long and 1.7 meters wide, which provided enough space for 18 1-meter plots (6 plots lengthwise and 3 plots widthwise). Rye grass was planted in between the plot rows to maintain the structure of the sandy beds. Weeds were suppressed throughout the season with regular herbicide application. The field was watered with overhead irrigation. In 2016, plots with more than 50 plants were thinned to 50 plants per plot. This research station is in the central sands region of central Wisconsin, which is the third largest carrot producing state in the U.S. (\$8.5M; 4,100 acres; 92K metric tons produced), making this a highly relevant target environment for identifying high-performing accessions (USDA Economic Research Service (ERS), 2023). This region has a characteristic sandy soil that is amenable to carrot seedling emergence and reflects optimal conditions for carrot cultivation.

CarrotOmics shoot-growth ontology and trait evaluation

We included new shoot growth traits and descriptions, as well as elaborations to previously defined traits in the CarrotOmics database (Rolling et al., 2022). Traits measured (Table 1) include stand count (20 DAS), percent emergence (20 DAS), canopy height (40, 80, and 100 DAS), and canopy coverage (50, 80 and 100 DAS). In Table 1, these traits are described alongside existing trait descriptions in CarrotOmics. Stand count (20 DAS) refers to the total number of plants with cotyledons emerged in each plot for each accession. Emergence percentage is a transformation of stand count, calculated from the observed stand count divided by the expected number of plants (i.e., the number of seeds planted; approx. 50) (Figure 2). Canopy height (80 and 100 DAS) was measured on each plot at three randomly selected points within the plot, from root shoulder to top of leaf canopy (Turner et al., 2018a). We defined 'canopy coverage' as a visual estimate of the proportion of the soil obscured by carrot top-growth vegetation when viewed from a single point above the plot (Figure 3). We evaluated canopy coverage (50 DAS) using a five-point scoring system (0%, 25%, 50%, 75%, or 100%).

Data management

All statistical analyses were performed using RStudio Version 2023.6.1.524 (Posit team, 2023) and R Version 4.3.1 (R Core Team, 2023). Rosner's Test in the *EnvStats* identified multiple simultaneous potential outliers for each trait in each year (Millard, 2013). Data was subset and manipulated with the

TABLE 1 CarrotOmics shoot-growth trait ontology developed in this paper, compared with current trait descriptions in GRIN-Global This paper elaborates on traits that were previously recognized as important in CarrotOmics and provides standard methodologies that carrot researchers can follow, enabling collaboration across programs.

CarrotOmics Trait Ontology for Shoot-Growth	2023 GRIN-Global Shoot-Growth Trait Descriptors
Stand Count Absolute number of seedlings emerged within a field plot. 14 (early vigor), 20 (standard measurement)	Seedling Vigor "1=good" (GRIN-Global)
Percentage Emergence Stand count divided by the number of seeds planted in plot 14 (early vigor), 20 (standard measurement)	- -
Canopy Coverage Visual estimate of the proportion of the soil obscured by carrot top-growth vegetation when viewed from a single centered point above the field plot at notetaker's eyeline. Measured visually on a 0% - 100% scale in increments of 25%. 50 (early-season), 80 (late-season), 100 (harvest day)	Shoot Biomass "Estimate of mass of all shoot tissue more than 4 cm above the crown, obtained from image analysis" (Turner 2018a).
Canopy Height Measured at various times throughout the season, with three random measurements taken per plot. 40 (early vigor), 80 (late-season vigor), 100 (harvest day)	Canopy Height at Harvest "Canopy height was measured just before harvest with three measurements taken per plot" (Turner 2018a).

Trait evaluation time is given in days after seeding (DAS) that the measurement was taken. Expected data collection times for the same plant growth stage in other programs may vary by location, cultivar, market type, and length of growing season.

tidyverse suite of packages (Wickham et al., 2019). Rosner's Test implemented in *EnvStats* was used to identify multiple possible outliers for each trait in each year (Millard, 2013). A variety of utility packages were critical to data analysis, including *ggthemes* (Arnold, 2021), *beepr* (Bååth, 2018), and *flextable* (Gohel and Skintzos, 2023).

Analysis of variance

F-tests of significance were performed using fixed effects models in a two-way Analysis of Variance (ANOVA) with Type III sums of squares using the *car* package (Fox and Weisberg, 2019). For each year, a fixed effects model was structured to calculate the proportion of phenotypic variance for each trait attributable to genotype: $T_{ik} = u + g_i + b_k + e_{ik}$, where T = phenotypic measurement of the trait of interest (emergence, canopy coverage, or canopy height), g_i = genotype, b_k = block, and e_{ik} = error with $e_{ik} \sim \text{i.i.d. } N(0, \sigma^2)$.



FIGURE 2 Carrot Seedling Development. (top) Week 1 seedlings (cv. Bolero). Coleoptile and mesocotyl approx. 2-2.5 cm in length. Radicle approx. 1-1.5 cm in length. First true leaf at the base of coleoptiles is barely visible to the naked eye. Scale marker on left in centimeters. Artifacts in the background are seedling shadows. (middle) Week 2 seedlings (cv. Bolero). Coleoptile and mesocotyl approx. 3.5-4 cm in length. Radicle approx. 3-6 cm in length. True leaves (1-2) are clearly visible. Scale marker on left in centimeters. (bottom) Week 3 seedlings (cv. Bolero). Coleoptile and mesocotyl approx. 6-8 cm in length. Radicle approx. 5-7 cm in length. True leaves (2-4) are clearly visible. Scale marker on left in centimeters.

The multi-year fixed effects model included trait data from multiple years and calculated the proportion of phenotypic variance in each trait attributable to g_i = genotype, y_j = year, $(gy)_{ij}$ = genotype*year interaction, $b_{k(j)}$ = block within year: $T_{ijk} = u + g_i + y_j + (gy)_{ij} + b_{k(j)} + e_{ijk}$, where T = phenotypic measurement of the trait of interest and e_{ijk} = error with $e_{ijk} \sim \text{i.i.d. } N(0, \sigma^2)$. Due to unbalanced data from abnormal weather events (destructive hail), we ran two multi-year ANOVAs: one that included the 2017 shoot-growth data and one that excluded the 2017 shoot-growth data.



FIGURE 3

Canopy coverage is defined by the CarrotOmics shoot-growth ontology as the proportion of the ground covered by carrot foliar biomass. Photograph taken approx. 150 cm above the meter-length plot. Photos below are from five independent plots with canopy coverage in descending order from 100% carrot canopy coverage (left) to 0% carrot canopy coverage (right).

Broad-sense heritability and variance components

Variance components were estimated for each trait within-year (single-year) and across-years (multi-year) using random effects models with the lme4 package (Bates et al., 2015). Variance components for each trait were then used to calculate broad-sense heritability (H^2) within and across years. Statistical analysis used the same models as for the Analysis of Variance described in the previous section, but with all effects random. Broad-sense heritability (H^2) for each trait, within years (single-year model) and across years (multi-year model), was calculated from variance components, including genotypic variance (V_g) and phenotypic variance (V_p). As in the fixed effects ANOVA, we ran two multi-year analyses: one that included the 2017 data and one that excluded the 2017 data.

Single-year broad-sense heritability was calculated for each trait:

$$H^2 = \frac{V_g}{V_p} = \frac{V_g}{V_g + \frac{V_{error}}{\# \text{ rep}}}$$

Multi-year broad-sense heritability was calculated for each trait:

$$H^2 = \frac{V_g}{V_p} = \frac{V_g}{V_g + \frac{V_{gy}}{\# \text{ years}} + \frac{V_{error}}{\# \text{ years} * \# \text{ reps}}}$$

Mixed models and estimated marginal means

For each year and each trait, mixed models were fit using the same models as above with genotype as fixed effect and year and block as random effect (Bates et al., 2015). Estimated marginal means for each trait within each year were extracted from this model using the emmeans package (Lenth, 2023). Pearson's sample correlation was used to calculate trait relationships and stability across years. Smoothed curves between traits were fit using Locally

Estimated Scatterplot Smoothing (LOESS). Correlation coefficients (upper panels), curvilinear regression (lower panels), and trait distributions across years (diagonals) are summarized in a correlation matrix (Figure 4). These single-year estimated marginal means were used as phenotypes and summarized in Table 2. As in the prior models, we ran two multi-year mixed models: one that included the 2017 shoot-growth data and one that excluded the 2017 shoot-growth data.

Results

Descriptive statistics

Germination data was collected by NCRPIS for accessions planted in this trial ($79.7\% \pm 15.9\%$). Trait averages were consistent in 2016 and 2018 (42% - 46% emergence, 49 - 54 cm height, and 52% canopy coverage) (Table 2). Average emergence at 20 DAS (Table 2) is consistent with previous reports in carrot (35% - 77%) (Heydecker, 1956). Average canopy coverage (50 DAS) is within range of previous reports (32.5% - 80%) (Colquhoun et al., 2017), though we report a wider breadth of canopy coverage observations (0% - 100%). In 2017, average emergence (17%), canopy height at 80 and 100 DAS (28 cm and 35 cm, respectively), and canopy coverage (37%) were extremely low compared to 2016 and/or 2018. Diagonal panels in Figure 4 convey that trait distributions are visually consistent with descriptive statistics, demonstrating that trait performance (average and variance) are similar for 2016 and 2018, while 2017 observations are lower for all traits.

Analysis of variance and broad-sense heritability

ANOVA results for all traits and years indicated that genotype was a highly significant factor ($p < 0.001$) influencing phenotypic variation. Broad-sense heritability estimates consistently indicated

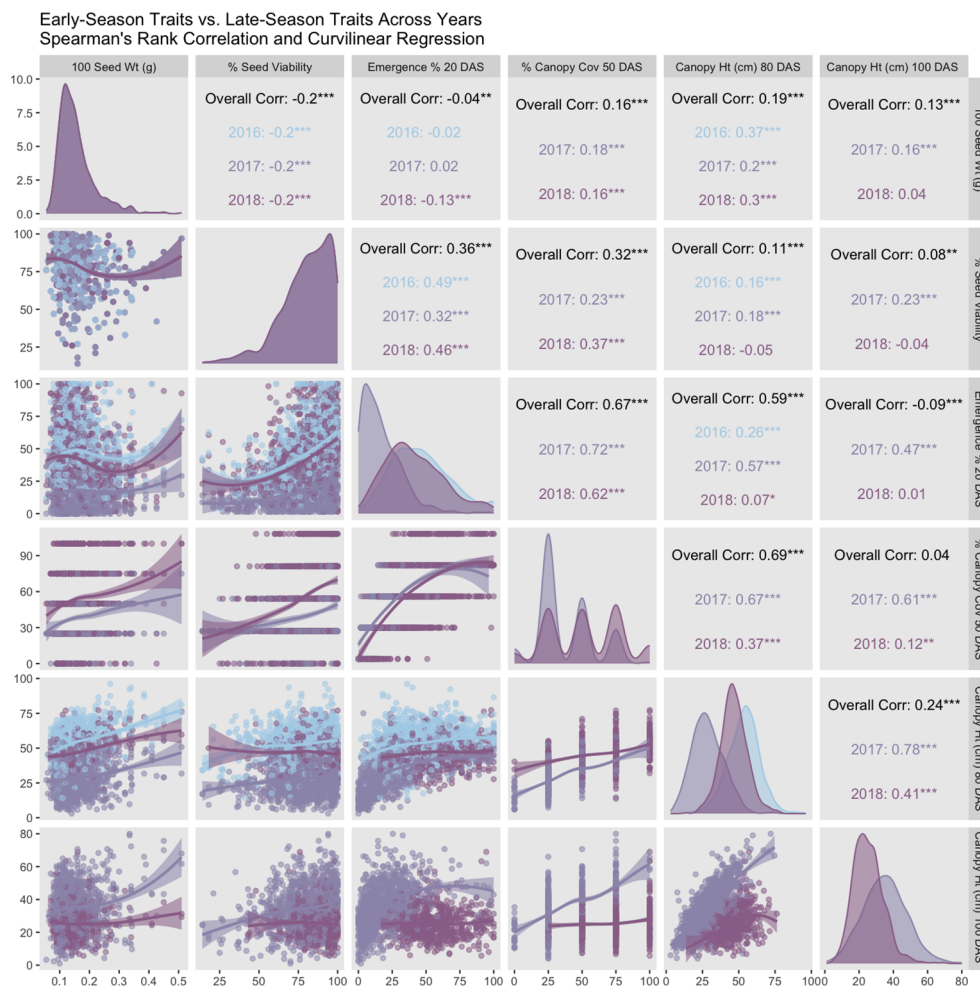


FIGURE 4

Pearson correlation matrix and curvilinear regression of early-season vs. late-season traits (2016–2018). Coefficients in the matrix indicate high correlation between both early-season traits. Similarly, both late-season traits are highly correlated. However, in normal years (2016 and 2018), both early-season traits are correlated weakly, if at all, with late-season traits. Anomalously, correlation is high between all traits in 2017, which is a year with poor stand establishment.

moderately high heritability across all years (Table 3) for emergence ($0.61 < H^2 < 0.72$) and canopy coverage ($H^2 = 0.65$). Because emergence percentage is a transformation of stand count, their ANOVA outputs broad-sense heritability values were identical; we chose to present data only on emergence percentage, which is more generalizable to other studies and more intuitive than stand count. Broad-sense heritability estimate for canopy height was moderately high for all years at 80 DAS ($0.64 < H^2 < 0.82$) (Table 4A) and at harvest day (100 DAS) ($0.77 < H^2 < 0.78$) (Table 4B). These estimates are consistent with previous studies of end-of-season canopy height heritability estimates in carrot ($0.65 < H^2 < 0.82$) (Luby et al., 2016; Turner et al., 2018b). ANOVA results indicated that genotype is a highly significant source of canopy height variation, both in single-year (Tables 4A, B) and multi-year (Table 5) models. When excluding 2017 data, the year and block in year terms are not significant, however genotype and genotype \times year effects remained highly significant (Table 6). Calculated p-values for shoot-growth ANOVA results are available in Supplementary Tables 1–5.

Trait correlations

Seedling emergence (20 DAS) and canopy coverage (50 DAS) demonstrated moderately strong correlation ($0.62 < r < 0.72$) across 2017 and 2018, despite poor emergence and poor stand in 2017. Seed viability (germination %) correlated moderately with seedling emergence ($0.38 < r < 0.55$).

In 2016 and 2018, emergence correlated poorly with canopy height 80 DAS ($0.07 < r < 0.26$) and canopy height 100 DAS ($r < 0.013$), while in 2017, correlations were moderate ($0.47 < r < 0.58$). Correlations with canopy coverage (50 DAS) were more variable with canopy height 80 DAS ($0.37 < r < 0.67$) and 100 DAS ($0.12 < r < 0.62$), with lower values representing normal years 2016 and 2018 and the higher value representing 2017 with poor stands overall. Curvilinear regression (Figure 4, lower panels) varies by year between emergence and canopy height, ranging from $0.07 < r < 0.57$. Because 2017 had poor emergence, and 2018 did not record height on plots with fewer than 20 plants, it is likely that the 2016 dataset conveys the most

TABLE 2 Descriptive statistics for early-mid season traits and late season traits.

CarrotOmics Shoot Growth Traits	2016		2017		2018		2016 & 2018	
	mean	SD	mean	SD	mean	SD	mean	SD
Emergence % (20 DAS)	46.38	21.82	17.23	14.98	41.73	21.69	44.67	16.63
Stand Count (20 DAS)	23.19	10.91	9.48	8.24	22.95	11.93	22.33	8.32
Canopy Coverage % (50 DAS)	-	-	37.55	20.61	52.47	26.73	-	-
Canopy Height (cm) (80 DAS)	53.42	10.98	28.76	10.84	46.62	9.15	49.47	9.49
Canopy Height (cm) (100 DAS)	-	-	35.08	12.65	25.46	8.73	-	-

Trait evaluation time is given in days after seeding (DAS) that the measurement was taken. Canopy coverage was measured on one replication in 2017. Late season canopy height was only recorded on accessions with stand count > 20 plants in 2018.

accurate correlation ($r = 0.256$) between emergence and canopy height (80 DAS) in this study. Overall, early-season traits correlated highly with each other, and late-season traits correlated highly with each other (Figure 4, upper panels). Hundred-seed weight (g) did not correlate remarkably with any other traits in this study.

Discussion

This study is the first multi-year study to systematically evaluate shoot-growth traits, from seed to harvest, in a global carrot diversity panel. This germplasm characterization has resulted in the identification of weed-competitive accessions that are of immediate utility to carrot breeders targeting improvement of stand establishment, particularly in organic carrots, and will result in reduced labor and cost associated with weeding, herbicide applications, and seed treatments. Strong broad-sense heritability for all traits measured indicates their potential to be improved through plant breeding. A program targeting early-season crop success would do well to focus on improving emergence. Strong correlation between emergence and canopy coverage suggests that improvement of seedling emergence has great potential to increase yield (through increased total number of individual carrot plants) and weed competitive ability (as all viable carrot plants contribute to canopy coverage). Accessions with vigorous emergence and canopy coverage provide breeders with raw materials for improving stand establishment in elite germplasm, increasing

cultivar options for organic carrot farmers. Moderate correlation between germination percentage and seedling emergence suggests that, while necessary to produce a viable plant, it does necessarily predict successful emergence and crop establishment.

Broad sense heritability

This study has demonstrated the breadth of variation for top growth traits present in a global carrot diversity panel, and is the first study to provide broad-sense heritability estimates for early-season seedling and shoot vigor. The broad sense heritability estimates we obtained demonstrate sufficient genetic control of emergence and vigor to be useful to breeding programs, however, this useful genetic variation may require breeders to use germplasm outside of current elite pools. We have also provided improved descriptions for agronomically-important shoot-growth traits in carrot, including standard methodologies and time-frames for trait evaluation. Data collection times were measured for central Wisconsin and will need to be adjusted according to location, cultivars, market type, and length of growing season. In addition, researchers will need to determine the optimal time-frames for trait evaluation for germplasm in their target locations. An inherent challenge of working with diverse germplasm of outcrossing crop species is intra-accession genetic variability as a source of unquantified variation in this study. Controlling for the level of inbreeding across accessions would correct the bias of inbreeding

TABLE 3 ANOVA and broad-sense heritability (H^2) for seedling emergence and canopy coverage.

Source of Variation	Emergence % (20 DAS)									Can. Cov. (50 DAS)		
	2016			2017			2018			2018		
	df	F	p	df	F	p	df	F	p	df	F	p
Genotype	684	3.50	***	693	2.59	***	693	2.70	***	694	2.80	***
Block	1	3.78	0.052	1	90.44	***	1	1.17	NS	1	10.39	***
Residuals	684			691			674			682		
H^2	0.72			0.61			0.63			0.65		

ANOVA results indicate that genotype is a highly significant factor in all years. Broad-sense heritability (H^2) is moderately high. P-values available in [Supplementary Table 1](#).

TABLE 4A ANOVA and broad-sense heritability (H^2) of canopy height (80 DAS) (2016–2018).

Source of Variation	Canopy Height (cm) 80 DAS								
	2016			2017			2018		
	df	F	p	df	F	p	df	F	p
Genotype	648	5.46	***	674	2.72	***	458	3.24	***
Block	1	0.74	NS	1	106.54	***	1	0.14	NS
Residuals	616			591			270		
H^2	0.82			0.64			0.74		

Statistically significant at * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; NS = otherwise. ANOVA results indicate that genotype is a highly significant factor for late-season canopy height and broad-sense heritability (H^2) is moderately high for all three years studied. P-values available in [Supplementary Table 2](#).

depression or hybrid vigor from recent outcrossing and high levels of accession heterozygosity.

Genotype was a highly significant factor in all traits in this study, demonstrating moderate to high broad-sense heritability for several agronomically important carrot shoot growth traits, and reporting the first broad-sense heritability estimates for seedling emergence ($0.68 < H^2 < 0.80$) and canopy coverage ($0.60 < H^2 < 0.66$). Heritability estimates for emergence fill the gap in heritability estimates identified by [Dowker \(1978\)](#). Our study contradicts [Gray \(1984\)](#), who claimed that genetic factors are not important when evaluating carrot seedling vigor on one carrot cultivar (cv. Red-cored Chantenay). Our results also disagree with a recent estimation that 0.6% of variation in emergence was explained by varietal identity, while 70% of variation was due to environment ([Hundertmark-Bertaud et al., 2019](#)). A major reason for their low estimate of genotypic variation could be due to their population under study, which included five F1 varieties of the market class Nantaise. Their conclusion that environmental conditions are more important than genetic background does not apply when evaluating a large genetically diverse and representative panel. All three years of our study indicated that genotype was a significant source of variation, and moderately-high broad-sense heritability for emergence supports our hypothesis that genetics have an important influence on variation in seedling vigor.

Broad-sense heritability for early-season canopy height ($0.76 < H^2 < 0.82$) was similar to our broad-sense heritability estimate for late-season canopy height, and within range of plant height and

canopy height estimates in previous studies (0.67 and 0.82, respectively) ([Luby et al., 2016](#); [Turner et al., 2018b](#)). While these previous studies evaluated late-season carrot canopy height, our study provides the first broad-sense heritability estimates for early-season seedling vigor and early-season shoot vigor, which demonstrated high broad-sense heritability. Consequently, we recommend selection on early-season characteristics to improve weed competitiveness, with a goal of balancing rapid early season growth and moderate end of season biomass.

Correlations among shoot growth traits

In 2017, a post-emergence hailstorm severely reduced stand, height, and canopy coverage. In 2016 and 2018, years in which hail damage did not occur, emergence correlated poorly with canopy height 80 DAS ($0.07 < r < 0.26$) and canopy height 100 DAS ($r < 0.013$), while in low-stand years like 2017, where hail damage occurred, correlations were more moderate ($0.47 < r < 0.58$) ([Figure 4](#), upper panels). Years 2016 and 2018 were more normal years and comparable for stand count and late-season canopy height. However, the 2018 dataset excluded height on plots with fewer than 20 plants, and consequently the 2016 dataset likely conveys the most accurate correlation ($r = 0.256$) between emergence and canopy height (80 DAS) in this study. Emergence correlates weakly with late-season canopy height, which was unexpected, given the well-known density-dependent shade etiolation response, or shade avoidance, in plants.

TABLE 4B ANOVA and broad-sense heritability (H^2) of canopy height (100 DAS) (2017 & 2018).

Source of Variation	Canopy Height (cm) 100 DAS					
	2017			2018		
	df	F	p	df	F	p
Genotype	624	4.09	***	401	3.82	***
Block	1	5.95	*	1	13.04	***
Residuals	508			226		
H^2	0.77			0.78		

Statistically significant at * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; NS = otherwise. ANOVA results indicate that genotype is a highly significant factor for late-season canopy height and that broad-sense heritability (H^2) is moderately high for all years studied, despite the presence of the disease, ALB. P-values available in [Supplementary Table 3](#).

TABLE 5 Multi-year ANOVA and broad-sense heritability (H^2) of emergence (20 DAS) and canopy height (80 DAS) (2016-2018), and canopy height (100 DAS) (2017 & 2018).

Source of Variation	Emergence (20 DAS)			Canopy Ht. (80 DAS)			Canopy Ht. (100 DAS)		
	df	F	p	df	F	p	df	F	p
Genotype	663	3.81	***	648	4.43	***	380	4.90	***
Year	2	9.53	***	2	15.76	***	1	0.22	NS
Genotype x Year	1328	1.63	***	1093	1.44	***	380	1.68	***
Block in Year	3	19.08	***	3	47.97	***	2	7.67	***
Residual	1974			1469			554		
H^2	0.72			0.79			0.80		

Statistically significant at * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; NS = otherwise
ANOVA results indicated that genotype and genotype x year interaction were highly significant factors across all years. Broad-sense heritability (H^2) is moderately high for all traits. P-values available in [Supplementary Table 4](#).

Canopy coverage did not correlate with either canopy height measurements in years with normal weather. Low correlation between emergence and canopy height may also be explained by the nature of working with a diversity panel – there may be a variety of genetic responses to plant competitive conditions (Ballaré et al., 1994), including density-dependent self-thinning and shade tolerance (Westoby, 1984; Lonsdale, 1990). Similarly, canopy coverage has low correlation with both canopy height measurements. However, given that carrot tops have a unique morphology, with no internodes and long flexible petioles that bend under the weight of their own foliar growth (known as ‘canopy closure’), this may be unsurprising. This observation is consistent with Turner et al. (2018b), who also observed low correlation ($0.3 < r < 0.4$) between biomass and shoot height, as well as with shoot area and shoot height. Shoot biomass and shoot area, however, were highly correlated ($r = 0.91$).

Correlations between early-season traits and late-season traits were higher in 2017. This response is not consistent with low densities plots in 2016 or 2018. The 2017 results may accurately represent the kind of correlation that is typical of poor-stand years or when intentionally planted at low planting density. This response could indicate a tendency for these carrot shoots to grow into the space available, or to thrive in the absence of weed competition. Given that cultivated carrot competes poorly with weeds, this could be a

reasonable and adaptive response to lack of competition. Further studies with multiple controlled densities would clarify this idea.

Previous studies measuring carrot canopy coverage reported an average canopy coverage of 66% and range of 32.5% - 80% at 55 DAS (Colquhoun et al., 2017). Our reported average canopy coverage (37% - 52%) was lower (which can be expected for unadapted germplasm) though we report a higher upper range (100%), which has important implications for improving canopy coverage through breeding. The greater variation reported for canopy coverage and emergence indicates that there are accessions in this collection with greater emergence and canopy coverage potential than some commercially available cultivars. The high correlation between emergence and canopy coverage suggests that seedling vigor may be an early-season predictor of mid-season canopy cover, crop vigor, and crop competitiveness.

Factors interacting with measurement of shoot-growth traits

While height across the season may confer weed-competitiveness, it is not sufficient to select only for populations with the largest plants because excessive foliar biomass can impede

TABLE 6 Multi-Year ANOVA and broad-sense heritability (H^2) of emergence (20 DAS) and canopy height (80 DAS) (2016 & 2018).

Source of Variation	Emergence % (20 DAS)			Canopy Height (80 DAS)		
	df	F	p	df	F	p
Genotype	663	3.21	***	648	5.55	***
Year	1	1.26	NS	1	0.87	NS
Genotype x Year	662	1.54	***	439	1.30	***
Block in Year	2	1.65	NS	2	0.44	NS
Residual	1310			886		
H^2	0.67			0.83		

Statistically significant at * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; NS = otherwise
ANOVA results indicated that, when excluding the abnormal 2017 data, year and block in year terms are not significant.-Genotype and genotype x year remained significant and broad-sense heritability (H^2) is moderately high for all traits. P-values available in [Supplementary Table 5](#).

the inner workings of machine harvesters. Furthermore, the advantage of early-season vigor can become a liability by late-season, as large canopies create a humid microclimate in which fungal pathogens, such as *Alternaria dauci*, a necrotrophic fungus that can readily infect leaves, causing Alternaria Leaf Blight (ALB) (Prohens and Nuez, 2008). ALB was present during all three years of our study, and is the most economically devastating carrot pathogen that is present in most carrot production areas (Tas, 2016). Therefore whole-plot mortality or foliage reduction caused by ALB had a confounding effect in late-season carrot shoot trait evaluation. Despite ALB's destructive impact on foliar biomass, broad-sense heritability estimate for canopy height was still moderately high for all years at 80 DAS ($0.64 < H^2 < 0.82$) (Table 4A) and at harvest ($0.77 < H^2 < 0.78$) (Table 4B). It is not yet clear how these accessions would perform in an environment where this pathogen is well-controlled. Mitigating ALB as a source of noise would strengthen genetic signal for late-season canopy growth in future evaluations.

Despite high germination rates in this collection ($79.7\% \pm 15.9\%$), average field emergence ranged from 42% - 52%, even in the presence of sufficient moisture and amenable soil and bed conditions. The gap between potential emergence and actual emergence has long been known in carrot (Heydecker, 1956; Hegarty, 1971; Finch-Savage and Pill, 1990; Finch-Savage et al., 1998), and the values reported in our study are consistent with previous studies (35% - 77% emergence) on untreated carrot seeds (Heydecker, 1956). We observed accessions with germination rates upward of 90% in replicated lab tests that demonstrated very low emergence and canopy coverage in the field. There are many potential mechanisms and points of failure between germination and emergence, that warrant significant attention in future studies (Steiner and Zuffo, 2019) in diverse carrot germplasm. Accessions or seed lots that have high germination under standardized or controlled laboratory conditions will not necessarily germinate or emerge under field conditions. Therefore, while high seed germination is a prerequisite to seedling growth, it does not necessarily predict successful seedling growth, development, and emergence in the field. Therefore, while studies on seed priming may improve carrot seed germination in controlled test conditions (Aazami and Zahedi, 2018; Sowmeya et al., 2018; Mahmood-ur-Rehman et al., 2020; Guragain et al., 2021; Muhie et al., 2021; Muhie et al., 2024), which is an important part of the puzzle, these studies will have stronger potential to identify vigorous, agronomically useful germplasm when combined with a field emergence study, to develop a complete package for early-season agronomic performance under real world conditions (Simon et al., 2021). Similarly, recent germination studies could be improved upon (Aazami and Zahedi, 2018; Bolton and Simon, 2019; Bolton et al., 2019; Nijabat et al., 2023) by screening germplasm in field emergence trials, especially given the long-known and well-established gap between lab germination and field emergence (Heydecker, 1956; Hegarty, 1971; Finch-Savage and Pill, 1990; Finch-Savage and McQuistan, 1988). Additional traits to measure include seedling vigor and emergence speed (Acosta-Motos et al., 2021).

Breeders have traditionally relied on well-adapted germplasm for development of improved cultivars. Our study empowers breeders to identify accessions with desirable top-growth traits

that can be leveraged to invigorate breeding or pre-breeding programs with useful genetic diversity for shoot-growth and weed-competitive traits. We recommend that breeders interested in improving season-long weed competitiveness incorporate these trait measurements in their breeding and variety trial evaluations. Additional metrics of emergence that incorporate growth uniformity have important implications for end-of-season root yield. Future studies on seedling vigor would benefit from additional measurements of uniformity, such as emergence timing and seedling growth rate. While studying emergence requires uniform seeding rate, studying canopy height requires uniform emergence to achieve uniform planting density, and consequently, uniform intraplant competitive conditions. This would necessitate overplanting accessions with low germination or low emergence to achieve uniform planting density conditions across all accessions.

Furthermore, optimal planting density varies depending on carrot root shape. While all accessions in this study had the same number of seeds per plot, future studies may benefit from considering the relationship between market class and planting density. This carrot germplasm collection was recently evaluated for root shape and market class, though not all cultivars fit cleanly into a particular market class (Brainard et al., 2021b). Bulkier carrots are grown at lower population densities than slimmer fresh market types at higher densities (1,500,000 - 3,000,000 plants per hectare) – these densities were established to produce high levels of biomass on a *per hectare* and a *per root* basis (Goldman, 2019). However, planting densities above the optimal rate can reduce individual root biomass by 50% (Vega and Goldman, 2023). Integrating this data with known optimal planting densities for carrot market classes in this germplasm collection, to the extent possible, would enable optimal plant density, rather than uniform seed rate – this is important because it is not yet known how planting density of various market types affects shoot growth. Accounting for planting density would enable accessions with similar optimal planting densities to be compared, given that planting density is well-documented as a source of variation in above-ground biomass and morphology (Duthie et al., 1999; Peil and López-Gálvez, 2002; Aziz et al., 2007; Goss, 2012; Khan et al., 2017; Postma et al., 2021). Non-optimal planting densities for the certain root shapes in this collection could partially explain the unexpected lack of correlation between stand count and height in our study.

Measurement accuracy and labor

Stand count provides very valuable information about genotype performance because early emergence correlates with root uniformity at harvest (Mann and MacGillivray, 1949), one of the primary components of marketable root yield. The moderately high broad-sense heritability estimates reported in our study suggest that the current phenotyping methods we presented are capable of detecting a genetic signal among a diverse set of germplasm. However, stand count is the most physically laborious and time-consuming trait to measure in this study. The time required to phenotype one plot increases with emergence and planting density.

Because carrot seedlings at 20 DAS are still very small (2–7 cm) and typically densely planted, stand count data collection requires technicians to bend, kneel, or squat over the plot. Moreover, counting requires manually separating the plants, which is disruptive to established seedlings.

Canopy coverage phenotyping is the fastest of all methods described, requiring fewer than five seconds to assign a value to each plot and can be recorded from an upright position (i.e., no bending or squatting required). Canopy height data collection requires less than one minute per plot with one technician, but is more efficient with two technicians, as one records data while the other reports the data. Moreover, because canopy height measurements can also require bending or squatting, sharing the load between two technicians reduces laborious repetitive motions. Other carrot studies have used a similar scoring method to visually estimate canopy coverage on a continuous scale (0%–100%) to measure ‘ground cover’ or ‘carrot canopy development’ in an experiment that included nine entries (Colquhoun et al., 2017). In contrast to a continuous scale, the five-point visual scale improved phenotyping efficiency, which was critical on an experiment of this size. It is not clear how much more accuracy is gained from a continuous vs. categorical visual estimation. Evaluation of canopy coverage on a continuous scale could smooth the distribution for canopy coverage (Figure 4, diagonal panel for canopy coverage). Drone imaging could convert this measurement to a continuous trait, potentially improving estimates of canopy coverage, but requires significant investment in training, equipment, software, and analysis that not all programs necessarily have access to.

The canopy-coverage phenotyping method presented in this study demonstrated sufficiently high broad-sense heritability for canopy coverage ($H^2 = 0.65$) to begin to make progress on evaluations, selections, and genetic gain on canopy coverage. Beyond potential gains in phenotyping accuracy, image-based phenotyping of field plots would eliminate the risk of repetitive motion injuries while collecting data in the field. Unpiloted aerial vehicles (UAV) with a RGB camera would provide a high-throughput phenotyping method to evaluate shoot growth traits in carrot field trials. This method would increase measurement speed and provide a three-dimensional rendering of other shoot architecture traits, such as canopy height and canopy coverage, using common surface-from-motion algorithms. It is not yet known at what planting density high-quality cameras can resolve among individual carrot seedlings to accurately measure stand count or distinguish carrot seedlings from weeds. In our study, the few weeds in our field were visually ignored when making canopy coverage estimates. Digital phenotyping methods are under development to distinguish weeds from carrot shoots (Miao et al., 2019; Ying et al., 2021).

The CarrotOmics shoot-growth ontology definition of canopy coverage (Table 1) in the field is similar to Turner’s description of postharvest shoot biomass (Turner et al., 2018a). However, our definition of canopy coverage is not intended to supplant Turner’s methodology – both methods are appropriate for evaluating shoot biomass in different context, with the present method providing visual estimates of whole-plot shoot tissue biomass evaluated in the field, and Turner’s method estimating shoot biomass from images obtained in the lab after harvesting roots from the field. Both

estimate carrot shoot biomass using two-dimensional data, and Turner’s method is suitable for postharvest evaluation in the lab, which can be photographed in standard lighting conditions and analyzed with imaging software (Turner et al., 2018b). The method described in the present study is suitable for season-long canopy estimation in the field.

Conclusions and recommendations

We encourage carrot researchers to utilize and expand upon the descriptive terminology that we have developed. More shoot growth traits can be added to CarrotOmics shoot-growth ontology and more detailed aspects of crop growth and seedling morphology have been described and evaluated. Some traits can be studied as properties of emergence, such as emergence speed and emergence uniformity (Egli et al., 2010; Samfield et al., 1991; TeKrony and Egli, 1991). Additionally, studies in rice, wheat, and castor bean have evaluated seedling length, seedling weight, and growth rate (Hughes and Mitchel, 1987; Zhou et al., 2010; Abe et al., 2012; Guo et al., 2019). In wheat, coleoptile length has been implicated in seedling vigor (Rebetzke et al., 2007; Li et al., 2014), while in rice and pearl millet, mesocotyl elongation has been implicated in seedling vigor – an important aspect of successful stand establishment (Mohamed et al., 1989; Lee et al., 2012; Ohno et al., 2018). Genetic studies in rice and wheat have found quantitative trait loci (QTL) associated with seedling vigor (Zhang et al., 2005; Spielmeier et al., 2007; Zhou et al., 2007; Landjeva et al., 2010).

Public availability of multi-year or multi-environment phenotypic data facilitates selection of accessions with desirable agronomic traits and can be used by researchers to create core collections (Berger et al., 2013; Byrne et al., 2018). The wealth of data available on this carrot germplasm collection enables identification of germplasm across a suite of agronomically important traits, such as flowering habit (Loarca et al., 2024), ALB resistance (Tas, 2016), beta-carotenes (Ellison et al., 2018), plant growth traits (Acosta-Motos et al., 2021), taproot shape (Brainard et al., 2021), antioxidant capacity (Pérez et al., 2023), and seed germination under abiotic stress (Simon et al., 2021). While results are specific to central Wisconsin, ranked correlations were high for two out of three years of our studies. It is unknown how these ranks would shift when evaluating this germplasm in other significant carrot growing regions globally, such as in other temperate growing regions, in subtropical climates, or in the other significant growing regions in the U.S. (Washington and California). We highly recommend application of our methodology to evaluate other global carrot germplasm collections and identify ecoregional adaptation for shoot growth vigor in each target environment.

This cultivated diversity panel was curated from a *Daucus* collection to increase the relevance of our germplasm evaluation to commercial breeders. In addition to variation for shoot growth phenotypes, this collection contains annual, biennial, and mixed flowering habits and has now been characterized with the CarrotOmics flowering habit trait ontology in the companion to this article (Loarca et al., 2024), in which relationships between

carrot shoot vigor and flowering habit have also been elucidated. We have found locally adapted accessions with consistent performance over multiple years to start breeding pools for stand establishment, thereby lowering the barrier to utilization of carrot genetic resources. This list of accessions can be further optimized for efficient breeding in combination with phenological data using methods from this study's companion paper to identify ideotypes based on global market needs, such as biennial accessions for temperate breeding programs or late-flowering annual accessions for semi-arid or subtropical breeding programs. Carrot global per capita production has increased 2.7-fold in the last fifty years (Simon, 2019), making this question of multi-environment trialing of diverse germplasm relevant to all global regions of carrot production.

Data availability statement

A list of accessions used in this study is provided in [Supplementary Table 6](#). Researchers may request these accessions through the USDA Germplasm Resources Information Network (GRIN) database of the U.S. National Plant Germplasm System (NPGS). <https://npgsweb.ars-grin.gov/gringlobal/search>. The datasets for this study are hosted on the CarrotOmics database <https://www.carrotomics.org/file/409952>.

Author contributions

JL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. ML: Formal analysis, Software, Validation, Writing – review & editing. JD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing. PS: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2016-51181-25400.

Acknowledgments

The authors would like to thank Kathleen Reitsma for sharing her extensive knowledge of carrot germplasm; Lucia Gutiérrez, Irwin Goldman, and Edgar Spalding for their insightful comments on thesis chapters that led up to this manuscript; Keo

Corak, Ken Owens, David Spooner, and Bill Rolling for meaningful discussions on carrot stand establishment and crop genetic diversity; Doug Senalik for managing the CarrotOmics database; Tom Horesji, Shelby Ellison, Kevser Özel, Sarah Turner, Gunay Yildiz, Sarah Acosta, Edie Africano, Annelise Atwood, Alexis Lightner, and Hayley Stoneman for data collection and assistance in the field.

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.2024.1342512/full#supplementary-material>

SUPPLEMENTARY TABLE 1

P-values for ANOVA of seedling emergence and canopy coverage. ANOVA results indicate that genotype is a highly significant factor in all years. Broad-sense heritability (H^2) is moderately high for early-season vigor in all years studied. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 2

P-values for ANOVA of canopy Height (80 DAS) (2016–2018). ANOVA results indicate that genotype is a highly significant factor for late-season canopy height in all three years studied. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 3

P-values for ANOVA of canopy height (100 DAS) (2017 & 2018). ANOVA results indicate that genotype is a highly significant factor for late-season canopy height despite the presence of the disease ALB. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 4

P-values for multi-year ANOVA for emergence and canopy height (80 DAS) (2016–2018) and canopy height (100 DAS) (2017 & 2018). ANOVA results indicated that genotype and genotype x year interaction were highly significant factors across all years. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 5

P-values for multi-year ANOVA of emergence (20 DAS) and canopy height (80 DAS) (2016 & 2018). ANOVA results indicated that the year and block in year terms became non-significant when excluding 2017 data from the model. Genotype and genotype x year interaction were statistically significant factors across all years. Statistically significant p-values in bold.

SUPPLEMENTARY TABLE 6

List of Plant Introductions (also known as accessions or genotypes) (N = 695) used in this study.

References

- Aazami, M. A., and Zahedi, S. M. (2018). Germination of carrot (*Daucus carota* L.) seeds in response to osmotic priming. *Thai J. Agric. Sci.* 51 (4), 188–194.
- Abdulrahmani, B., Ghassemi-Golezani, K., Valizadeh, M., and Asl, V. F. (2007). Seed priming and seedling establishment of barley (*Hordeum vulgare* L.). *J. Food Agric. Environ.* 5, 179.
- Abe, A., Takagi, H., Fujibe, T., Aya, K., Kojima, M., Sakakibara, H., et al. (2012). OsGA20ox1, a candidate gene for a major QTL controlling seedling vigor in rice. *Theor. Appl. Genet.* 125, 647–657. doi: 10.1007/s00122-012-1857-z
- Acosta-Motos, J. R., Díaz-Vivancos, P., Becerra-Gutiérrez, V., Hernández Cortés, J. A., and Barba-Espin, G. (2021). Comparative characterization of eastern carrot accessions for some main agricultural traits. *Agronomy* 11, 2460. doi: 10.3390/agronomy11122460
- Alessandro, M. S., Galmarini, C. R., Iorizzo, M., and Simon, P. W. (2013). Molecular mapping of vernalization requirement and fertility restoration genes in carrot. *Theor. Appl. Genet.* 126, 415–423. doi: 10.1007/s00122-012-1989-1
- Allender, C. (2019). “Genetic resources for carrot improvement,” in *The carrot genome, compendium of plant genomes*. Eds. P. Simon, M. Iorizzo, D. Grzebelus and R. Baranski (Springer International Publishing, Cham), 93–100. doi: 10.1007/978-3-030-03389-7_6
- Arif, M., Jan, M. T., Marwat, K. B., and Khan, M. A. (2008). Seed priming improves emergence and yield of soybean. *Pakistan J. Bot.* 40, 1169–1177.
- Arnold, J. B. (2021) ggthemes: Extra Themes, Scales and Geoms for ‘ggplot2’. Rpackage version 4.2.4. Available at: <https://CRAN.R-project.org/package=ggthemes>.
- Arscott, S. A., and Tanumihardjo, S. A. (2010). Carrots of many colors provide basic nutrition and bioavailable phytochemicals acting as a functional food. *Compr. Rev. Food Saf. Food* 9, 223–239. doi: 10.1111/j.1541-4337.2009.00103.x
- Austin, R. B., and Longden, P. C. (1967). Some effects of seed size and maturity on the yield of carrot crops. *J. Hortic. Sci.* 42, 339–353. doi: 10.1080/00221589.1967.11514219
- Aziz, A., Rehman, H. U., and Khan, N. (2007). Maize cultivar response to population density and planting date for grain and biomass yield. *Sarhad J. Agric.* 23, 25.
- Bääth, R. (2018) beerp: Easily Play Notification Sounds on any Platform. Rpackage version 1.3. Available at: <https://CRAN.R-project.org/package=beerp>.
- Ballaré, C. L., Scopel, A. L., Jordan, E. T., and Vierstra, R. D. (1994). Signaling among neighboring plants and the development of size inequalities in plant populations. *Proc. Natl. Acad. Sci. U.S.A.* 91, 10094–10098. doi: 10.1073/pnas.91.21.10094
- Banga, O. (1957). Origin and distribution of the western cultivated carrot. *Euphytica* 6 (1), 54–63. doi: 10.1007/BF00179518
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bellinder, R. R., Kirkwyland, J. J., and Wallace, R. W. (1997). Carrot (*Daucus carota*) and Weed Response to Linuron and Metribuzin Applied at Different Crop Stages. *Weed Technol.* 11, 235–240. doi: 10.1017/S0890037X00042895
- Bennett, M. A., Fritz, V. A., and Callan, N. W. (1992). Impact of seed treatments on crop stand establishment. *horttech* 2, 345–349. doi: 10.121273/HORTTECH.2.3.345
- Berger, J. D., Hughes, S., Snowball, R., Redden, B., Bennett, S. J., Clements, J. C., et al. (2013). Strengthening the impact of plant genetic resources through collaborative collection, conservation, characterisation, and evaluation: a tribute to the legacy of Dr Clive Francis. *Crop Pasture Sci.* 64, 300. doi: 10.1071/CP13023
- Beveridge, J. L., and Wilsie, C. P. (1959). Influence of depth of planting, seed size, and variety on emergence and seedling vigor in alfalfa¹. *Agron. J.* 51, 731–734. doi: 10.2134/agronj1959.000219620051001200011x
- Bhasi, A., Senalik, D., Simon, P. W., Kumar, B., Manikandan, V., Philip, P., et al. (2010). RoBuST: an integrated genomics resource for the root and bulb crop families Apiaceae and Alliaceae. *BMC Plant Biol.* 10, 161. doi: 10.1186/1471-2229-10-161
- Bolton, A., Nijabat, A., Mahmood-ur-Rehman, M., Naveed, N. H., Mannan, A. M., Ali, A., et al. (2019). Variation for heat tolerance during seed germination in diverse carrot [*Daucus carota* (L.)] germplasm. *HortScience* 54, 1470–1476. doi: 10.21273/HORTSCI14144-19
- Bolton, A., and Simon, P. (2019). Variation for salinity tolerance during seed germination in diverse carrot [*Daucus carota* (L.)] germplasm. *HortScience* 54, 38–44. doi: 10.21273/HORTSCI13333-18
- Brainard, S. H., Bustamante, J. A., Dawson, J. C., Spalding, E. P., and Goldman, I. L. (2021). A digital image-based phenotyping platform for analyzing root shape attributes in carrot. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.690031
- Brocklehurst, P. A., and Dearman, J. (1980). The germination of carrot (*Daucus carota* L.) seed harvested on two dates: a physiological and biochemical study. *Journal of Experimental Botany* 31 (6), 1719–1725. doi: 10.1093/jxb/31.6.1719
- 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
- Coe, K., Bostan, H., Rolling, W., Turner-Hissong, S., Macko-Podgórn, A., Senalik, D., et al. (2023). Population genomics identifies genetic signatures of carrot domestication and improvement and uncovers the origin of high-carotenoid orange carrots. *Nat. Plants.* 9 (10), 1643–1658.
- Colquhoun, J. (2020). “Integrated weed management in carrot,” in *Carrots and related apiaceae crops*. Eds. G. Emmanuel and W. S. Philipp (CABI, Wallingford), 103–114. doi: 10.1079/9781789240955.0103
- Colquhoun, J. B., Rittmeyer, R. A., and Heider, D. J. (2017). Tolerance and suppression of weeds varies among carrot varieties. *Weed Technol.* 31, 897–902. doi: 10.1017/wet.2017.54
- Colquhoun, J. B., Rittmeyer, R. A., and Heider, D. J. (2019). Carrot weed management programs without linuron herbicide. *Weed Technol.* 33, 490–494. doi: 10.1017/wet.2018.122
- Corak, K. E., Ellison, S. L., Simon, P. W., Spooner, D. M., and Dawson, J. C. (2019). Comparison of representative and custom methods of generating core subsets of a carrot germplasm collection. *Crop Sci.* 59, 1107–1121. doi: 10.2135/cropsci2018.09.0602
- Corak, K. E., Genger, R. K., Simon, P. W., and Dawson, J. C. (2023). Comparison of genotypic and phenotypic selection of breeding parents in a carrot (*Daucus carota*) germplasm collection. *Crop Sci.* 63, 1998–2011. doi: 10.1002/csc2.20951
- De Boer, T., Smith, P., Chandler, K., Nurse, R., Obeid, K., and Swanton, C. (2019). A linuron-free weed management strategy for carrots. *Weed Technol.* 33, 464–474. doi: 10.1017/wet.2018.112
- Dowker, B. D. (1978). “Genetic variation in seedling establishment and subsequent root size variation in carrots,” in *Acta Horticulturae*, Netherlands. 72, 49–56. doi: 10.17660/ActaHortic.1978.72.5
- Drewnowski, A. (2010). The cost of US foods as related to their nutritive value. *Am. J. Clin. Nutr.* 92, 1181–1188. doi: 10.3945/ajcn.2010.29300
- Drewnowski, A. (2013). New metrics of affordable nutrition: which vegetables provide most nutrients for least cost? *J. Acad. Nutr. Dietet.* 113, 1182–1187. doi: 10.1016/j.jand.2013.03.015
- Duthie, J. A., Shreffler, J. W., Roberts, B. W., and Edelson, J. V. (1999). Plant density-dependent variation in marketable yield, fruit biomass, and marketable fraction in watermelon. *Crop Sci.* 39, 406–412. doi: 10.2135/cropsci1999.0011183X0039000200018x
- Egli, D. B., Hamman, B., and Rucker, M. (2010). Seed vigor and uniformity of seedling emergence in soybean. *Seed Technol.*, 87–95. Available at: <https://www.jstor.org/stable/23433600>.
- Ella, E. S., Dionisio-Sese, M. L., Ismail, A. M., and Box, D. (2011). Seed priming improves crop establishment of rice in flooded soils. *Int. Rice Res. Notes* 36, 6.
- Ellison, S. (2019). “Carrot domestication,” in *The carrot genome, compendium of plant genomes*. Eds. P. Simon, M. Iorizzo, D. Grzebelus and R. Baranski (Springer International Publishing, Cham), 77–91. doi: 10.1007/978-3-030-03389-7_5
- Ellison, S. L., Luby, C. H., Corak, K. E., Coe, K. M., Senalik, D., Iorizzo, M., et al. (2018). Carotenoid presence is associated with the *or* gene in domesticated carrot. *Genetics* 210, 1497–1508. doi: 10.1534/genetics.118.301299
- El-Sanatawy, A. M., El-Kholy, A. S. M., Ali, M. M. A., Awad, M. F., and Mansour, E. (2021). Maize seedling establishment, grain yield and crop water productivity response to seed priming and irrigation management in a mediterranean arid environment. *Agronomy* 11, 756. doi: 10.3390/agronomy11040756
- Eskandari, H., and Kazemi, K. (2011). Effect of seed priming on germination properties and seedling establishment of cowpea (*Vigna sinensis*). *Not. Sci. Biol.* 3, 113–116. doi: 10.15835/nsb346338
- Farooq, M., Barsa, S. M. A., and Wahid, A. (2006). Priming of field-sown rice seed enhances germination, seedling establishment, allometry and yield. *Plant Growth Regul.* 49, 285–294. doi: 10.1007/s10725-006-9138-y
- Finch-Savage, W. E., and McQuistan, C. I. (1988). Performance of carrot seeds possessing different germination rates within a seed lot. *J. Agric. Sci.* 110, 93–99. doi: 10.1017/S0021859600079739
- Finch-Savage, W. E., and Pill, W. G. (1990). Improvement of carrot crop establishment by combining seed treatments with increased seed-bed moisture availability. *J. Agric. Sci.* 115, 75–81. doi: 10.1017/S0021859600073949
- Finch-Savage, W. E., Steckel, J. R. A., and Phelps, K. (1998). Germination and post-germination growth to carrot seedling emergence: predictive threshold models and sources of variation between sowing occasions. *New Phytol.* 139, 505–516. doi: 10.1046/j.1469-8137.1998.00208.x
- Fox, J., and Weisberg, S. (2019). *An R Companion to applied regression*. 3rd ed. (Thousand Oaks CA: Sage).
- Gepts, P. (2006). Plant genetic resources conservation and utilization: the accomplishments and future of a societal insurance policy. *Crop Sci.* 46, 2278–2292. doi: 10.2135/cropsci2006.03.0169gas
- Gohel, D., and Skintzos, P. (2023) _flextable: Functions for Tabular Reporting. R package version 0.9.1. Available at: <https://CRAN.R-project.org/package=flextable>.
- Goldman, I. L. (2019). The University of Wisconsin carrot breeding and genetics program: 69 cycles of breeding for improved quality, productivity, and accessibility in processing carrot. *Acta Hortic.* 1264, 35–44. doi: 10.17660/ActaHortic.2019.1264.5

- Goss, M. (2012). A study of the initial establishment of multi - purpose moringa (*Moringa oleifera* Lam) at various plant densities, their effect on biomass accumulation and leaf yield when grown as vegetable. *Afr. J. Plant Sci.* 6 (3), 125–129. doi: 10.5897/AJPS11.259
- Grassbaugh, E. M., and Bennett, M. A. (1998). Factors affecting vegetable stand establishment. *Sci. Agric. (Piracicaba, Braz.)* 55, 116–120. doi: 10.1590/S0103-9016199800500021
- Gray, D. (1984). The performance of carrot seeds in relation to their viability. *Ann. Appl. Biol.* 104, 559–565. doi: 10.1111/j.1744-7348.1984.tb03039.x
- Gray, D., Brocklehurst, P. A., Steckel, J. R. A., and Dearman, J. (1984). Priming and pre-germination of parsnip (*Pastinaca sativa* L.) seed. *J. Hortic. Sci.* 59, 101–108. doi: 10.1080/00221589.1984.11515175
- Gray, D., and Steckel, J. R. A. (1983a). Some effects of umbel order and harvest date on carrot seed variability and seedling performance. *J. Hortic. Sci.* 58, 73–82. doi: 10.1080/00221589.1983.11515092
- Gray, D., and Steckel, J. R. A. (1983b). Seed quality in carrots: the effects of seed crop plant density, harvest date and seed grading on seed and seedling variability. *J. Hortic. Sci.* 58, 393–401. doi: 10.1080/00221589.1983.11515135
- Gray, D., and Steckel, J. R. A. (1985). Variation in flowering time as a factor influencing variability in seedling size in the subsequent carrot (*Daucus carota* L.) crop. *J. Hortic. Sci.* 60, 77–81. doi: 10.1080/14620316.1985.11515603
- Gray, D., Steckel, J. R., Jones, S. R., and Senior, D. (1986). Correlations between variability in carrot (*Daucus carota* L.) plant weight and variability in embryo length. *J. Hortic. Sci.* 61, 71–80. doi: 10.1080/14620316.1986.11515674
- GRIN-Global. (2023). Germplasm resources information network - global, USDA-ARS (National Plant Germplasm System). Available at: <https://www.grin-global.org/> (Accessed April 26, 2023).
- Guo, T., Yang, J., Li, D., Sun, K., Luo, L., Xiao, W., et al. (2019). Integrating GWAS, QTL, mapping and RNA-seq to identify candidate genes for seed vigor in rice (*Oryza sativa* L.). *Mol. Breed.* 39, 1–16. doi: 10.1007/s11032-019-0993-4
- Guragain, R. P., Baniya, H. B., Pradhan, S. P., Dhungana, S., Chhetri, G. K., Sedhai, B., et al. (2021). Impact of non-thermal plasma treatment on the seed germination and seedling development of carrot (*Daucus carota sativus* L.). *J. Phys. Commun.* 5, 125011. doi: 10.1088/2399-6528/ac4081
- Hadjichristodoulou, A., Della, A., and Photiades, J. (1977). Effect of sowing depth on plant establishment, tillering capacity and other agronomic characters of cereals. *J. Agric. Sci.* 89, 161–167. doi: 10.1017/S0021859600027337
- Hegarty, T. W. (1971). A relation between field emergence and laboratory germination in carrots. *J. Hortic. Sci.* 46, 299–305. doi: 10.1080/00221589.1971.11514410
- Hegarty, T. W. (1979). Factors influencing the emergence of calabrese and carrot seedlings in the field. *J. Hortic. Sci.* 54, 199–207. doi: 10.1080/00221589.1979.11514871
- Heydecker, W. (1956). Establishment of seedlings in the field: I. Influence of sowing depth on seedling emergence. *J. Hortic. Sci.* 31, 76–88. doi: 10.1080/00221589.1956.11513859
- Hughes, K. A., and Mitchell, W. J. P. (1987). "The relationship of coleoptile length and plant height with establishment of cereals under zero-tillage," in *Proceedings Annual Conference New Zealand Agronomy Society*, New Zealand (New Zealand: DSIR), Vol. 17, 67–70.
- Hundertmark-Bertaud, M., Boizard, S., Bregier, G., Rouby, P., Geslin, S., and Jauvion, V. (2019). Vigour tests show little evidence for predictivity of carrot stand establishment in field trials. *Acta Hortic.* 1264, 9–18. doi: 10.17660/ActaHortic.2019.1264.2
- Khan, A. A., Abawi, G. S., and Maguire, J. D. (1992). Integrating matricconditioning and fungicidal treatment of beet seed to improve stand establishment and yield. *Crop Sci.* 32, 231–237. doi: 10.2135/cropsci1992.0011183X003200010047x
- Khan, A., Wang, L., Ali, S., Tung, S. A., Hafeez, A., and Yang, G. (2017). Optimal planting density and sowing date can improve cotton yield by maintaining reproductive organ biomass and enhancing potassium uptake. *Field Crops Res.* 214, 164–174. doi: 10.1016/j.fcr.2017.09.016
- Lada, R., Stiles, A., Surette, M. A., Caldwell, C., Nowak, J., Sturz, A. V., et al. (2004). Stand establishment technologies for processing carrots. *Acta Hortic.* 105–116. doi: 10.17660/ActaHortic.2004.631.12
- Landjeva, S., Lohwasser, U., and Börner, A. (2010). Genetic mapping within the wheat D genome reveals QTL for germination, seed vigour and longevity, and early seedling growth. *Euphytica* 171, 129–143. doi: 10.1007/s10681-009-0016-3
- Lee, H.-S., Sasaki, K., Higashitani, A., Ahn, S.-N., and Sato, T. (2012). Mapping and characterization of quantitative trait loci for mesocotyl elongation in rice (*Oryza sativa* L.). *Rice* 5, 13. doi: 10.1186/1939-8433-5-13
- Lenth, R. (2023). *emmeans: estimated marginal means, aka least-squares means*. R. Available at: <https://cran.r-project.org/web/packages/emmeans/index.html>.
- Li, X. M., Chen, X. M., Xiao, Y. G., Xia, X. C., Wang, D. S., He, Z. H., et al. (2014). Identification of QTLs for seedling vigor in winter wheat. *Euphytica* 198, 199–209. doi: 10.1007/s10681-014-1092-6
- Loarca, J. S. (2021). *Identifying phenotypes and markers in diverse cultivated carrot germplasm (Daucus carota) to deliver improved stand establishment to growers* (United States: The University of Wisconsin - Madison).
- Loarca, J., Liou, M., Dawson, J. C., and Simon, P. W. (2024). Advancing utilization of carrot (*Daucus carota* L.) germplasm resources with flowering habit trait ontology. *Front. Plant Sci.* 15, 1342513. doi: 10.3389/fpls.2024.1342513
- Lonsdale, W. M. (1990). The self-thinning rule: dead or alive? *Ecology* 71, 1373–1388. doi: 10.2307/1938275
- Luby, C. H., Dawson, J. C., and Goldman, I. L. (2016). Assessment and Accessibility of Phenotypic and Genotypic Diversity of Carrot (*Daucus carota* L. var. *sativus*) Cultivars Commercially Available in the United States. *PLoS One* 11, e0167865. doi: 10.1371/journal.pone.0167865
- Mahmood-ur-Rehman, M., Amjad, M., Ziaf, K., and Ahmad, R. (2020). Seed priming with salicylic acid improve seed germination and physiological responses of carrot seeds. *Pak. J. Agric. Sci.* 57, 351–359. doi: 10.21162/PAKJAS.20.8975
- Mann, L., and MacGillivray, J. (1949). On of carrot root sizes: Studies made of spacing and seed germination to determine possible cause of size variation. *California Agric.* 3 (10), 9–13.
- Marchi, J. L., and Cicero, S. M. (2017). Use of the software Seed Vigor Imaging System (SVIS®) for assessing vigor of carrot seeds. *Sci. Agric. (Piracicaba, Braz.)* 74, 469–473. doi: 10.1590/1678-992x-2016-0220
- Maynard, D. N., Hochmuth, G. J., and Knott, J. E. (2006). *Knott's handbook for vegetable growers*. 5th ed (Hoboken, New Jersey: J. Wiley).
- Mezghani, N., Khoury, C. K., Carver, D., Achicanoy, H. A., Simon, P., Flores, F. M., et al. (2019). Distributions and conservation status of carrot wild relatives in Tunisia: A case study in the western mediterranean basin. *Crop Sci.* 59, 2317–2328. doi: 10.2135/cropsci2019.05.0333
- Miao, F., Zheng, S., and Tao, B. (2019). "Crop weed identification system based on convolutional neural network, in: 2019 IEEE 2nd international conference on electronic information and communication technology (ICEICT)," in *2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT)*, IEEE, Harbin, China. 595–598. doi: 10.1109/ICEICT.2019.8846268
- Millard, S. P. (2013). *EnvStats: an R package for environmental statistics*. (New York: Springer). Available at: <https://www.springer.com>, doi: 10.17660/ActaHortic.2002.588.10
- Mohamed, A., Barnett, F. L., Vanderlip, R. L., and Khaleeq, B. (1989). Emergence and stand establishment of pearl millet as affected by mesocotyl elongation and other seed and seedling traits. *Field Crops Res.* 20, 41–49. doi: 10.1016/0378-4290(89)90022-1
- Mou, F. I., Hossain, M. M., HaqueB, T., and Yasmin, A. (2023). Impact of tillage depth and planting spacing on growth and root yield of carrot (*Daucus carota* L.). *J. Trop. Crop Sci.* 10 (3), 186–195. doi: 10.29244/jtcs.10.03.186-195
- Muhie, S., Memiş, N., Özdamar, C., Gökdaş, Z., and Demir, İ. (2021). Biostimulant priming for germination and seedling quality of carrot seeds under drought, salt and high temperature stress conditions. *Int. J. Agric. Environ. Food Sci.* 5, 352–359. doi: 10.31015/jaefs.2021.3.13
- Muhie, S. H., Akele, F., and Yeshiwas, T. (2024). Economic feasibility of carrot (*Daucus carota* L.) production in response to different seed priming techniques under deficit irrigation. *Scientia Hortic.* 325, 112662. doi: 10.1016/j.scienta.2023.112662
- Nijabat, A., Manzoor, S., Faiz, S., Naveed, N. H., Bolton, A., Khan, B. A., et al. (2023). Variation in Seed Germination and Amylase Activity of Diverse Carrot [*Daucus carota* (L.)] Germplasm under Simulated Drought Stress. *HortScience* 58, 205–214. doi: 10.21273/HORTSCI16806-22
- Ohno, H., Banayo, N. P. M. C., Bueno, C. S., Kashiwagi, J., Nakashima, T., Corales, A. M., et al. (2018). Longer mesocotyl contributes to quick seedling establishment, improved root anchorage, and early vigor of deep-sown rice. *Field Crops Res.* 228, 84–92. doi: 10.1016/j.fcr.2018.08.015
- Peil, R. M., and López-Gálvez, J. (2002). Fruit growth and biomass allocation to the fruits in cucumber: effect of plant density and arrangement. *Acta Hortic.* 588, 75–80. doi: 10.17660/ActaHortic.2002.588.10
- Pérez, M. B., Carvajal, S., Beretta, V., Bannoud, F., Fangio, M. F., Berli, F., et al. (2023). Characterization of purple carrot germplasm for antioxidant capacity and root concentration of anthocyanins, phenolics, and carotenoids. *Plants* 12, 1796. doi: 10.3390/plants12091796
- Posit team (2023). RStudio: integrated development environment for R (Boston, MA: Posit Software, PBC). Available at: <http://www.posit.co/>.
- Postma, J. A., Hecht, V. L., Hikosaka, K., Nord, E. A., Pons, T. L., and Poorter, H. (2021). Dividing the pie: A quantitative review on plant density responses. *Plant Cell Environ.* 44, 1072–1094. doi: 10.1111/pce.13968
- Prohens, J., and Nuez, F. (2008). "Vegetables II: fabaceae, liliaceae, solanaceae, and umbelliferae," in *Handbook of plant breeding* (Springer, New York). doi: 10.1007/978-0-387-74110-9
- R Core Team. (2023). *R: A language and environment for statistical computing*. (Vienna, Austria: R Foundation for Statistical Computing). Available at: <https://www.R-project.org/>, doi: 10.1534/g3.117.300235
- Rebetzke, G. J., Richards, R. A., Fittell, N. A., Long, M., Condon, A. G., Forrester, R. I., et al. (2007). Genotypic increases in coleoptile length improves stand establishment, vigour and grain yield of deep-sown wheat. *Field Crops Res.* 100, 10–23. doi: 10.1016/j.fcr.2006.05.001
- Rolling, W. R., Senalik, D., Iorizzo, M., Ellison, S., Van Deynze, A., and Simon, P. W. (2022). CarrotOmics: a genetics and comparative genomics database for carrot (*Daucus carota*). *Database* 2022, baac079. doi: 10.1093/database/baac079
- Rubatzky, V. E., Quiros, C. F., and Simon, P. W. (1999). "Carrots and related vegetable Umbelliferae," in *Carrots and related vegetable Umbelliferae*. CABI publishing.

- Salter, P. J., Currah, I. E., and Fellows, J. R. (1981). Studies on some sources of variation in carrot root weight. *J. Agric. Sci.* 96, 549–556. doi: 10.1017/S002185960003450X
- Salter, P. J., and Darby, R. J. (1976). Synchronization of germination of celery seeds. *Ann. Appl. Biol.* 84, 415–424. doi: 10.1111/j.1744-7348.1976.tb01784.x
- Samfield, D. M., Zajicek, J. M., and Cobb, B. G. (1991). Rate and uniformity of herbaceous perennial seed germination and emergence as affected by priming. *J. Am. Soc. Hortic. Sci.* 116, 10–13. doi: 10.21273/JASHS.116.1.10
- Sanders, D. C., Ricotta, J. A., and Hodges, L. (1990). Improvement of carrot stands with plant biostimulants and fluid drilling. *HortSci* 25, 181–183. doi: 10.21273/HORTSCI.25.2.181
- Seale, D. N., and Cantliffe, D. J. (1986). “Improved stand establishment and yield of sand land grown lettuce by seed treatment and soil amendments,” In *Proceedings of the Florida State Horticultural Society*, vol. 99, 365–369.
- Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., et al. (2010). Multifunctional crop trait ontology for breeders’ data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants* 2010, plq008. doi: 10.1093/aobpla/plq008
- Simon, P. W. (2019). Beyond the genome: carrot production trends, research advances, and future crop improvement. *Acta Hortic.* 1264, 1–8. doi: 10.17660/ActaHortic.2019.1264.1
- Simon, P. W., Navazio, J. P., Colley, M., McCluskey, C., Zystro, J., Hoagland, L., et al. (2017). The CIOA (Carrot Improvement for Organic Agriculture) project: location, cropping system and genetic background influence carrot performance including top height and flavour. *Acta Hortic.* 1153, 1–8. doi: 10.17660/ActaHortic.2017.1153.1
- Simon, P. W., Rolling, W. R., Senalik, D., Bolton, A. L., Rahim, M. A., Mannan, A. M., et al. (2021). Wild carrot diversity for new sources of abiotic stress tolerance to strengthen vegetable breeding in Bangladesh and Pakistan. *Crop Sci.* 61 (1), 163–176. doi: 10.1002/csc2.20333
- Sowmeya, T. V., Macha, S. I., Vasudevan, S. N., Shakuntala, N. M., and Ramesh, G. (2018). Influence of priming on seed quality of fresh and old seed lots of carrot (*Daucus carota* L.). *J. Pharmacog. Phytochem.* 7 (1), 1114–1117.
- Spielmeier, W., Hyles, J., Joaquim, P., Azanza, F., Bonnett, D., Ellis, M. E., et al. (2007). A QTL on chromosome 6A in bread wheat (*Triticum aestivum*) is associated with longer coleoptiles, greater seedling vigour and final plant height. *Theor. Appl. Genet.* 115, 59–66. doi: 10.1007/s00122-007-0540-2
- Spooner, D. M., Widrechner, M. P., Reitsma, K. R., Palmquist, D. E., Rouz, S., Ghrabi-Gammar, Z., et al. (2014). Reassessment of practical subspecies identifications of the USDA *daucus carota* L. Germplasm collection: morphological data. *Crop Sci.* 54, 706–718. doi: 10.2135/cropsci2013.04.0231
- Steiner, F., and Zuffo, A. M. (2019). Drought tolerance of four vegetable crops during germination and initial seedling growth. *Biosci. J.* 35 (1), 177–186. doi: 10.14393/BJ-v32n1a2016
- Suchánková, M., Kapounová, Z., Dofková, M., Ruprich, J., Blahová, J., and Kouřilová, I. (2015). Selected fruits and vegetables: comparison of nutritional value and affordability. *Czech J. Food Sci.* 33, 242–246. doi: 10.17221/353/2014-CJFS
- Swanton, C. J., O’Sullivan, J., and Robinson, D. E. (2010). The critical weed-free period in carrot. *weeds* 58, 229–233. doi: 10.1614/WS-09-098.1
- Szafirowska, A., Khan, A. A., and Peck, N. H. (1981). Osmoconditioning of carrot seeds to improve seedling establishment and yield in cold soil ¹. *Agron. J.* 73, 845–848. doi: 10.2134/agronj1981.00021962007300050023x
- Tas, P. M. (2016). *Evaluating resistance to Alternaria deuce and related traits among diverse germplasm of Daucus carota* Ph.D. thesis (United States: University of Wisconsin).
- TeKrony, D. M., and Egli, D. B. (1991). Relationship of seed vigor to crop yield: A review. *Crop Sci.* 31, 816–822. doi: 10.2135/cropsci1991.0011183X003100030054x
- Turner, S. D., Ellison, S. L., Senalik, D. A., Simon, P. W., Spalding, E. P., and Miller, N. D. (2018a). An automated image analysis pipeline enables genetic studies of shoot and root morphology in carrot (*Daucus carota* L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01703
- Turner, S. D., Maurizio, P. L., Valdar, W., Yandell, B. S., and Simon, P. W. (2018b). Dissecting the genetic architecture of shoot growth in carrot (*Daucus carota* L.) using a diallel mating design. *G3 Genes[Genomes]Genet.* 8, 411–426. doi: 10.1534/g3.117.300235
- USDA Economic Research Service (ERS). (2023). Organic production. Available at: <https://www.ers.usda.gov/data-products/organic-production/documentation/>.
- Van Heemst, H. D. J. (1985). The influence of weed competition on crop yield. *Agric. Syst.* 18, 81–93. doi: 10.1016/0308-521X(85)90047-2
- Vavilov, N. I. (1951). Centres of origin, variation, immunity and breeding of cultivated plants. *Chronica Botanica* 13, 1, 366.
- Vega, A., and Goldman, I. (2023). Planting density does not affect root shape traits associated with market class in carrot. *horts* 58, 996–1004. doi: 10.21273/HORTSCI17232-23
- Walls, R. L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M. A., Jaiswal, P., et al. (2012). Ontologies as integrative tools for plant science. *Am. J. Bot.* 99, 1263–1275. doi: 10.3732/ajb.1200222
- Westoby, M. (1984). The self-thinning rule. In *Advances in ecological research*. (Academic press) 14, 167–225. doi: 10.1016/S0065-2504(08)60171-3
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Soft.* 4*, 1686. doi: 10.21105/joss.01686
- Ying, B., Xu, Y., Zhang, S., Shi, Y., and Liu, L. (2021). Weed detection in images of carrot fields based on improved YOLO v4. *TS* 38, 341–348. doi: 10.18280/ts.380211
- Zhang, Z.-H., Yu, S.-B., Yu, T., Huang, Z., and Zhu, Y.-G. (2005). Mapping quantitative trait loci (QTLs) for seedling-vigor using recombinant inbred lines of rice (*Oryza sativa* L.). *Field Crops Res.* 91 (2-3), 161–170. doi: 10.1016/j.fcr.2004.06.004
- Zhou, G., Ma, B. L., Li, J., Feng, C., Lu, J., and Qin, P. (2010). Determining salinity threshold level for castor bean emergence and stand establishment. *Crop Sci.* 50, 2030–2036. doi: 10.2135/cropsci2009.09.0535
- Zhou, L., Wang, J. K., Yi, Q., Wang, Y. Z., Zhu, Y. G., and Zhang, Z. H. (2007). Quantitative trait loci for seedling vigor in rice under field conditions. *Field Crops Res.* 100, 294–301. doi: 10.1016/j.fcr.2006.08.003



OPEN ACCESS

EDITED BY

Xiyang Zhao,
Jilin Agricultural University, China

REVIEWED BY

Kundapura Ravishankar,
Indian Institute of Horticultural Research
(ICAR), India
Yuepeng Song,
Beijing Forestry University, China
Xiang Li,
Ningxia University, China

*CORRESPONDENCE

Zhangqi Yang
✉ yangzhangqi@163.com

RECEIVED 12 January 2024

ACCEPTED 02 April 2024

PUBLISHED 24 April 2024

CITATION

An Q, Feng Y, Yang Z, Hu L, Wu D and
Gong G (2024) Differences in *Albizia*
odoratissima genetic diversity between Hainan
Island and mainland populations in China.
Front. Plant Sci. 15:1369409.
doi: 10.3389/fpls.2024.1369409

COPYRIGHT

© 2024 An, Feng, Yang, Hu, Wu and Gong.
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.

Differences in *Albizia odoratissima* genetic diversity between Hainan Island and mainland populations in China

Qi An^{1,2}, Yuanheng Feng^{1,2,3}, Zhangqi Yang^{1,2,3*}, La Hu^{1,2},
Dongshan Wu^{1,2} and Guifang Gong^{1,2}

¹Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation, Guangxi Zhuang Autonomous Region Forestry Research Institute, Nanning, Guangxi, China, ²Masson Pine Engineering Technology Research Center of Guangxi, Nanning, Guangxi, China, ³Key Laboratory of Central South Fast-growing Timber Cultivation of Forestry Ministry of China, Nanning, Guangxi, China

Background: This study aimed at exploring unique population genetic characteristics of *Albizia odoratissima* (Linn. f) Benth on Hainan Island to provide a scientific basis for its rational utilization and protection.

Methods: It analyzed the genetic diversity and structure of 280 individuals from 10 subpopulations of *A. odoratissima* from Hainan Island and Baise City using 16 expression sequence markers - simple sequence repeat markers.

Results: The genetic diversity of Hainan population ($I = 0.7290$, $He = 0.4483$) was lower than that of the Baise population ($I = 0.8722$, $He = 0.5121$). Compared with the Baise population ($Nm = 2.0709$, $F_{ST} = 0.1077$), the Hainan Island population ($Nm = 1.7519$, $F_{ST} = 0.1249$) exhibited lower gene flow and higher degree of genetic differentiation. Molecular variance and genetic differentiation analyses showed that the main variation originated from individuals within the subpopulation. There were significant differences in the genetic structure between Hainan and Baise populations. It grouped according to geographical distance, consistent with the Mantel test results ($R^2 = 0.77$, $p = 0.001$). In conclusion, the genetic diversity of the island *A. odoratissima* population was lower than that distributed on land, the two populations exhibited obvious genetic structure differences. Both the degrees of inbreeding and genetic differentiation were higher in the island population than in the land population.

KEYWORDS

Albizia odoratissima, EST-SSR molecular markers, genetic diversity, genetic structure, Hainan Island

Introduction

Albizia odoratissima (Linn. f) Benth is a perennial evergreen tree of *Albizia Durazz* that belongs to Leguminosae sp., Mimosaceae, commonly found in low-altitude sparse forests (Yuan et al., 2011; Ai, 2016). The plant has no spines, the twigs are initially puberulous; leaflet oblong, apex blunt; the flowers are sessile, yellowish, fragrant, capitulum arranged terminal, open panicles; pod flat, oblong, with 6–12 seeds; flowering period from April to July, fruit period from June to October. This tree species has excellent quality, rapid growth, medicinal value, is a rare fast-growing precious timber tree species (Liang, 2012; Liang, 2020; Wei et al., 2020). However, the *A. odoratissima* distribution area is shrinking due to human interference, among other reasons, and the threat of endangerment is becoming increasingly serious (Luo et al., 2020). Therefore, population genetics research is needed to effectively protect this precious species.

As a typical monsoonal evergreen broad-leaved forest species, *A. odoratissima* is mainly distributed in Guangxi, Yunnan, Guizhou, Guangdong, Sichuan, Hainan, and other provinces in China, it had also been reported in India, Vietnam, Malaysia, and other countries (Wei et al., 2020). In China, the natural population of *A. odoratissima* shows an “island-like” discontinuous distribution, spanning the South subtropical and tropical climate zones. Its distribution area covers a variety of landforms, such as the Yunnan-Guizhou Plateau, Nanling Mountain, the Hilly area of South China, and Hainan Island Mountain. This distribution pattern is highly likely to hinder the gene exchange between *A. odoratissima* populations in different regions, resulting in rich genetic variation between them.

Hainan Island is the second-largest island in China. It was completely separated from the Asian mainland owing to the formation of the Qiongzhou Strait in the middle Pleistocene. Since then, it has faced the southernmost end of mainland China across the sea. With an area of 35400 km² (Wang et al., 2022), the island has a tropical monsoon climate, rich rainfall, and sufficient heat. Known as a “species gene pool” and “natural museum,” it is also the most special distribution area of *A. odoratissima*. The Hainan population is the only *A. odoratissima* group distributed in the tropical areas of China. As a typical “continental island,” Hainan Island is an ideal place to study species genetics and evolution (Del Valle et al., 2020). For species with intermittent island and land distributions, the limited island area, long-term geographical isolation, and greatly different climatic conditions lead to obvious genetic differentiation and phenotypic differences between the populations on the island and those on the mainland during evolution. From the formation of unique or even excellent germplasm resources, research different populations of the same species distributed on both islands and the mainland can provide important clues for the intraspecific evolution that determines the early differentiation and recent spread of species (Fernández-Mazuecos and Vargas, 2011).

In general, in order to research the genetic differences between island populations and mainland populations, it is more appropriate to compare the populations in the land distribution area closest to the island. In addition, comparative studies can also be conducted using

the core distribution area or origin center area populations on the mainland. The current survey results showed that there is no large-scale distribution of *A. odoratissima* populations in the Leizhou Peninsula which is closest to Hainan Island and the coastal areas of southern China. The Baise population, located in the northwest of Guangxi, China, is adjacent to the two distribution areas of eastern Yunnan and southwest Guizhou, is the closest to Hainan Island among the large-scale distribution areas of *A. odoratissima* found so far. The geographical distribution area of Baise is 36,000 km² (Liu et al., 2017), which is similar to Hainan Island. From the perspective of biogeography, Hainan Island was once connected to Guangxi, China (Zhu, 2016). Therefore, selecting the population of Baise area as the control experimental material has good representativeness.

The research of genetic diversity is of great significance for revealing the evolutionary history of species, evaluating their survival status, and predicting their future development trends. At present, research on population genetics of species mostly relies on experimental materials from island or terrestrial populations, while there are relatively few populations from both islands and continents. As a species of important research value distributed on both islands and land, the study of population genetics of *A. odoratissima* was still limited. Simple sequence repeat (SSR) molecular marker has the advantages of abundant quantity, codominant inheritance, high polymorphism, high specificity, high universality and good repeatability (Garrido-Cardenas et al., 2018). According to their origin, they can be divided into genomic SSR (G-SSR) and expression sequence label SSR (EST-SSR). Compared with G-SSR, EST-SSR has the advantages of simple development process, low cost, high sequence conservation, and inter-species transferability (Manoj et al., 2021). The EST-SSR originates from transcription regions, which is often linked to functional genes, it has obvious advantages in population genetics research and marker-assisted breeding. In view of this, it used EST-SSR molecular markers as a tool to comprehensively evaluate the genetic diversity and structure of *A. odoratissima* populations in four regions of Hainan Island. Baise population as a representative of terrestrial population. To provide reference for the genetic evaluation of this germplasm resource and the development of high-quality genetic resources.

Materials and methods

Experimental materials of *A. odoratissima* population

Based on relevant literature, the natural population of *A. odoratissima* is mainly distributed in the west and southwest Hainan Province (Figure 1). Yingge Ling, Sanya City, Jianfengling, and Bawangling of Hainan Island were selected as the locations for test materials collection, representing the east, south, west, and north of the main distribution area of *A. odoratissima*. The population of Baise Prefecture in Guangxi were collected from six regions, including Youjiang District, Xilin County, Tianlin County, Tiandong County, Longlin County, and Leye County. At each collection site, according to the standard

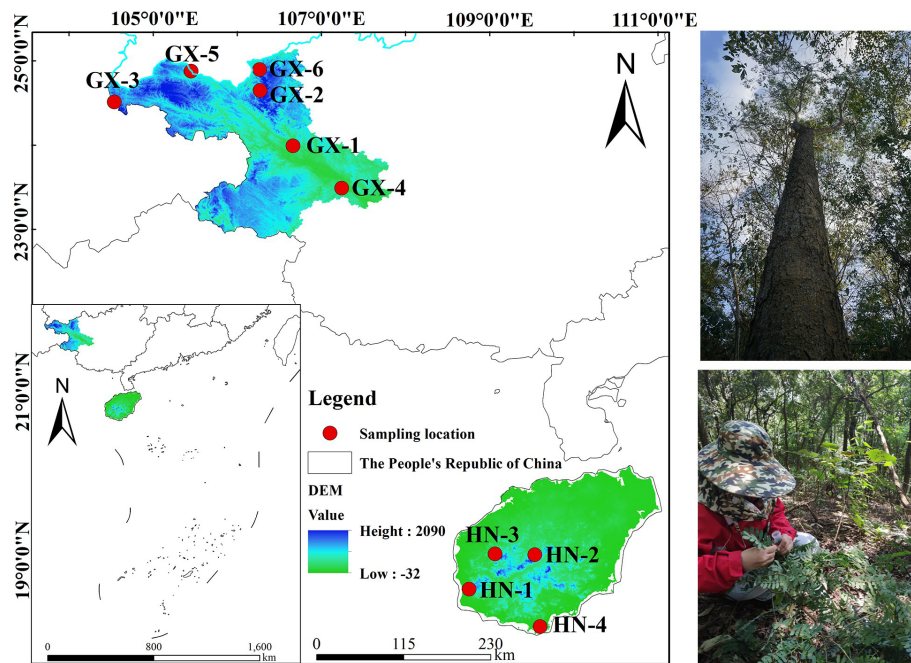


FIGURE 1
Geographical distribution of the plant materials.

distance between plants of no less than 30 m, individual plants with normal growth, no obvious defects, diseases, and pests were selected. The top tender tree leaves from the current year were collected, silica gel was used to dry and preserve them. The collection sites for the experimental materials were evenly distributed, showing good representation. A total of 280 samples were collected from each of the ten sites (Table 1).

Extraction of genomic DNA from *A. odoratissima*

Fresh plant tissues (50 – 100 mg) were ground into a powder in liquid nitrogen, and genomic DNA was extracted using an EZ – 10 Spin Column Plant Genomic DNA Purification Kit, which Providing by Sangon Biotech (Shanghai) Co., Ltd.

SSR (simple sequence repeat)-PCR (polymerase chain reaction) amplification and detection of amplification products

Transcriptome sequencing was performed on *A. odoratissima* leaves. Considering the RNA-seq results, Novofinder software was used to search for SSR loci in the transcriptome Universal Gene of *A. odoratissima*. The SSR loci were screened following these criteria: sequence length: 18–26 bp and number of nucleotide repeats ≥ 5 . 16 pairs of EST-SSR primers were selected for this study (Table 2). All primers were synthesized by Guangzhou Ige Biotechnology LTD, Guangzhou, China.

The PCR system was 10.00 μ L, including template DNA (1.00 μ L; 50 ng/stock), 10 \times buffer (1.00 μ L; 1 \times Buffer), dNTPs (including Mg^{2+}) (0.20 μ L; 0.2 mmol/L), positive and negative primers (0.25 μ L; 0.25 μ mol/L), Taq DNA enzyme (0.10 μ L 0.25 U/stock), and ddH₂O (7.20 μ L). The PCR procedure was as follows: 94°C pre-denaturation for 4 min; 25 cycles of 94°C denaturation for 15 s, 58°C annealing for 15 s, 72°C extension for 30 s; extension at 72°C for 20 min; and storage at 12°C. The amplified products were separated using 8% polyacrylamide gel electrophoresis.

Data analysis

POPGENE 32 (Yeh et al., 1999) software was used to calculate the following genetic parameters: average number of alleles (A), observed number of alleles (N_a) (Fang et al., 2018), effective number of alleles (N_e) (Hartl and Clark, 1989), Shannon's Information Index (I), and Nei's gene diversity index (H) (Shannon and Weaver, 1949; Shannon et al., 1950), observed heterozygosity (H_o), expected heterozygosity (H_e) (Nei, 1973), and Wright's fixation index (F) could be obtained from the formula $1-F_{IT} = (1-F_{IS})(1-F_{ST})$, used to determine F_{IS} , F_{IT} , and genetic differentiation index (F_{ST}) (Wright, 1978). Nei's standard genetic distance (GD) (Nei, 1972), genetic identity, gene flow (N_m) (Slatkin and Barton, 1989), and genetic differentiation coefficient (G_{ST}).

STRUCTURE 2.3.4 (Pritchard et al., 2000) software was used to perform Bayesian clustering analysis on individuals from different populations. Subpopulation grouping was determined based on the optimal K value obtained via STRUCTURE Harvester (<https://taylor0.biology.ucla.edu/structureHarvester>) analysis.

TABLE 1 Basic information of natural *A. odoratissima* populations.

Sample col- lection site	Subpopulation names	Subpopulation number	Longitude and latitude	Altitude(m)	Sample size	Climate
Hainan population	Jianfengling	HN-1	108°50'E	96~534	17	Tropical monsoon climate, dry rainy season obvious
			18°42'N			
	Yinggeling	HN-2	109°36'E	271~536	16	Tropical maritime monsoon climate, long summer without winter
			18°59'N			
	Bawangling	HN-3	109°08'E	126~600	25	Tropical rainforest climate, mild cli-mate, abundant rainfall
			19°15'N			
	Sanya	HN-4	109°40'E	45~260	20	Tropical maritime monsoon climate
			18°16'N			
Baise population	Chengbi Lake	GX-1	106°39'E	170~330	31	Subtropical monsoon climate
			24°00'N			
	Tianlin county	GX-2	106°10'E	350~450	30	Subtropical monsoon climate, long summer, and short winter
			24°21'N			
	Xilin county	GX-3	104°34'E	750~1100	30	Subtropical continental monsoon climate, no cold winter, and summer heat
			24°28'N			
	Tiandong county	GX-4	107°11'E	330~520	31	South Tropical monsoon climate, hot summer, mild winter, wet summer, and dry winter
			23°43'N			
	Longlin county	GX-5	105°28'E	520~580	30	Subtropical alpine climate, more obvious four seasons, no hot summer, and no cold winter
			24°50'N			
	Leye county	GX-6	106°17'E	670~710	50	Subtropical monsoon climate, summer without heat, and winter without cold
			24°51'N			

The polymorphism information content (*PIC*) for each locus was calculated using the online program *PIC* calc (Botstein et al., 1980).

Arlequin 3.5 (Excoffier and Lischer, 2010) software was used to analyze the molecular variation at different levels within and between populations.

Mantel test was performed using GenAlEx 6.5 (Peakall and Smouse, 2012). The Mantel test was used to assess the correlation between geographic distance and genetic distance.

The UPGMA (unweighted pair-group method with arithmetic means) clustering tree was constructed using NTSYS PC (Rohlf, 2000) software based on Nei’s standard genetic distance.

The ape (Paradis and Schliep, 2019) and ggplot 2 (Wickham, 2016) packages of R 2.3.4 were analyzed for principal coordinates analysis (PCoA).

Results

Screening of EST-SSR primers for *A. odoratissima*

The results of the previous study showed that among the 243 pairs of developed primers of *A. odoratissima*, the effective amplification rate in *Albizia odoratissima* *Albizia procera*, *Albizia falcataria*, *Acacia melanoxylon* and *Erythrophloeum fordii* was 63.79%, 33.75%, 45.68%, 41.56% and 14.81% respectively, the polymorphism ratio in them was 23.87%, 12.20%, 9.01%, 3.96% and 2.78% respectively (An et al., 2022). A total of 16 materials were randomly selected from various subpopulations of *A. odoratissima* in Guangxi and Hainan, the polymorphisms of 155 pairs of effective primers were re-detected. Finally, 16 pairs of primers with good generality, high polymorphism and clear bands were selected for the genetic diversity study of the germplasm resources of *A. odoratissima*. The polymorphism information content was analyzed, as shown in Table 2. The effective amplification rate of 16 selected primers in *Albizia procera*, *Albizia falcataria* and *Acacia melanoxylon* was 50%, and the proportion of polymorphic primers in effective primers was 50%, 12.5% and 12.50%, respectively. Only one pair of primers can have effective amplification rate and no polymorphism in *Erythrophloeum fordii*. It can be seen that the polymorphic information content of the 16 SSR primers used in this

experiment ranged from 0.08 to 0.70, with an average value of 0.48. The 16 pairs of primers selected were all medium-high polymorphic primers, except for AO-194, which was a low polymorphic primer. Indicating that the screened primers could be used for the genetic diversity study of the germplasm resources of *A. odoratissima*. It had certain universality in *Albizia procera*, *Albizia falcataria*, *Acacia melanoxylon* and *Erythrophloeum fordii*.

Genetic diversity of *A. odoratissima* population

A total of 52 alleles were amplified by 16 SSR loci in 280 individuals from 10 *A. odoratissima* subpopulations (Table 3), the number of alleles detected at each site ranged from 2.00 to 5.00.

In terms of gene abundance (Table 3), 43 and 52 alleles were detected at 16 SSR loci in Hainan and Baise populations, respectively. Compared with the Baise population, the Hainan population had allele deletions at 50.00% of loci and no specific alleles (Table 4). The average number of observed alleles (*Na*) in Hainan population was 2.69, which was 17.23% lower than that in Baise population. The average number of effective alleles (*Ne*) in Hainan population was 1.96, which was 12.50% lower than that in the Baise population. To some extent, it showed that after geographical isolation from the mainland, Hainan population were greatly limited in introducing new alleles through gene communication.

The average *He*, *Ho*, *I*, and *H* values of the Hainan *A. odoratissima* population were 0.4483, 0.2075, 0.7290, and 0.4452, respectively (Table 3). which were 87.54%, 74.10%, 83.58%, and 87.16% lower than those of Baise population, respectively. Thus, genetic complexity was lower in the Hainan population than in the

TABLE 2 EST-SSR primer information of *A. odoratissima*.

Primer	Repeat the primitive	Primer (5' ~ 3')	Tm/°C	PIC
AO-37	(GAA)8	F: ACGATGGAACAGTAACCGGA	59.93	0.48
		R: GTGCTGTTTGGATCCTCCAT		
AO-53	(TAT)8	F: AGGAGGAGGAGGCGTTGTAT	59.95	0.70
		R: TTCAGCTCAGCCCTGATTTT		
AO-62	(TTCG)6	F: TGCCTCACACTACACGCTTC	60.02	0.50
		R: GCGTTGCTTGAGGACTAAGG		
AO-75	(GAG)8	F: ATGCATGAGGAATGGAGGAG	60.03	0.53
		R: CCTCTCCTTATGCCTTTCCC		
AO-130	(ATT)7	F: AGCTCTAAAAGCAGGTGGCA	59.83	0.40
		R: GCCTGTGTGCATCATCGCTTA		
AO-133	(GCC)7	F: AGGATTAAGCAAAGCGCTGA	59.96	0.43
		R: CGGAGTTGGCAGTGATATT		
AO-141	(AGA)7	F: AGGAAGTGTCCAACGGGTG	59.98	0.46
		R: GGCGTCTTCGCTATTCAAAG		
AO-146	(CAG)7	F: ATCTGAGATGGCTTGTTGGG	59.98	0.54

(Continued)

TABLE 2 Continued

Primer	Repeat the primitive	Primer (5' ~ 3')	Tm/°C	PIC
		R: TTTGCTGCATATCTCGTTGC		
AO-166	(ATG)7	F: TTCGTGGAATCGATCAATCA	59.93	0.47
		R: TGGCTCCAACATCCCTTAAC		
AO-184	(TCTT)5	F: TGGGGGAACAGTGGTTATGT	59.98	0.50
		R: TCTCTGTTTCGTCATTCTGTCG		
AO-188	(TTAA)5	F: GCTCCCAATATCCATGTGCT	59.92	0.61
		R: TGAAGGATATCACCGCATCA		
AO-189	(TTAA)5	F: ATGCAGGTTGCAATCAATCAA	59.98	0.60
		R: TTTGGGAATTGGGGATTACCA		
AO-194	(AAGA)5	F: CTTCACCGGATCTAGGACCA	59.97	0.08
		R: ATTTCGGAACGAACCAGTTG		
AO-199	(TACA)5	F: TCATCAATGTGCTTCCCAAA	59.89	0.57
		R: AGCTCAAGCAGCTCAGGAAC		
AO-210	(GAAA)5	F: GTTTCATGGTGATATGGGC	60.07	0.28
		R: ATGTCCCAGAGAATGCCAAG		
AO-217	(GCAG)5	F: TCTCCCATCAAAATCCAAGC	60.00	0.58
		R: CTGGAGAATCCCATCGAAA		

Baise population. In the four subpopulations of Hainan, *Ho* ranged from 0.1869 to 0.2246, *He* ranged from 0.3820 to 0.4150, *I* ranged from 0.6088 to 0.6537, and *H* ranged from 0.3706 to 0.4021. In the six subpopulations of Baise, *Ho* ranged from 0.2542 to 0.3737, *He* ranged from 0.4436 to 0.4855, *I* ranged from 0.7261 to 0.7927, and *H* ranged from 0.4357 to 0.4765. The *He*, *Ho*, *I* and *H* values of the

four Hainan subgroups were all lower than those of the six subgroups of Baise.

There were certain differences in allele frequencies between the two populations at different loci (Table 4), with a maximum of 67.84%, indicating that the two populations have different evolutionary directions due to differences in natural environment.

TABLE 3 Genetic diversity in 10 populations of *A. odoratissima*.

population	<i>N</i>	<i>Na</i>	<i>Ne</i>	<i>Ho</i>	<i>He</i>	<i>I</i>
HN-1	41	2.56	1.82	0.2246	0.4150	0.6537
HN-2	38	2.38	1.84	0.2126	0.4132	0.6434
HN-3	38	2.38	1.85	0.2095	0.4092	0.6452
HN-4	39	2.44	1.7	0.1869	0.3820	0.6088
Population level	43	2.69	1.96	0.2075	0.4483	0.7290
GX-1	46	2.88	1.96	0.2542	0.4473	0.7474
GX-2	47	2.94	2.01	0.2786	0.4476	0.7378
GX-3	47	2.94	2.04	0.3737	0.4855	0.7927
GX-4	48	3.00	2.06	0.2555	0.4600	0.7830
GX-5	44	2.75	1.98	0.2744	0.4436	0.7261
GX-6	46	2.88	2.12	0.2591	0.4715	0.7872
Population level	52	3.25	2.24	0.2800	0.5121	0.8722
Species level	52	3.25	2.36	0.2615	0.5366	0.9079

N, Total number of alleles; *Na*, Observed number of alleles; *Ne*, Effective number of alleles; *Ho*, Observed Heterozygosity; *He*, Expected Heterozygosity; *I*, Shannon's Information Index; *H*, Nei's gene diversity index.

TABLE 4 Allele frequencies of *A. odoratissima* populations from Hainan and Baise.

Allele	Hainan population					Baise population				
	A	B	C	D	E	A	B	C	D	E
AO-39	34.87%	65.13%				29.95%	53.81%	16.24%		
AO-53	0.67%	2.00%	74.00%	23.33%		52.01%	14.07%	6.28%	16.08%	11.56%
AO-62	28.57%	51.95%	19.48%			30.25%	61.50%	8.25%		
AO-75	22.08%	49.35%	28.57%			10.05%	68.81%	21.13%		
AO-130	22.67%	63.33%	14.00%			11.36%	81.31%	7.32%		
AO-133		24.67%	75.33%			2.99%	78.61%	18.41%		
AO-141		98.08%	1.92%			26.41%	51.54%	22.05%		
AO-146	38.67%	26.67%	34.67%			19.49%	27.95%	52.56%		
AO-166	4.86%	36.11%	59.03%			1.02%	35.03%	63.96%		
AO-184	14.67%	85.33%				45.75%	41.75%	11.75%	0.75%	
AO-188	11.46%		10.42%	78.13%		42.94%	6.18%	40.59%	10.29%	
AO-189	7.29%		35.42%	57.29%		45.86%	2.37%	30.77%	21.01%	
AO-194	94.23%	5.77%				99.75%	0.25%			
AO-199	30.92%	15.13%	53.95%			40.86%	20.05%	38.83%	0.25%	
AO-210	49.36%	50.64%				25.00%	75.00%			
AO-217	59.46%	17.57%	22.97%			33.60%	43.55%	22.85%		

Population genetic differentiation and genetic variation

The N_m (0.9976) between the two populations of *A. odoratissima* was lower than that of Hainan ($N_m=1.7559$) and

Baise ($N_m=2.0709$), respectively. This indicates that gene exchange between the two populations was low, but the degree of gene exchange within the population was relatively high.

The H_o values of both populations were lower than the H_e value, the inbreeding coefficients of each subgroup were greater

TABLE 5 Analysis of molecular variance (AMOVA) for population of *A. odoratissima*.

Population	Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Fixation Indices
Hainan population	Among populations	3	33.255	0.20478 Va	8.14	$F_{IS}=0.4671$
	Among individuals	74	235.649	0.87427 Vb	34.76	$F_{ST}=0.1249$
	Within populations					
	Within individuals	78	112	1.43590 Vc	57.09	$F_{IT}=0.5336$
Baise population	Total	155	380.904	2.51495	100	$G_{ST}=0.1000$
	Among populations	5	134.244	0.35007 Va	11.47	$F_{IS}=0.3768$
	Among individuals	196	686.053	0.79840 Vb	26.16	$F_{ST}=0.1077$
	Within populations					
Total population	Within individuals	202	384.5	1.90347 Vc	62.37	$F_{IT}=0.4439$
	Total	403	1204.797	3.05194	100	$G_{ST}=0.0980$
	Among populations	9	305.738	0.55071 Va	17.26	$F_{IS}=0.4130$
	Among individuals	270	946.005	0.86347 Vb	27.06	$F_{ST}=0.2004$
	Within populations					
	Within individuals	280	497.500	1.77679 Vc	55.68	$F_{IT}=0.5306$
	Total	559	1749.243	3.19096	100	$G_{ST}=0.1838$

Df, Degree of freedom.

than zero, indicating that both populations had a certain degree of homozygosity and inbreeding. Among them, the inbreeding coefficient of the GX-3 subgroup was close to zero, showing that it was closest to the Hardy Weinberg equilibrium.

From the F_{ST} and G_{ST} values of the *A. odoratissima* population in Hainan Province were 0.1249 and 0.1000, respectively, while the population in Baise City were 0.1077 and 0.0980, respectively. Therefore, the genetic differentiation of the Hainan population was higher than that of the Baise population. The genetic variation of the two populations mainly comes from subpopulations.

Further analysis of variance was performed at the population, subpopulation and individual levels (Table 5). Both populations exhibit within individuals > among individuals Within populations> among populations. It indicated that the genetic variation of *A. odoratissima* mainly came from the variation between individuals.

Genetic structure of *A. odoratissima* population

The genetic structure of *A. odoratissima* was predicted using STRUCTURE software. When $K = 2$, the ΔK value was the highest (Figures 2A, B), indicating that it was most reasonable to divide 280

samples into two groups of independent evolutionary units (Figure 2C). Based on the Q value, it also plotted the cluster member proportion of each subpopulation when $K = 2$. The four Hainan subpopulations were composed of individuals with red clusters, and the six Baise subpopulations were composed of individuals with green clusters. The results were consistent with those of UPGMA cluster analysis (Figure 3A) and PCoA (Figure 3B) based on GD . This shows obvious differences in genetic structure between the Hainan and Baise populations of *A. odoratissima*. It further analyzed the correlation between GD and geographic distance using Mantel test (Figure 4). GD was significantly correlated with geographical distance ($R^2 = 0.77$, $p = 0.001$), indicating an obvious geographical origin structure or main distance isolation among the investigated populations.

Discussion

Genetic diversity of the *A. odoratissima* population

The genetic diversity of trees directly affects their evolutionary potential and adaptability to environmental changes. Higher genetic diversity indicates stronger evolutionary potential and ability to match the environment (Xia, 1999; Booy et al., 2000). The genetic

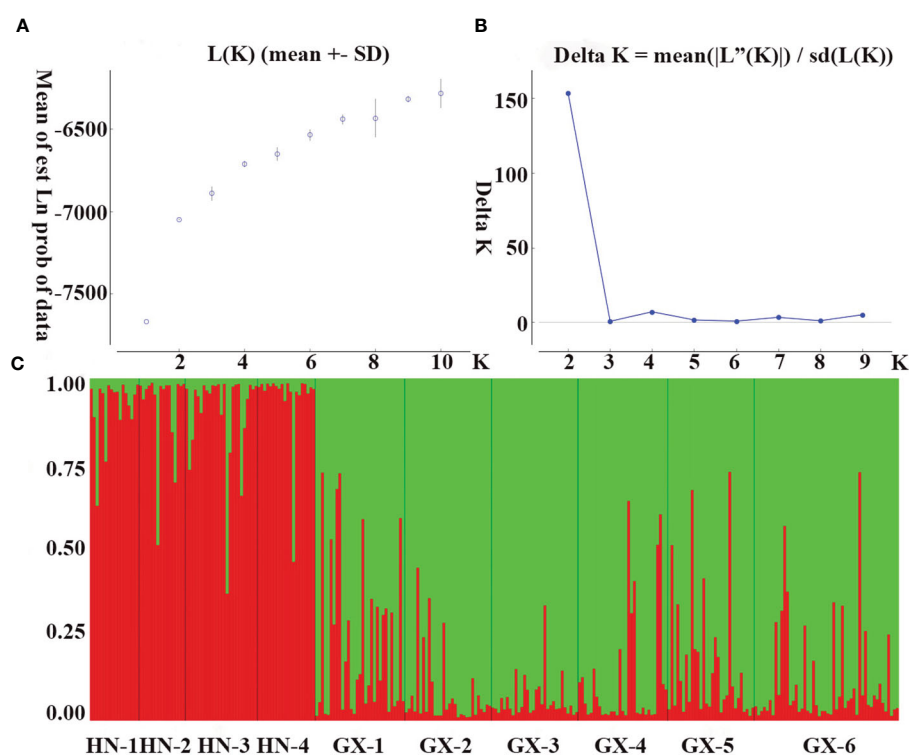


FIGURE 2

Population genetic structure. (A) Mean log-likelihood $[Ln(K) \pm SD]$ against the number of K ; (B) Relations between the rational groups number K and Estimated value ΔK . (C) Genetic structural plot of 10 *A. odoratissima* subpopulations based on structure analysis (Each individual is represented by a single vertical bar, which is partitioned into two different colors).

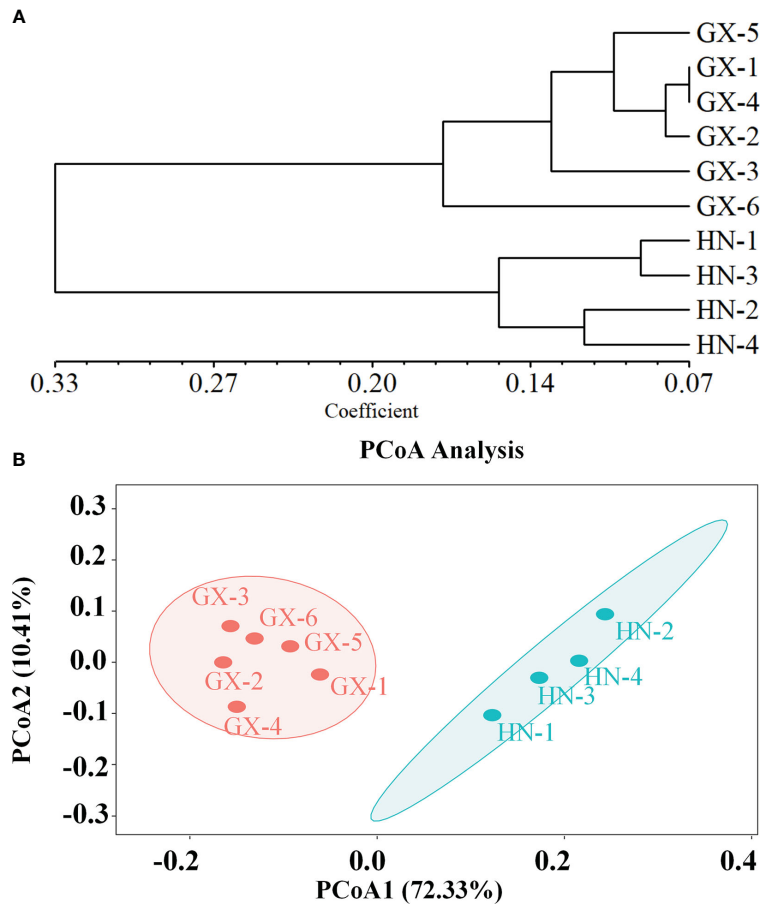


FIGURE 3 (A) Genetic divergence among 10 subpopulations of *A. odoratissima* based on UPGMA clustering analysis. (B) Principal coordinate analysis (PCoA) of 10 *A. odoratissima* subpopulations.

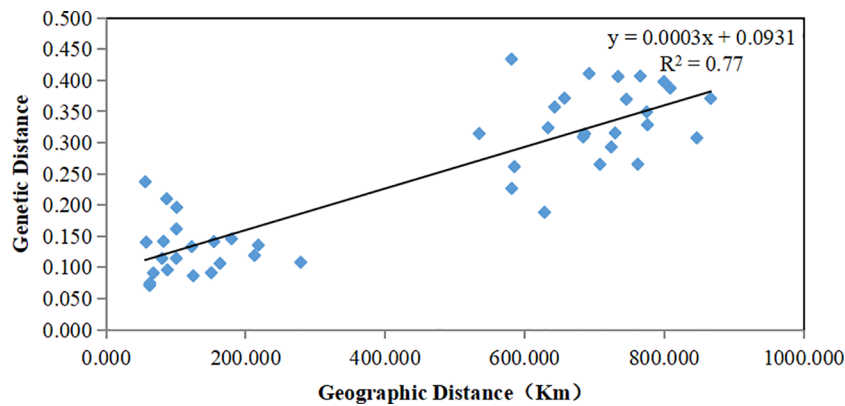


FIGURE 4 Mantel between genetic distance and geographical distance of 10 *A. odoratissima* subpopulations.

diversity of a species is closely related to its geographical distribution, mating system, gene flow, genetic drift, and human interference (Jaenike, 1973; Fan et al., 2019; Jiang et al., 2019).

This study revealed that the genetic abundance and complexity of the *A. odoratissima* population in Hainan were lower than those of the Baise population. This phenomenon that the genetic diversity of island population was lower than that of land population also exists in some species such as *Mussaenda kwangtungensis* (Guangdong) (Hufford et al., 2014; Shi et al., 2020; Hunt et al., 2022). Long-term geographic isolation is probably the main reason explaining the lower genetic diversity of the Hainan population. This phenomenon is related to factors, such as Nm blocking, genetic drift, and small population bottleneck effect allele loss (Avice, 1996; Rieseberg and Swensen, 1996; Balmford et al., 1997).

Compared to the Baise population, allele deletion occurred in more than half of the loci in the Hainan population. It found that the two groups had lower levels of genetic communication. Indicating that the existence of geographical isolation makes it difficult for Hainan population to acquire new genes through gene exchange from land population. In addition, the population size on Hainan Island was small, which may complicate gene mutation fixing in this population. The geographical distribution area of Baise population is close to Hainan, located at the junction of Guangxi, Guizhou, and Yunnan provinces in China, adjacent to Vietnam to the south. The population size of Baise was large, there were fewer gene exchange barriers between nearby populations, which helps to accept new genes through gene exchange and fix mutated genes. It may also be the main factor why a unique haplotype was not found in the Hainan population. It is also the reason that genetic diversity of Hainan population is lower than that of Baise population.

The genetic diversity of species dominated by outbreeding is generally higher than that of species dominated by self-breeding, the inter-population variation of the former accounts for more than 50% of the total variation (Gao and Li, 2008; Lu et al., 2021). The genetic diversity of the total population ($H_o = 0.2615$, $H_e = 0.5366$) was higher than that of wild soybean ($H_o = 0.0310$, $H_e = 0.4260$) (He et al., 2012), as well as the average of the major self-crossing species ($H_o = 0.1200$, $H_e = 0.3030$) (Hamrick and Godt, 1996). In addition, 18.38% of the genetic variation existed among the subpopulations, it can be reasonably inferred that the breeding system was dominated by out-crossing. The blossoms of *A. odoratissima* are light yellow, fragrant, terminal, and dispersed panicles, conforming to the characteristics of insect pollination. Affected by the flight distance of insects, the ability of the wild population to spread pollen among subpopulations was limited by distance, terrain, and other factors (Kwon and Morden, 2002). The Hainan population is more about gene exchange among subpopulations on the island, the richness of genetic diversity is lower than that of the Baise population located at the border of the three provinces and in the core distribution area.

Genetic differentiation and genetic structure of the *A. odoratissima* population

The genetic structure of a population is the result of the interaction between ecological and genetic processes. Related

research is of great significance for understanding population genetic characteristics and dynamics, and developing effective protection measures (Leberg, 1990; Cheng et al., 2020). There were significant differences in genetic structure between the Hainan and Baise populations, the genetic variation mainly came from within the population. The coefficient of genetic differentiation between the two populations ($F_{ST} = 0.2004$) was significantly higher than that of Hainan and Baise populations ($F_{ST} = 0.1249$ and $F_{ST} = 0.1077$). In the cluster analysis, six Baise subpopulations were grouped into one cluster, and four Hainan subpopulations were grouped into another cluster. Hainan had different geographical, topographic, water, thermal conditions from Baise. Under the effect of selection, the Hainan population had taken a different evolutionary direction from that of the Baise population in order to accommodate the local environment. The existence of geographical isolation resulted in genetic differentiation between Hainan and mainland populations, while the outcrossing breeding system and the characteristics of insect pollination further intensified the genetic differentiation between island and land populations. In the wild, the *A. odoratissima* are mainly propagated by seeds, when the seeds mature, the pods crack. Affected by seed weight, wind speed, and media, the seeds generally fall not far from the mother plants, can also spread to further places along rivers. Nevertheless, the long-term geographical isolation makes it difficult for the Hainan population to rely on water flow or some media to break through the geographical isolation barrier and smoothly communicate genes with the land population. The degree of pollen dispersal among long-distance subpopulations was also lower, therefore, the Hainan population is more about gene exchange among subpopulations on the island. The isolation of the sea resulted in a lower degree of gene exchange between island and land populations than within the population, making the genetic differentiation between island and land populations more pronounced (Wright, 1951; Grant, 1986; Manel, 2003; Saro et al., 2015).

In addition, the degree of inbreeding and genetic differentiation were higher in Hainan than in Baise, consistent with the conclusion of Chen et al (Chen et al., 2008). on *Ficus pumila*. The small size and scattered distribution of the natural *A. odoratissima* population may be underlying reasons for inbreeding in this population. The resources of *A. odoratissima* were abundant in Hainan, their distribution range was limited, and the population is smaller than that in Baise. There are many tall mountains on the island, creating a mountain isolation effect greater than natural transportation effect (Frankham, 1998). It prevents gene exchange among subpopulations and facilitating inbreeding between individuals. The existence of inbreeding reduces population heterozygosity and further increases genetic differentiation of the populations.

Conservation and utilization of *A. odoratissima* genetic resources

Some alleles were found to exist in only a few subpopulations, there was abundant genetic variation among individuals within the populations. However, due to human interference, the number of its

natural population is decreasing dramatically. In view of this, it can further expand the select range of superior trees, increase their number, and establish seed orchards with a wider genetic basis than the original. Establishing seed orchards with a wider genetic basis can better preserve *A. odoratissima* resources. Seeds with better genetic quality can also be produced to establish high-quality *A. odoratissima* plantations to meet production needs and increase the breeding value of high-quality genetic resources. In addition, the study found that different subpopulations had different levels of genetic diversity, there were rare alleles with allele frequencies less than 0.01 in the *A. odoratissima* population, indicating that the rare alleles carrying certain genetic variants in the population were likely to exist in only one or a few individuals, these genetic variants were facing loss. Therefore, it is necessary to select some subpopulations with high genetic diversity and establish nature reserves to protect them in situ.

Conclusions

This article used EST-SSR molecular markers as a research tool to comprehensively evaluate the genetic diversity and structure of two *A. odoratissima* populations from Hainan Island and Baise, Guangxi, China. The genetic variation of the population mainly comes from individual to individual. The *A. odoratissima* population distributed in Hainan Island has lower genetic diversity than the Baise population, the degree of genetic differentiation. The Hainan Island *A. odoratissima* population had a 50% loss of alleles compared to the Baise population, and there are no specific alleles. There were significant differences in the genetic structure of the two populations, with different evolutionary directions. These findings have improved our current understanding of the genetic diversity and population genetic structure of *A. odoratissima*. It can provide a theoretical basis for the improvement of seed orchard, further development, utilization, and protection of this species.

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

QA: Conceptualization, Data curation, Formal analysis, Investigation, Validation, Writing – original draft, Writing – review

& editing. YF: Conceptualization, Methodology, Validation, Writing – review & editing. ZY: Conceptualization, Data curation, Funding acquisition, Project administration, Validation, Writing – review & editing. LH: Investigation, Validation, Writing – review & editing. DW: Investigation, Validation, Writing – review & editing. GG: Formal analysis, Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Guangxi Science and Technology Base and Special Funds for Talents (AD19254004), Special Program of Bagui Scholar (2019A026) and National “Thirteenth Five-Year” Key R&D Project (2017YFD0600303).

Acknowledgments

We appreciate senior engineer Jie Jia and engineer Qunfeng Luo for their seedling cultivation assistance, and appreciate Bailin Forest Farm in Baise, Guangxi, China for collection of germplasm resources assistance.

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.2024.1369409/full#supplementary-material>

References

- Ai, T. M. (2016). *Annals of medicinal plants of China* Vol. 5 (Beijing: Peking University Medical Press), 143.
- An, Q., Feng, Y. H., Yang, Z. Q., and Hu, L. (2022). EST-SSR marker development and interspecific generality of *Albizia odoratissima*. *Guihaia* 42 (8), 1374–1382.
- Avise, J. C. (1996). Conservation genetics: case histories from nature. *J. Appl. Ecol.* 78.
- Balmford, A., Avise, J. C., and Hamrick, J. L. (1997). Conservation genetics: case histories from nature. *J. Appl. Ecol.* 34 (3), 829. doi: 10.2307/2404927
- Booy, G., Hendriks, R. J. J., Smulders, M. J. M., van Groenendaal, J. M., and Vosman, B. (2000). Genetic diversity and the survival of populations. *Plant Biol.* 2, 379–395. doi: 10.1055/s-2000-5958
- Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.
- Chen, Y., Shi, M., Ai, B., Gu, J., and Chen, X. (2008). Genetic variation in island and mainland populations of *Ficus pumila* (Moraceae) in eastern Zhejiang of China. *Symbiosis* 45, 37–44.
- Cheng, J., Kao, H., and Dong, S. (2020). Population genetic structure and gene flow of rare and endangered *Tetraena mongolica* Maxim. revealed by reduced representation sequencing. *BMC Plant Biol.* 20, 391. doi: 10.1186/s12870-020-02594-y
- Del Valle, J. C., Herman, J. A., and Whittall, J. B. (2020). Genome skimming and microsatellite analysis reveal contrasting patterns of genetic diversity in a rare sandhill endemic (*Erysimum teretifolium*, Brassicaceae). *PLoS One* 15, e0227523. doi: 10.1371/journal.pone.0227523
- Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x
- Fan, J. J., Zhang, X. P., Liu, K., Liu, H. J., Zhang, L., Wang, X. P., et al. (2019). The population genetic diversity and pattern of *Pteroceltis tatarinowii*, a relic tree endemic to China, inferred from SSR markers. *Nordic J. Bot.* 37 (2). doi: 10.1111/njb.01922
- Fang, S. S., Xie, X. F., Qi, J. M., Zhang, L. M., Xu, J. T., Lin, L. H., et al. (2018). Universality of simple sequence repeat (SSR) markers from Cotton (*Gossypium hirsutum*) to Kenaf (*Hibiscus cannabinus*). *Chin. J. Trop. Crops* 39, 1373–1382. doi: 10.3969/j.issn.1000-2561.2018.07.017
- Fernández-Mazuecos, M., and Vargas, P. (2011). Genetically Depauperate in the Continent but Rich in Oceanic Islands: *Cistus monspeliensis* (Cistaceae) in the Canary Islands. *PLoS One* 6, e17172. doi: 10.1371/journal.pone.0017172
- Frankham, R. (1998). Inbreeding and extinction: Island population. *Conserv. Biol.* 12, 665–675. doi: 10.1046/j.1523-1739.1998.96456.x
- Gao, J., and Li, Q. M. (2008). Genetic diversity of natural populations of *Acacia pennata* in Xishuang-banna, Yunnan. *Biodiversity Sci.* 16, 271–278. doi: 10.3321/j.issn:1005-0094.2008.03.009
- Garrido-Cardenas, J. A., Mesa-Valle, C., and Manzano-Agugliaro, F. (2018). Trends in plant research using molecular markers. *Planta* 247, 543–557. doi: 10.1007/s00425-017-2829-y
- Grant, V. (1986). The evolutionary process: a critical study of evolutionary theory. *Stud. Hist. Phil. Sci.* 17, 65–98.
- Hamrick, J. L., and Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philos. Trans. R. Soc. Lond.* 351, 1291–1298. doi: 10.1098/rstb.1996.0112
- Hartl, D. L., and Clark, A. G. (1989). *Principles of population genetics*. 2nd ed (Sunderland, MA: Sinauer Associates).
- He, S., Wang, Y., Volis, S., Li, D., and Yi, T. (2012). Genetic diversity and population structure: implications for conservation of wild soybean (*Glycine soja* Sieb. et Zucc.) based on nuclear and chloroplast microsatellite variation. *Int. J. Mol. Sci.* 13, 12608–12628. doi: 10.3390/ijms131012608
- Hufford, K. M., Mazer, S. J., and Hodges, S. A. (2014). Genetic variation among mainland and island populations of a native perennial grass used in restoration. *AoB Plants* 6, plt055. doi: 10.1093/aobpla/plt055
- Hunt, D. A. G. A., DiBattista, J. D., and Hendry, A. P. (2022). Effects of insularity on genetic diversity within and among natural populations. *Ecol. Evol.* 12, e8887. doi: 10.1002/ece3.8887
- Jainike, J. R. (1973). A steady state model of genetic polymorphism on islands. *Am. Nat.* 107, 793–795. doi: 10.1086/282878
- Jiang, H., Long, W., Zhang, H., Mi, C., Zhou, T., and Chen, Z. (2019). Genetic diversity and genetic structure of *Decalobanthus boissianus* in Hainan Island, China. *Ecol. Evol.* 9, 5362–5371. doi: 10.1002/ece3.5127
- Kwon, J. A., and Morden, C. W. (2002). Population genetic structure of two rare tree species (*Colubrina oppositifolia* and *Alphitonia ponderosa*, Rhamnaceae) from Hawaiian dry and mesic forests using random amplified polymorphic DNA markers. *Mol. Ecol.* 11, 991–1001. doi: 10.1046/j.1365-294X.2002.01497.x
- Leberg, P. L. (1990). Genetic considerations in the design of introduction programs transactions of the north American wildlife. *Natural Resource Conf.* 55, 609–619.
- Liang, S. H. (2020). Analysis of technical measures for building *Albizia odoratissima* forest in Yachang forest area. *Agric. Technol.* 40, 87–88. doi: 10.19754/j.nyys.20201030028
- Liang, S. Y. (2012). *Sylva guangxigensis. Volume I* (Beijing: China Forestry Publishing House), 382.
- Liu, X., Liu, Z., Zhang, Y., and Jiang, B. (2017). The effects of floods on the incidence of bacillary dysentery in baize (Guangxi province, China) from 2004 to 2012. *Int. J. Environ. Res. Public Health* 14, 179. doi: 10.3390/ijerph14020179
- Lu, J., Zhang, Y., Diao, X., Yu, K., Dai, X., Qu, P., et al. (2021). Evaluation of genetic diversity and population structure of *Fragaria nilgerrensis* using EST-SSR markers. *Gene* 796–797, 145791. doi: 10.1016/j.gene.2021.145791
- Luo, Q. F., Hu, L., Tan, J. H., Jia, J., Feng, Y. H., and Yang, Z. Q. (2020). Phenotypic diversity of seeds of *Albizia odoratissima* from Baize district. *J. For. Environ.* 40, 62–67. doi: 10.13324/j.cnki.jfcf.2020.01.009
- Manel, S. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* 18, 189–197. doi: 10.1016/S0169-5347(03)00008-9
- Manoj, K. G., Ravindra, D., Sabarinathan, S., Gayatri, G., Goutam, K. D., Pallabi, P., et al. (2021). Microsatellite markers from whole genome and transcriptomic sequences. *Bioinf. Rice Res.*, 387–412. doi: 10.1007/978-981-16-3993-7_18
- Nei, M. (1972). Genetic distance between populations. *Am. Nat.* 106, 283–292. doi: 10.1086/282771
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321
- Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633
- Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460
- 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
- Rieseberg, L. H., and Swensen, S. M. (1996). Conservation genetics of endangered island plants. *Conservation Genetics Case Histories from Nature*. doi: 10.1007/978-1-4757-2504-9
- Rohlf, F. J. (2000). NTSYS-pc numerical taxonomy and multivariate analysis system solutions manual. *Am. Stat.* 41, 330.
- Saro, I., González-Pérez, M. A., García-Verdugo, C., and Sosa, P. A. (2015). Patterns of genetic diversity in *Phoenix canariensis*, a widespread oceanic palm (species) endemic from the Canarian archipelago. *Tree Genet. Genomes* 11, 1–13. doi: 10.1007/s11295-014-0815-0
- Shannon, C. E., and Weaver, W. (1949). The mathematical theory of communication. *Philos. Rev.* 93 (3), 31–32. doi: 10.1063/1.3067010
- Shannon, C. E., Weaver, W., and Wiener, N. (1950). The mathematical theory of Communication. *Phys. Today*. 9 (3), 31–32. doi: 10.1063/1.3067010
- Shi, M., Wang, Y., Duan, T., Qian, X., Zeng, T., and Zhang, D. (2020). *In situ* glacial survival maintains high genetic diversity of *Mussaenda kwangtungensis* on continental islands in subtropical China. *Ecol. Evol.* 10, 11304–11321. doi: 10.1002/ece3.6768
- Slatkin, M., and Barton, N. H. (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* 43, 1349–1368. doi: 10.2307/2409452
- Wang, M., Zhang, X., Zhang, Y., and Xiao, M. (2022). Prevalence and genetic analysis of thalassemia and hemoglobinopathy in different ethnic groups and regions in hainan island, Southeast China. *Front. Genet.* 13. doi: 10.3389/fgene.2022.874624
- Wei, S. X., Liang, R. L., Lin, J. Y., He, Y. H., and Jiang, Y. (2020). Geographical distribution and community characteristics of *Albizia odoratissima* in China. *Guangxi For. Sci.* 49, 71–75. doi: 10.19692/j.cnki.gfs.2020.01.014
- Wickham, H. (2016). *ggplot2: elegant graphics for data analysis* (New York: Springer-Verlag New York). doi: 10.1007/978-3-319-24277-4
- Wright, S. (1951). The genetic structure of populations. *Ann. Eugen.* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x
- Wright, S. (1978). *Variability within and among natural populations* (Chicago: The Univ. Of Chicago Press).
- Xia, M. (1999). Research progress of genetic diversity. *Chin. J. Ecol.* (3), 59–65. Available at: <http://www.cje.net.cn/EN/Y1999/V13/I5/59>.
- Yeh, F. C., Yang, R. C., and Boyle, T. (1999). *POPGENE version1.32, microsoft window-bass software for population genetic analysis: a quick user's guide university of alberta, center for international forestry research* (Canada: Alberta).
- Yuan, T. X., Huang, Y. Q., and Liang, R. L. (2011). *Main native tree species in Guangxi* (Nanning: Guangxi Science and Technology Press), 149.
- Zhu, H. (2016). Biogeographical evidences help revealing the origin of Hainan Island. *PLoS One* 11, e0151941. doi: 10.1371/journal.pone.0151941



OPEN ACCESS

EDITED BY

Petr Smýkal,
Palacký University in Olomouc, Czechia

REVIEWED BY

Enoch G. Achigan-Dako,
University of Abomey-Calavi, Benin
Oldřich Trněný,
Agricultural Research Ltd., Czechia

*CORRESPONDENCE

Monica Carvajal-Yepes

✉ m.carvajal@cgiar.org

Miguel Correa Abondano

✉ m.correa@cgiar.org

[†]These authors have contributed equally to this work

RECEIVED 14 November 2023

ACCEPTED 19 June 2024

PUBLISHED 11 July 2024

CITATION

Correa Abondano M, Ospina JA, Wenzl P and Carvajal-Yepes M (2024) Sampling strategies for genotyping common bean (*Phaseolus vulgaris* L.) Genebank accessions with DArTseq: a comparison of single plants, multiple plants, and DNA pools. *Front. Plant Sci.* 15:1338332. doi: 10.3389/fpls.2024.1338332

COPYRIGHT

© 2024 Correa Abondano, Ospina, Wenzl and Carvajal-Yepes. 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.

Sampling strategies for genotyping common bean (*Phaseolus vulgaris* L.) Genebank accessions with DArTseq: a comparison of single plants, multiple plants, and DNA pools

Miguel Correa Abondano^{*†}, Jessica Alejandra Ospina[†], Peter Wenzl and Monica Carvajal-Yepes^{*}

Genetic Resources Program, International Center for Tropical Agriculture (CIAT), Palmira, Colombia

Introduction: Genotyping large-scale gene bank collections requires an appropriate sampling strategy to represent the diversity within and between accessions.

Methods: A panel of 44 common bean (*Phaseolus vulgaris* L.) landraces from the Alliance Bioversity and The Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT) gene bank was genotyped with DArTseq using three sampling strategies: a single plant per accession, 25 individual plants per accession jointly analyzed after genotyping (*in silico-pool*), and by pooling tissue from 25 individual plants per accession (*seq-pool*). Sampling strategies were compared to assess the technical aspects of the samples, the marker information content, and the genetic composition of the panel.

Results: The *seq-pool* strategy resulted in more consistent DNA libraries for quality and call rate, although with fewer polymorphic markers (6,142 single-nucleotide polymorphisms) than the *in silico-pool* (14,074) or the single plant sets (6,555). Estimates of allele frequencies by *seq-pool* and *in silico-pool* genotyping were consistent, but the results suggest that the difference between pools depends on population heterogeneity. Principal coordinate analysis, hierarchical clustering, and the estimation of admixture coefficients derived from a single plant, *in silico-pool*, and *seq-pool* successfully identified the well-known structure of Andean and Mesoamerican gene pools of *P. vulgaris* across all datasets.

Conclusion: In conclusion, *seq-pool* proved to be a viable approach for characterizing common bean germplasm compared to genotyping individual plants separately by balancing genotyping effort and costs. This study provides insights and serves as a valuable guide for gene bank researchers embarking on genotyping initiatives to characterize their collections. It aids curators in effectively managing the collections and facilitates marker-trait association studies, enabling the identification of candidate markers for key traits.

KEYWORDS

genotyping, sampling, genetic resources, common bean, DArTseq

Introduction

Germplasm banks are repositories of crop genetic diversity. These collections include landraces, cultivars, wild forms, and closely related species. Not only do they serve a conservation purpose, but these plants and seeds are also a vital source of novel and underused genetic variation, an important input for national and private plant breeding programs to tackle the challenges faced by the agricultural sector (Byrne et al., 2018; Swarup et al., 2021). However, in the lengthy process of introducing novel genetic variation into a program, the first step requires field trials to identify candidates to start testing crosses with elite cultivars. This increases the cost of characterizing gene bank collections for complex traits like tolerance to abiotic stresses, considering that collections may number in the tens of thousands of accessions. To address this, multiple tools have been developed to improve the characterization of germplasm collections such as using passport and climate data to identify candidate accessions for abiotic stress tolerance (Smith et al., 1994; Greene et al., 1999; Cortés et al., 2013; Khoury et al., 2015; Haupt and Schmid, 2020).

As DNA sequencing and genotyping has become increasingly prevalent, they have been used to characterize germplasm collections of cultivated species worldwide. Examples include cowpea [*Vigna unguiculata* (L.) Walp.; Wamalwa et al., 2016], rice (*Oryza sativa* L.; Wang et al., 2018), forages [*Elymus tangutorum* (Nevski) Hand.-Mazz; Wu et al., 2019], cassava (*Manihot esculenta* Crantz; Adjebo-Danquah et al., 2020), and common bean (Martins et al., 2006; Ariani et al., 2018; Nadeem et al., 2018). Emerging techniques have been developed, involving the use of one or more restriction enzymes to fragment genomic DNA, that enable the selection of specific genomic representations for subsequent sequencing and marker identification (Sansaloni et al., 2011). These advances significantly reduce the cost associated with genotyping numerous accessions. Nevertheless, genotyping thousands of plants still requires significant resources.

However, there is more to consider in a large-scale genotyping effort than just the sequencing strategy. A prime example is the seed bank of *Phaseolus* species conserved at the Genetic Resources Program of the Bioversity-CIAT Alliance (“the Alliance” or “ABC” hereafter). This remarkable collection encompasses approximately 38,000 plant materials, comprising all five cultivated species within the genus: the common bean (*P. vulgaris* L.), lima bean (*P. lunatus* L.), runner bean (*P. coccineus* L.), tepary bean (*P. acutifolius* A. Gray), and year bean (*P. dumosus* Macfady), along with approximately 40 wild species. The conventional practice of selecting a single random plant per accession for genotyping may not adequately represent the entire population (Gouda et al., 2020). This limitation arises because *Phaseolus* species exhibit a wide spectrum of mating behaviors, ranging from strictly allogamous to fully autogamous (Bitocchi et al., 2017). Moreover, there exists substantial variation within species themselves (Ibarra-Perez et al., 1997; Ferreira et al., 2000; Royer et al., 2002).

Genotyping more than 20–30 plants per population to obtain accurate allele frequencies and other population diversity estimates results in a significant increase (up to 30-fold) in genotyping costs, without accounting for additional space, labor, and time

requirements. As a result, alternative sampling schemes are imperative for genotyping large collections. Pooling DNA has emerged as a promising alternative to individual sampling [for a review, see the work of Schlötterer et al. (2014)]. This approach involves the collection of equal volumes of plant tissue into a single tube, followed by a single DNA extraction for subsequent sequencing. Previous research has been conducted to explore the genetic diversity of various species using pooled data (Farahani et al., 2019; Ketema et al., 2020; Dziurdziak et al., 2021; Gapare et al., 2021; Arca et al., 2023). Recent comparative studies have investigated individual sampling with bulks of different sizes in rice (*Oryza* spp.) using DArTseq (Gouda et al., 2020), comparing whole-genome individual and pool sequencing of honey bee (*Apis mellifera* L.) (Chen et al., 2022) and studying the population structure of the American lobster with either GBS, rapture, or whole-genome pool-seq (Dorant et al., 2019). Despite research exploring the genetic diversity of species using pool data, little work has been done on the viability of pooling DNA from the common bean.

This study addressed this gap by using a diversity panel comprised of 44 accessions of the common bean (*P. vulgaris*) to compare two distinct sampling methods: individual sequencing or pooled sequencing. Our aim is to determine whether pooling DNA represents a viable alternative for studying the genetic diversity of the common bean gene bank collection. To achieve this, we evaluate how individual and pooled sequencing compare in terms of the number of markers identified through DArTseq, estimates of allele frequencies and heterozygosity, and the exploration of population structure of accessions of the species. This investigation contributes valuable insights into optimizing genotyping strategies for large-scale germplasm collections.

Materials and methods

Plant material and sample pooling

A total of 44 cultivated accessions of *Phaseolus vulgaris* L. were included in this study: 43 landraces and one modern cultivar (G4489; Supplementary Table 1). These accessions were selected from various continents including Africa, the Americas, Asia, and Europe. They were selected from the bean germplasm collection of the Alliance for the purpose of comparing the impact of pooling samples on allele frequency estimates. Thirty seeds from each accession were sown in the greenhouse at 25°C and 60% relative humidity at the ABC campus in Palmira-Colombia. Young leaf tissue was collected 15 days after sowing from each individual plant using a leaf tissue punch to obtain standard-size leaf discs. Tissue leaf discs were stored individually or pooled together in a single tube, to create the pool for each accession. All samples were stored at –80°C until DNA extraction. A total of 1,140 samples, including 1,096 individual samples and 44 pooled samples, were collected. The samples were intended to compare two types of pools: *seq-pools*, consisting of the 22 to 25 tissue leaf discs from individual plants collected in one tube for DNA extraction and sequenced as single samples per accession, and the *in silico-pools*, which

comprise 22 to 25 individual plants each in single tubes for DNA extraction and sequenced independently. Subsequently, samples were analyzed together as *in silico*-pooled samples.

DNA extraction, sequencing, and genotyping

Genomic DNA was extracted from around 10 mg of lyophilized leaf tissue from 2-week-old seedlings according to a modified Cetrimonium bromide (CTAB) protocol (Dellaporta et al., 1983; Doyle and Doyle, 1990). Extracted DNA was resuspended in 100 µL of TE buffer and incubated with 2 U of Ribonuclease (RNase) (40 µg/mL). DNA integrity was verified on a 0.8% agarose gel, whereas the quantity and purity were measured by calculating the absorbance at 260-nm/280-nm ratio using the Epoch spectrophotometer (Epoch). The final samples were then stored at −80°C until they were sent for sequencing. Samples were diluted to a final concentration of 50 ng/µL and were sent to Diversity Arrays Technology Pty, Ltd., Australia, for genotyping by sequencing with the DArTseq platform, using a medium-sequencing density (generating approximately 1.25 million reads per sample). In summary, a representation of the genomic DNA was obtained by digesting DNA with two restriction enzymes (*Pst*I and *Mse*I) and the prepared libraries were sequenced on an Illumina HiSeq2000 (Illumina). A total of 77 cycles were run to produce single reads. The reference-free marker calling step was done with a Diversity Arrays Technology Pty, Ltd (DART P/L) proprietary method in the DS14 software. Reads were aligned to each other, with a threshold of two to three nucleotide mismatches, and used to call single-nucleotide polymorphisms (SNPs). Additionally, these reads were used to call presence/absence variations called SilicoDART.

Quality control and filtering loci

DArTseq SNP data csv files were read into R (V4.0.4; R Core Team, 2022) with the *gl.read.dart* function of the “*dartR*” package (V1.9.9.1; Gruber et al., 2018) and converted into genlight objects. Genlight objects were later split into three subsets: (i) one containing only individual samples and another, (ii) containing only pooled samples (*seq-pools*), and (iii) a single individual per accession (single plant).

A series of parameters were reviewed to identify potential samples and loci of low quality. This evaluation included the following: total reads per sample, total unique reads per sample, library quality (weak, downshifted, and good), sample call rate, loci call rate, minor allele frequency (maf), marker reproducibility, read depth, and polymorphism information content (Figure 1; Supplementary Figures 1–4).

Based on the descriptive statistics of the data, a set of filters was applied to all SNP subsets (individual samples and *seq-pools*) as follows: Replicability (RepAvg; the fraction of technical replicates at a locus with the same call) was set to 1; average read depth between 5 and 100 (as, unusually, high read depths can indicate paralogous regions of the genome mistakenly grouped together), and loci with call rate higher than 0.75 were retained (since samples cover a large geographical range, despite all belonging to the same species). Additionally, all monomorphic sites were removed from each dataset as they do not provide informative data.

To perform some estimations, we applied different filters. To estimate the expected heterozygosity (H_e), the dataset of 1,086 individual samples was split by accession, all missing data within the subset was removed and, following the recommendations of the work of Schmidt et al. (2021), we estimated H_e before and after removing all monomorphic sites (for further details, see Data analysis section below).

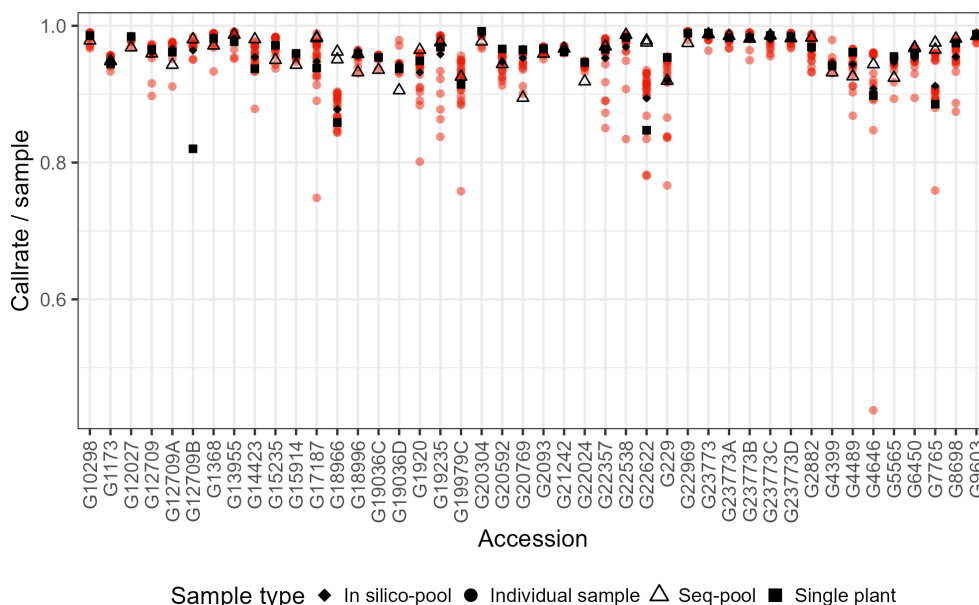


FIGURE 1

Comparison of the sample call rate between pools of *P. vulgaris* after filtering. Some accessions have two sequenced pools because of technical replicates.

An extra filter was incorporated to assess the resemblance between *seq-pools* and *in silico-pools* derived from the same accessions. This involved calculating the number of private alleles, the allele frequency difference (AFD), and a comparison of allele frequency estimates. Specifically, apart from the base filters mentioned above, an additional criterion was applied. Loci identified in both type of pools were retained by cross-referencing the AlleleIDs assigned by DArT P/L during the genotyping process. This additional step ensured a more rigorous comparison and enhanced accuracy of our analysis.

Data analysis

Sampling a random individual

To evaluate the efficacy of pooled data (either *in silico-pools* or *seq-pools*) in comparison to genotyping a single individual per accession, a random sample of 44 individuals was selected from the larger dataset of 1,086 individuals, and this dataset will be referred as the single plant subset. Each individual was drawn from each accession using a custom R script (R Core Team, 2022) with a predetermined seed to ensure replicability. To see how sampling affected the data, 10 runs of the random sampling described above were performed. The identical set of filters mentioned previously was applied to this subset to maintain consistency in the analysis. All analyses were conducted across all three datasets, except for the estimation of allele frequencies, of H_e , and the identification of private alleles (see below). Because the results from the 10 runs of sampling single plants were very consistent with each other, only the results of the first run are presented in the figures of the main text. The figures summarizing the results are available in the [Supplementary Material \(Supplementary Figures 5-10\)](#).

Allele frequency estimation and similarity between pools

To estimate the allele frequencies and assess the similarities between pools, we calculated allele frequencies within each accession for each kind of pool (*in silico-pools* and *seq-pools*). For *seq-pools*, DArT P/L provided an additional file alongside the standard report of SNPs and SilicoDArTs, containing the number of reads per allele per marker. Using these data, we calculated the frequencies as follows:

$$f_{ij} = \frac{\#reads_{ij}}{\sum_{j=1}^2 reads_j}$$

where f_{ij} is the allelic frequency of allele i at site j ; $\# reads_{ij}$ is the number of reads found for allele i at site j ; and $\sum_{j=1}^2 reads_j$ is the total number of reads at site j . Moreover, the allelic frequencies of *in silico-pools*' SNPs were estimated on a per-accession basis by using the following formula:

$$p_{ij} = f(AA) + \frac{1}{2}f(AB); q_{ij} = 1 - p_{ij}$$

where p_{ij} is the frequency of the reference allele p at locus i of accession j ; $f(AA)$ and $f(AB)$ are the frequencies of the AA and AB genotypes, respectively; and q_{ij} is the frequency of the SNP allele at locus i of accession j .

After estimating the allele frequencies of each dataset, we analyzed a series of key parameters within each pool. Specifically, we counted the number of called SNPs per pool, identified the number of missing sites, and determined the number of polymorphic sites within accessions.

To check the sampling effect on the estimate of allele frequencies, we used the technical replicates from DArT P/L for both *seq-pools* and single plants, assessing different read depth ranges. The average read depth per marker was estimated using the total read counts for the reference allele and the alternative allele, divided by the total of number of samples having reads for that marker. To compare the results from *seq-pools* and single plants derived from homogenous and heterogeneous accessions, we plotted the frequency of SNP allele reads at each marker across different read depth intervals.

Marker calling between pools and private and fixed alleles

To assess if there are differences between types of pools regarding the calling of markers, we compared the number of fixed and private alleles within each accession's pool. Following rigorous filtering and quality control procedures (as detailed in Quality control and filtering loci section), we counted those sites where a pool exhibited an exclusive allele (referred to as a private allele) in comparison to the other pool. Additionally, we assessed sites where opposite genotypes were called in each pool (referred to as fixed alleles). This comparison aimed to highlight differences in allele calling patterns between *seq-pools* and *in silico-pools*. These counts of private alleles were fit to a generalized linear model specified as follows:

$$\log(p_i) = \eta_i = \mu + \alpha_i$$

where $\log()$ is the logarithm link function between the linear predictor and the counts of private alleles (p_i); μ is the general mean; and α_i is the effect of the pool (*in silico-pool* or *seq-pool*). The model was applied using the "glm" function of R V 4.0.4 (R Core Team, 2022), utilizing the option "family = 'quasipoisson'" due to identified overdispersion. This conclusion was drawn from a preliminary analysis where the ratio between residual deviance and degrees of freedom exceeded 1. The effect of the pool (α) was tested with an analysis of deviance, as implemented in the Anova function of the car package (V3.0-12) (Fox and Weisberg, 2019), utilizing the option "test.statistic = 'F'." The estimated means from the model were back transformed to the scale of the response variable using the summary function utilizing the option "type = 'response'" in R.

In order to assess the similarity between allele frequency estimates across datasets, we calculated the AFD metric, as introduced by Berner (2019), which serves as an estimator of population differentiation to compare *in silico-pools*, *seq-pools*, and single plants. This measure was calculated using the following formula:

$$AFD = \frac{1}{2} \sum_{i=1}^n |f_{i1} - f_{i2}|$$

where f_{i1} and f_{i2} are the frequencies of allele i of an accession in datasets 1 and 2, respectively; and n is the number of markers.

Heterozygosity

Expected heterozygosity (H_e) was calculated before and after the removal of monomorphic markers, following guidelines recommended by Schmidt et al. (2021). Schmidt et al. (2021) categorized these estimates as autosomal (considering all markers) and SNP (considering only polymorphic markers) heterozygosities. To avoid confusion, especially as the term “autosomal” implies a distinction from sex chromosomes, we have referred to these estimates as H' [as per Schmidt et al. (2021)] and H for SNPs.

The H_e and H'_e were calculated using *in silico*-pools and the *seq-pools* dataset. The H_e was not estimated with the single plant dataset because this parameter is not commonly estimated on an individual basis, but rather on a population level, and we are working with accessions as populations. H_e (also known as gene diversity) is commonly defined as the expected frequency of the heterozygotes under Hardy-Weinberg equilibrium. Here, it was calculated as $H_{e_i} = 2p_iq_i$, where H_{e_i} is the expected heterozygosity at site i , and p_i and q_i are the allelic frequencies at site i . Calculations of the estimates of the heterozygosity were made with custom R scripts.

Modified Roger's distance and assessment of genetic patterns

The modified Roger's distance (MRD) was calculated both between pairs of accessions within datasets and between samples of the same accession but different subsets. This calculation was based on matrices of allelic frequencies, each corresponding to a specific type of pool (Wright, 1978, p. 91). The pairwise distances were calculated as follows:

$$MRD_{xy} = \frac{1}{\sqrt{2L}} \sqrt{\sum_{i=1}^L \sum_{j=1}^2 (\hat{p}_{ij(x)} - \hat{p}_{ij(y)})^2}$$

where MRD_{xy} is the distance between x and y ; L is the number of SNPs in the dataset; $\hat{p}_{ij(x)}$ is the frequency of the i th allele at the j th locus of sample x ; and $\hat{p}_{ij(y)}$ is the frequency of the i th allele at the j th locus of sample y . The matrices were calculated using a custom R script.

We employed various analytical techniques to unravel the genetic patterns within our dataset and to compare outputs across types of pools. Principal coordinate analysis (PCoA) was employed to understand the MRD matrix. PCoA, a dimensionality-reduction method, was executed using the “gl.pcoa” function from “dartR” package, generating a two-dimensional representation of the data. For clustering analysis, we utilized the complete linkage algorithm from the “stats” R package (V4.0.4) (R Core Team, 2022) to cluster the MRD matrix. The nodes of the resulting dendrogram were tested using a bootstrap analysis using the “boot.phylo” function of the “ape” package (V5.4.1; Paradis and Schliep, 2019) using parameters “rooted = FALSE” and “B = 1000.”

To explore population admixture, we compared the best estimation of K ancestral populations derived from all individuals, the *seq-pools*, or a single individual per accession. This comparison was conducted using the “LEA” package and the “snmf” function in R

(V3.2.0; Frichot and François, 2015). To run “snmf” with the *seq-pools*, the standard output from DArTseq was used because the input files for the “LEA” package are designed for allele counts, not allele frequencies. To run the analysis, the data (individuals, *seq-pools*, and single plants) as “genlight” objects were transformed into STRUCTURE input files using the “gl2structure” function of “dartR” package (using option “exportMarkerNames = FALSE” and all others as default). The STRUCTURE-formatted files were then converted into the geno format through the “struc2geno” function of “LEA” (parameters: “ploidy = 2, FORMAT = 2, extra.row = 0, extra.column = 1”), facilitating further in-depth analysis of genetic admixture patterns. The “snmf” method from the “LEA” package was executed for each dataset with specific parameters: “K = 1:20, ploidy = 2, entropy = TRUE, CPU = 20, repetitions = 5, iterations = 500, alpha = 100.” The optimal K , indicating the most likely number of ancestral populations given the data, was determined using the cross-entropy criterion, selecting the point where the cross entropy exhibited a plateau. Initially the ‘snmf’ run with individual samples did not display a plateau, leading to an additional run with K -values from 40 to 55. Visual representations, including bar plots of admixture coefficients and cross-entropy values plots across different K -values were generated using the ‘ggplot2’ package (V3.3.3, Wickham, 2016).

Results

Before applying any quality filters, a set of parameters, including total and unique read counts per sample, and the number of markers called, were assessed, and compared across different sample types.

For the 1,086 individual samples, the average total read count was 1,259,666 ($\pm 211,597$) and the average total unique read count was 201,500 ($\pm 48,604$). *Seq-pools*, consisting of 44 samples, exhibited a slightly higher average total and unique reads, reaching 1,271,141 ($\pm 107,025$) and 218,145 ($\pm 21,432$), respectively. In contrast, the 44 single plants showed the lowest mean counts of both total (1,241,579 $\pm 239,180$) and unique (199,673 $\pm 53,303$) reads along all the subsets. The counts of total and unique reads were more consistent across *seq-pools* samples (ranging from 985,347 to 1,443,516 and 167,534 to 267,046, respectively) than across individual samples (ranging from 594,075 to 1,744,258 and 91,370 to 364,280, respectively). The latter has a larger number of samples and a wider distribution across both variables, as reflected in the average and standard deviation of these counts on each dataset (Table 1).

After splitting the SNP data by datasets (*seq-pools*, *in silico*-pools, single plants) and removing markers with 100% missingness, the total number of called markers was very similar among the unfiltered datasets from the three sample types: 86,012 in *seq-pools*, 86,277 in *in silico*-pools, and 86,335 in the single plant subset. Among these markers, 31,677, 15,453 and 15,340 were polymorphic, respectively. Notably, the *in silico*-pools exhibited a higher average of markers called per accession (78,427 SNPs $\pm 2,150.6$) compared to either the *seq-pools* or the single plants, both of which had similar averages, 71,984 ($\pm 2,634.5$) and 71,909 ($\pm 4,711$), respectively (Table 1).

TABLE 1 Summary of the comparison between pools before and after filtering.

Dataset	Variable		<i>In silico</i> -pool	<i>Seq</i> -pool	Single plant
General information	Number of accessions	–	44	44	44
	Number of samples	–	1,086	52	44
	Count of unique sequence reads per sample	Mean	201,500	218,145	199,673
		Std. dev.	48,604	21,432	53,303
	Count of total sequence reads per sample	Mean	1,259,666	1,271,141	1,241,579
		Std. dev.	211,597	107,205	239,180
Unfiltered	Call rate/loci	Median	0.931	0.942	0.932
	Call rate per sample	Median	0.845	0.839	0.849
	maf	Mean	0.109	0.041	0.040
	Total number of SNPs	–	86,277	86,012	86,335
	Number of polymorphic SNPs across the dataset	–	31,677	15,453	15,340
Filtered	Call rate/loci	Median	0.983	1	0.977
	Call rate per sample	Median	0.963	0.969	0.951
	maf	Mean	0.110	0.241	0.240
	Number of polymorphic SNPs across the dataset	–	14,078	6,281	6,555

The effects of applying a series of filters to remove SNPs (reproducibility = 1, average read depth 5–100, call rate/locus ≥ 0.75 and removing monomorphic sites) were assessed on based on call rate, number of polymorphic sites and allele frequencies estimates (Table 1; Supplementary Figures 2–4). After filtering, the number of remaining SNPs numbered 14,078 in the *in silico*-pools, 6,281 in the *seq*-pools, and 6,555 in the single plant datasets (Table 1). A comparison of the median call rate per sample showed similarity between *seq*-pools (0.963) and the individually genotyped samples (0.969), despite differences in the number of markers and the significant variation of call rates among samples from the same accession (Figure 1). The median call rate for the single plant subset was slightly lower at 0.951 (Table 1). The number of polymorphic sites per pool/single plant varied across each dataset. In general, the *seq*-pools tended to have fewer polymorphic sites than the *in silico*-pools from the same accession and slightly more than a single plant (Figure 2; Supplementary Table 2). The number of polymorphic sites ranged from 4 to 1,357 in *seq*-pools, 372 to 3,492 in the *in silico*-pools, and 5 to 1,582 in the single plant datasets. The distribution of polymorphic SNPs varied little across resampling runs for most of the accessions, while other accessions had outlier individuals (Supplementary Figure 6).

The estimated allele frequencies from both pooled datasets revealed a wide range of homozygote markers within pools, from 75% to 97% in *in silico*-pools and 78% to 99.9% in *seq*-pools (Figure 3; Supplementary Table 2). Using the AlleleIDs from each pool type, we found that 6,142 (~97%) of the SNPs from the *seq*-pool data were also called in the *in silico*-pools. Comparing allele frequencies of these shared SNPs between types of pools showed that most markers coincide for the same allele in both pools

(Figure 3). The distribution of the homozygous SNPs within *in silico*-pools showed two groups of accessions, one highly homogeneous (i.e., over 92% of homozygous SNPs) and one heterogeneous (<92% of homogeneous SNPs, Supplementary Table 2). When comparing the frequency of SNP allele reads estimated between available technical replicates (provided by DArT P/L) of *seq*-pools (e.g. G1173, G6450, G17187) it was observed that SNPs with an average depth below 20 reads had a higher discrepancy across replicates than SNPs with higher read depth. This trend was more evident in heterogeneous accessions (e.g. G17187). The frequency of SNP allele reads of single plants replicates, was more consistent between replicates (Supplementary Figure 11).

Some SNPs that were found to be monomorphic on one pool were polymorphic in the other, i.e., one of the pools had private alleles with respect to the other (Figure 3; Supplementary Table 2). After fitting a generalized linear model with a quasi-Poisson distribution, the analysis of deviance revealed a significant effect of the type of pool on the number of private alleles (Analysis of Deviance; Dev. Residuals = 24,221, DF = 1, F = 92.5, p-value = 5.694×10^{-16}). The back-transformed estimated average of private alleles in *seq*-pools was 21.7, compared to an estimated 440.7 private alleles within *in silico*-pools. Fixed alleles (i.e., opposite alleles called in each pool) between pools were rare, for instance the highest observed count was 4 (Supplementary Table 2).

The AFD is an estimator similar to F_{st} to measure differentiation between populations (Bernier, 2019). The allele frequencies between the *in silico*-pools and the *seq*-pools two pools were highly similar, with a mean AFD of 0.008 (± 0.011) between pools. Accession G12709B, which showed a higher average AFD of 0.047 across

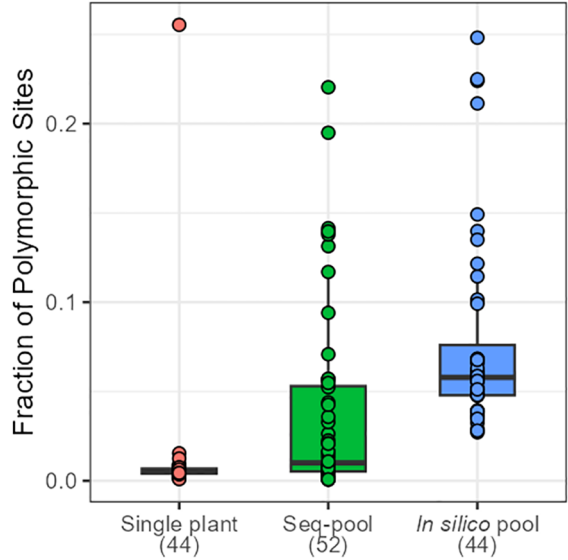


FIGURE 2
Distribution of the fraction of polymorphic SNPs across accessions of *P. vulgaris* on each dataset. Numbers in brackets indicate the number of samples per dataset.

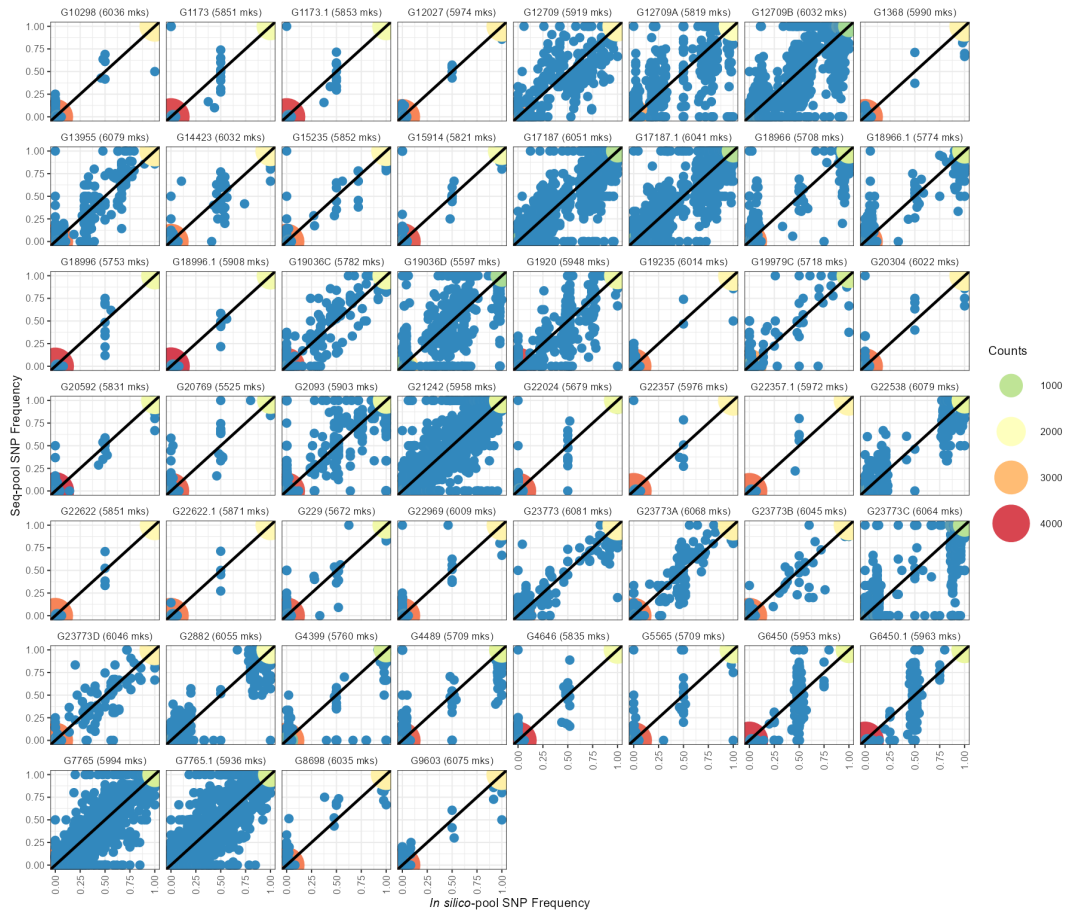


FIGURE 3
Comparison between allelic frequencies of the SNP allele between *in silico*-pools (X-axes) and *seq*-pools (Y-axes). Dot colors indicate the density of homozygous sites for the same allele in both pools. Blue dots indicate heterozygote sites on either or both pools. Next to each accession ID is the number of shared markers between pools after filtering, including monomorphic SNPs.

shared loci, behaving as an outlier (Supplementary Table 2). Meanwhile, the MRD between both pools and the single plant of the same accession (Figure 4A) showed that the smallest distances were estimated between the pools (0.034 ± 0.026), while the distances of the single plants with either the *seq-pools* of the *in silico-pools* tended to be larger (0.066 ± 0.067 and 0.057 ± 0.047 , respectively). When the data was split by homogeneous and heterogeneous accessions, the distances between *in silico-pools*, *seq-pools*, and single plants, tended to be smaller in the homogeneous group than in the heterogeneous group (Supplementary Figure 12). This pattern persisted even across all runs of resampling single plants (Supplementary Figure 8).

Although the shape of the distribution of the MRD (Wright, 1978, p. 91) was similar across datasets (Figure 4B; Supplementary Figure 13), the distances between *in silico-pools* were consistently smaller (Average MRD 0.341 ± 0.138) in comparison with either the *seq-pool* (Average MRD = 0.492 ± 0.203) or the single plant (Average MRD = 0.502 ± 0.209 ; Table 2). MRD was highly consistent across 10 runs of resampling single plants (Supplementary Figure 7).

The difference among datasets was attributed to the presence of unique SNPs detected in the *in silico-pools* but not in the *seq-pools* which, as shown in Figure 4C, tend to be markers with very low frequencies. The distance matrix based on the 6,142 shared SNPs between the *in silico-pools* and *seq-pools* showed an identical distribution to the *seq-pool* MRD matrix (Supplementary Figure 13). In contrast, estimating the distance matrix using markers exclusive to the *in silico-pool* data led to the lowest distances between *in silico-pools*, as shown in Supplementary Figure 13 with “unique markers only.” A similar pattern was observed when the AFD was calculated (Table 2), i.e., the average similarity between *in silico-pools* was higher in this dataset (0.142 ± 0.087) than either the *seq-pool* (0.313 ± 0.192) or the single plant data (0.308 ± 0.193).

The gene diversity ($H_e = 2pq$, expected heterozygosity) showed a significant variation between estimates (H_e and H'_e) and between *in silico-pool* and *seq-pool* (Figure 5). The mean H_e was 0.0026 for the *in silico-pools* and 0.0017 with the *seq-pool* data. In contrast, H'_e

was higher, averaging 0.09 and 0.31 in the *in silico-pool* and *seq-pool* datasets, respectively (Supplementary Table 3).

We employed SNP data and their corresponding distance matrices to investigate signs of population structure through PCoA, hierarchical clustering, and “snmf,” a method used to model admixture coefficients based on a given number of K ancestral populations.

In summary, all three analyses yielded consistent results across datasets (*in silico-pools*, *seq-pools*, and a single plant). They uniformly revealed the divergence and separation between the Andean and Mesoamerican gene pools of common bean. For the PCoA, this distinction was evident in the first axis, explaining 63%–64% of the variance (Figure 6) and clearly separated accessions into two distinct groups. This applies as well to the 10 resampling runs of the single plant dataset (Supplementary Figures 9, 10). Only two accessions, G21242 and G17187, were found in the space between the two groups, being more evident with the single plant subset (Figure 6). The second and third axes of the PCoA also showed an interesting pattern within each gene pool. Each axis split a group into two, with one composed mostly of accessions from the Americas and the other containing samples from other regions of the world (Supplementary Figure 14).

The hierarchical clustering analysis also separated two larger groups (Figure 7A). Although smaller groups were inconsistent, with low bootstrap support ($< 75\%$; Figure 7B). Whereas most accessions remained within the same two major clusters across the three sampling types, two accessions, G21242 y G17187, exhibited differential clustering patterns in *seq-pools* compared to *in silico-pools* and a single plant. Moreover, eight replicated *seq-pools* used by DArT P/L to estimate the replicability of the marker calling steps were also included into the tree and they confirmed the robustness of the clustering by being consistently groups together with their replicates (Figure 7A; *Seq-pool*). The panel of this study included three accessions that were subdivided into multiple accessions over time: G12709 (three accessions), G19036 (two accessions), and G23773 (five accessions). Of these, only G12709 was consistently clustered together across all trees (Figure 7A).

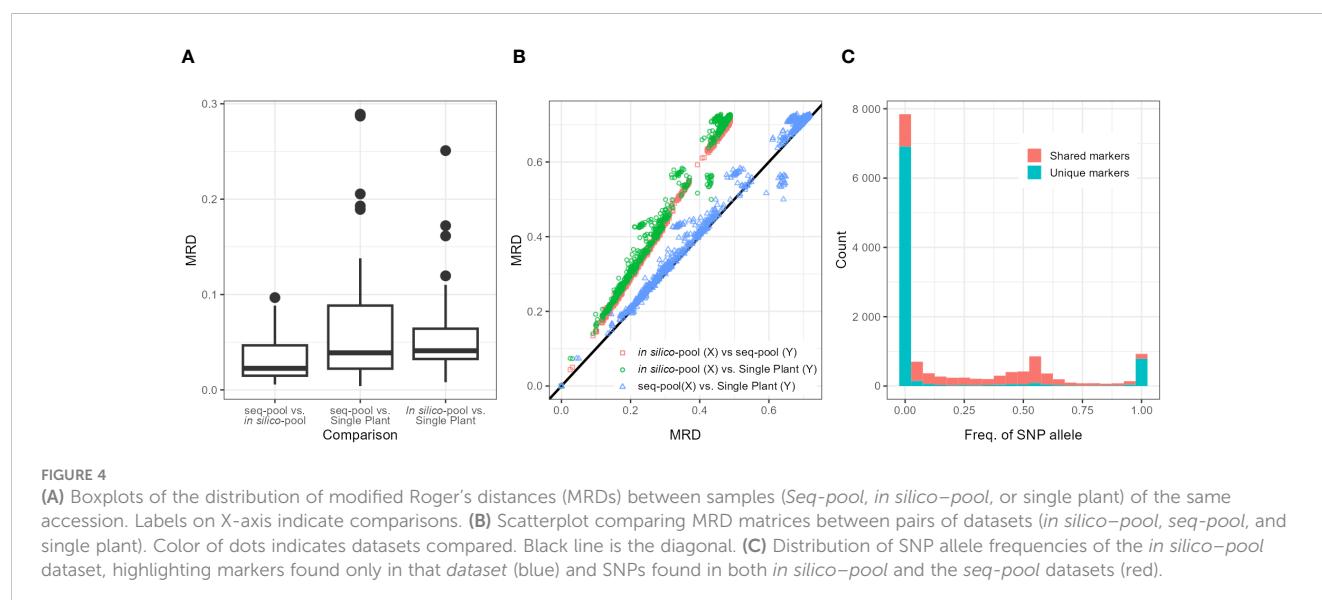


TABLE 2 Summary (mean ± std. deviation) of the allele frequency difference (AFD) and the modified Roger's distance (MRD) between accessions in each dataset.

Variable	<i>In silico-pool</i>	<i>Seq-pool</i>	Single plant
AFD	0.142 ± 0.087	0.313 ± 0.192	0.308 ± 0.193
MRD	0.341 ± 0.138	0.492 ± 0.203	0.502 ± 0.209

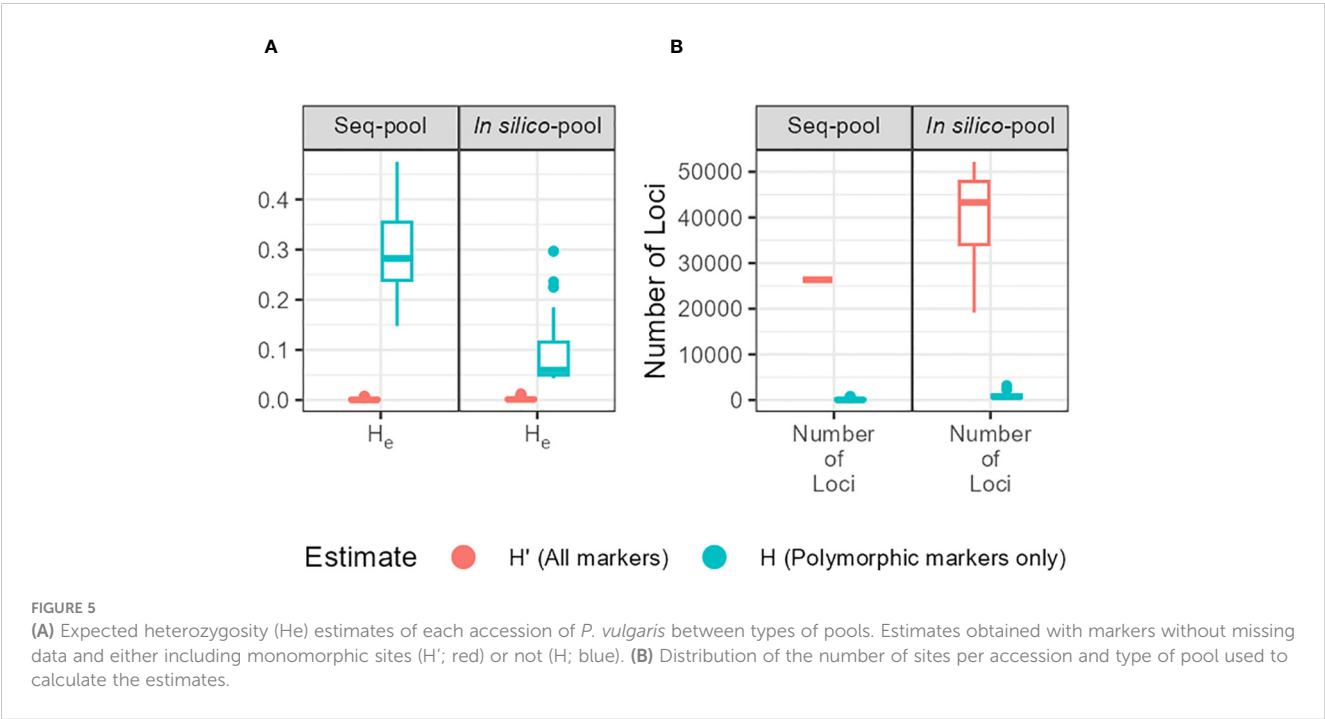
Furthermore, when studying the admixture coefficient of ancestral populations, the best fitting K-value accordingly to “snmf” was K = 2 for both the *seq-pool* and the single plant data, with a cross-entropy of the best run at 0.40 (Supplementary Figure 15). When mapping the admixture coefficients of the *seq-pool* data using the accessions’ passport data, the distribution of the ancestral populations across the Americas has a clear north-south split. That is, most Accessions originating to the south of Ecuador shared the same ancestral population, whereas accessions distributed across Central and North America shared the other ancestral population in common (Figure 8). Regarding the accessions from Africa, Asia, and Europe, most seem to share the same ancestral population with that of the South American accessions, but no clear pattern could be discerned (Figure 8A). These results are highly consistent with the two large clusters found with the hierarchical clustering (Figure 9).

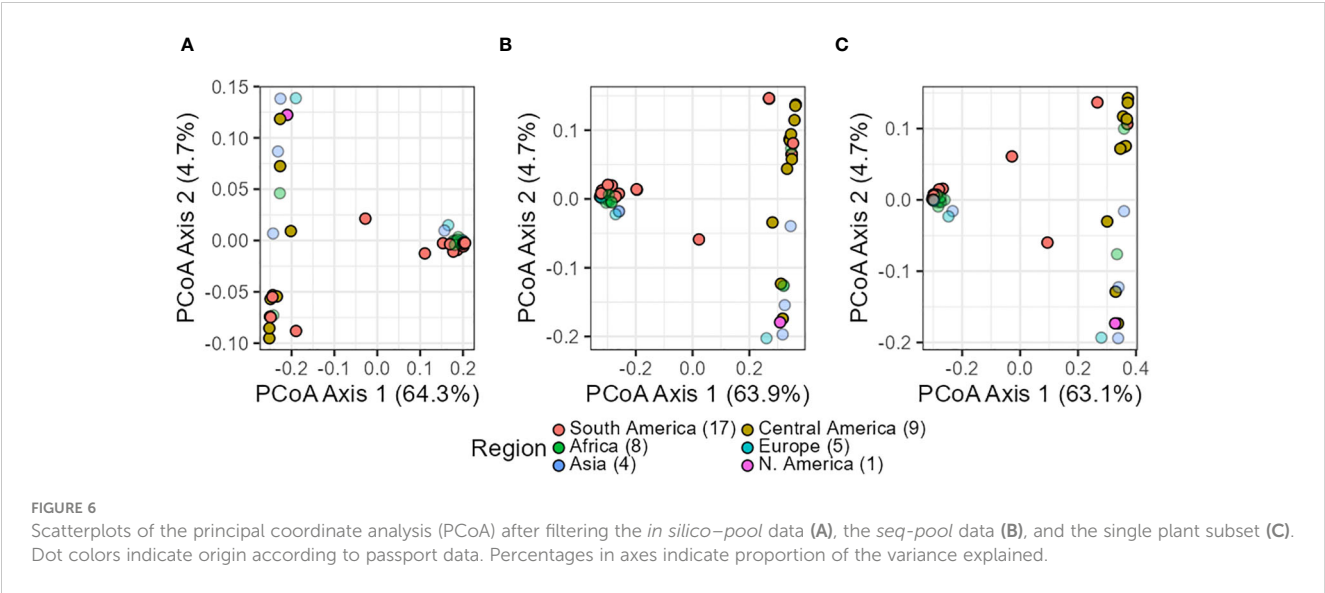
Discussion

In the last decade, there is been a notable increase in genomic characterization of long-preserved collections (Wang et al., 2018; Sansaloni et al., 2020). This trend is driven by cheaper sequencing costs and the increasing focus on maximizing the value of each accession in germplasm collections. The genetic data acquired offers

valuable insight to curators, aiding decision-making and improving access to alleles and genes linked to key traits. However, challenges persist, particularly in determining optimal sampling methods. Balancing the need for representing accessions or populations with cost-effectiveness is especially crucial for large germplasm collections managed by CGIAR. Achieving the right balance between scientific rigor and practicality is essential for effectively navigating these challenges. In this study, we genotyped 44 accessions of *P. vulgaris* using three sampling strategies to assess if analyses based on the genotype calls, estimated allele frequencies, diversity estimates, and population structure yielded consistent results across sampling methods. Our findings indicate that *in silico-pools* yielded a higher number of SNPs compared to both *seq-pools* and the single plant data. This is attributed to the individual genotyping of each member within the *in silico-pool*, which increases the likelihood of identifying rare alleles. However, calling SNPs from pooled DNA samples poses a challenge in distinguishing genuine rare variants from sequencing errors (Schlötterer et al., 2014; Anand et al., 2016). Similarly, there remains uncertainty when sampling a random individual per population/accession, as it may not accurately represent the entire population. Filtering and handling missing data are critical in genetic analyses. Methods have different tolerances to missing data, and strict filters can negatively impact downstream inferences (Wiens, 2006; Rubin et al., 2012; Huang and Knowles, 2014; Eaton et al., 2017). Conversely, some methods struggle when missingness is non-random, depending on factors like species or gene pools (Yi and Latch, 2022).

The overall population patterns observed in PCoA, snmf, and the hierarchical clustering across datasets (*seq-pool*, *in silico-pool*, and a single plant) after applying uniform filters (Reproducibility = 1, average read depth = 5–100, call rate/locus ≥ 0.75, no monomorphic sites) were similar. While these criteria may appear “lax” compared to general recommendations for filtering marker

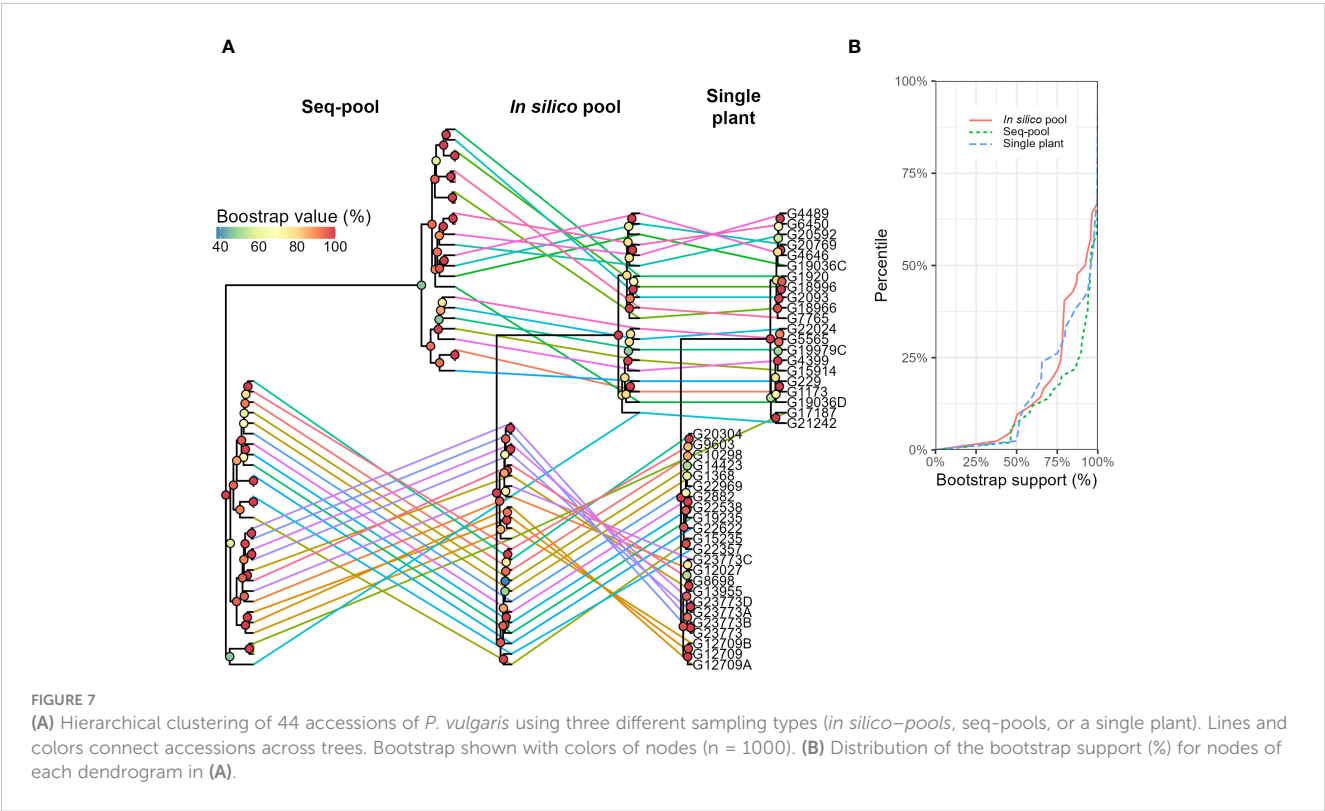


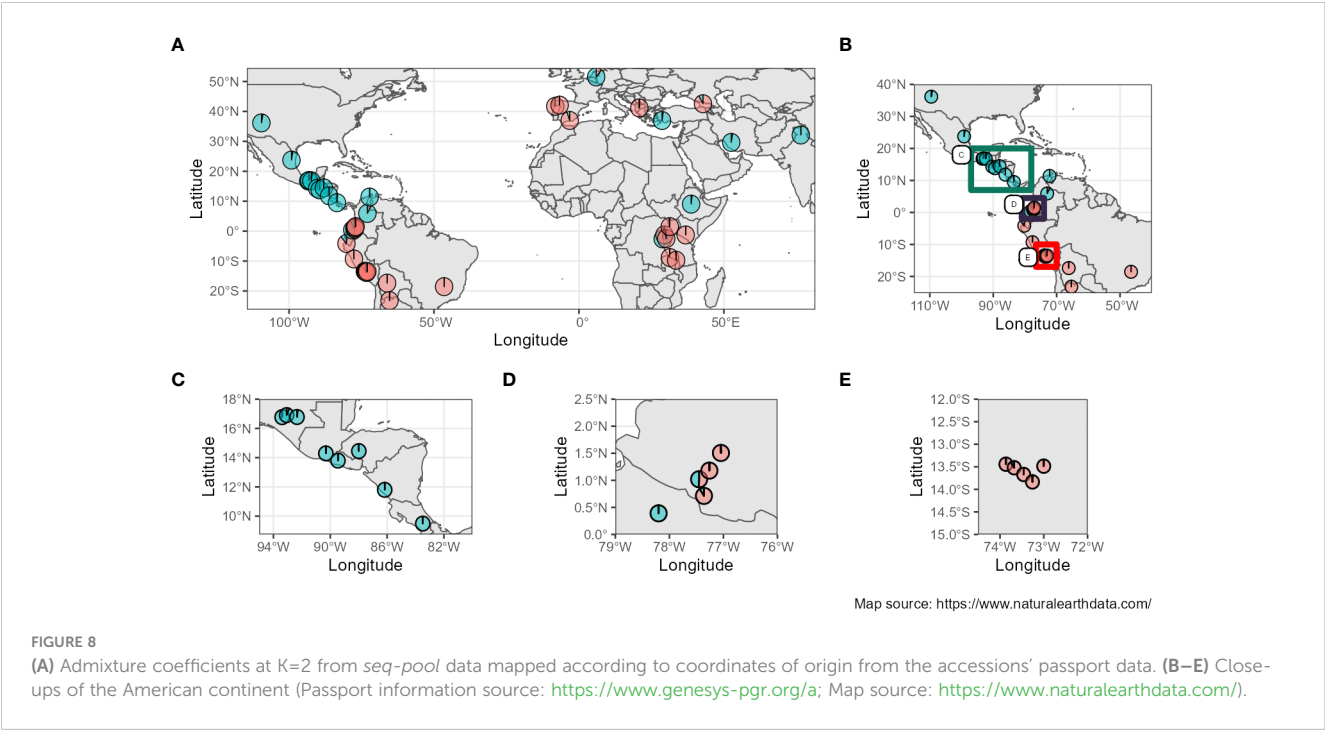


data (e.g., Carson et al., 2014; O’Leary et al., 2018; Pavan et al., 2020), our dataset encompasses a wide range of samples from diverse geographic origins, each subjected to different selection pressures and accumulating genetic differences. Similar lax filters have been employed in other studies investigating common bean genetic diversity (Valdisser et al., 2017; Nadeem et al., 2020; Gelaw et al., 2023). In this work, the aim was to retain sites displaying allele dropout, a common challenge in reduced representation approaches like DarTseq (Gautier et al., 2013), as they provide valuable insights information where they are present, making them informative across diverse populations (Wiens, 2006). Thus,

imputation was not performed to avoid assumptions about the cause of missing markers, acknowledging the biological nature of allele dropout.

Accurate estimation of allele frequencies is crucial, as it directly influences MRD matrices. While using single plants poses challenges due to varying call rates within an accession and potential bias from missing data (as depicted in Figure 1). Studies have found that estimating allele frequencies with pooled data can be more precise. This is attributed to reduced DNA contribution variance, particularly with larger pool sizes (Futschik and Schlötterer, 2010; Rellstab et al., 2013). In our study,

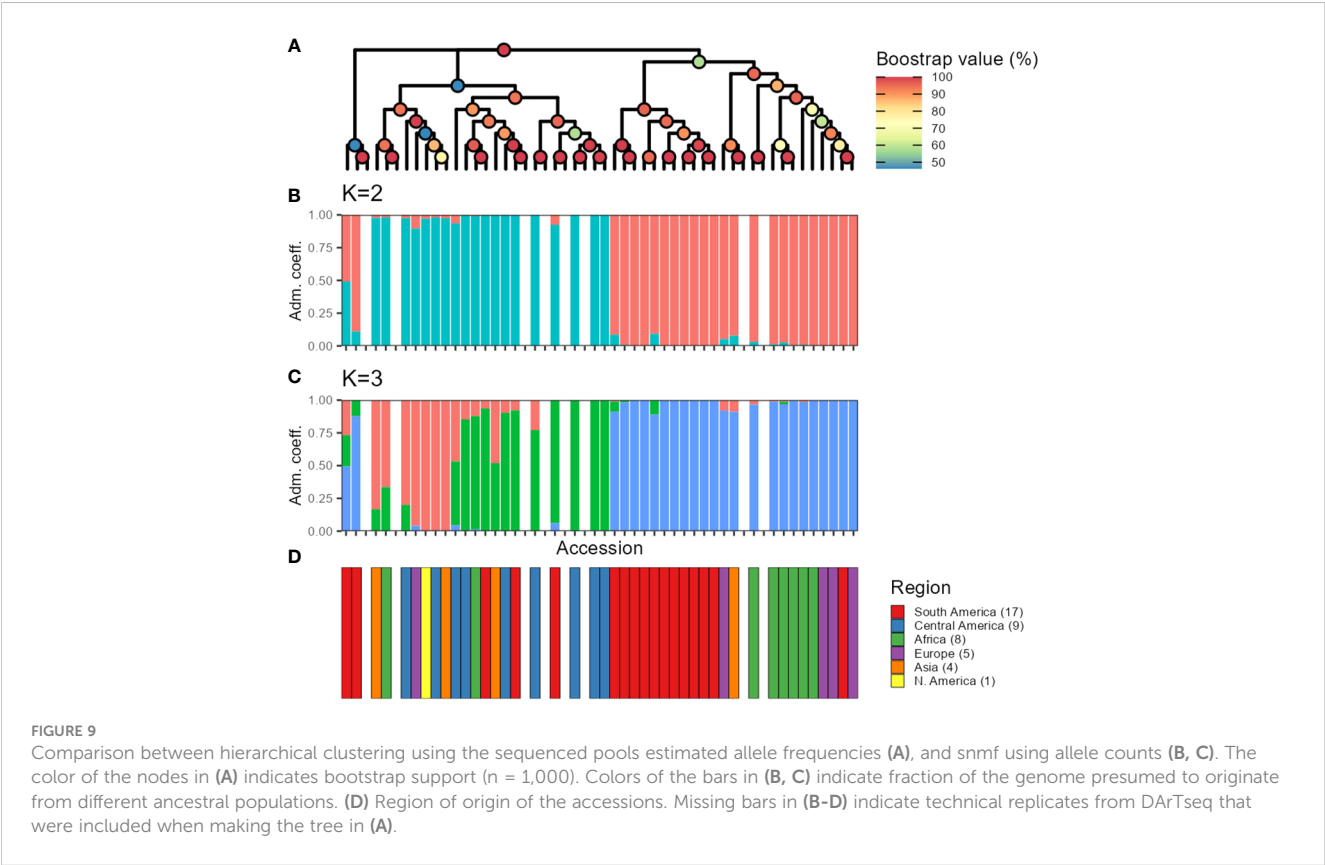




comparing allele frequencies between *seq-pools* and the *in silico-pools* revealed low AFD, suggesting minimal differentiation between pools of the same accession. Although allele frequency estimates from *seq-pools* and *in silico-pools* appear correlated, the large sample size and counts of fixed markers consistently return a

strong and significant correlation every time, which is why they are not shown here.

Seq-pools exhibit limitations in estimating intermediate (~0.5) frequencies (Figure 3), regardless of the population's polymorphic loci count. Theoretical and empirical research indicates that



variance and error of allele frequencies are highest at intermediate frequencies (Chen et al., 2012; Fung and Keenan, 2014), as is the difference between simulated and empirical allele frequencies (Hale et al., 2012). Other causes include technical artifacts such as random amplification of reads, insufficient locus depth, or uneven DNA contributions. The latter is unlikely due to meticulous control of sample tissue area per plant for consistency across individuals. When we compared the frequency of SNP allele reads between technical replicates, we observed that the least consistent estimates of allele frequencies were found in the SNPs with average depth <20 reads (Supplementary Figure 11). This difference between replicates was more evident in heterogeneous accessions for seq-pools that from single plants, suggesting a sampling effect on seq-pools, most likely due to random amplification of reads during library preparation or insufficient locus depth of rare alleles from individual including in the pool. Although a similar pattern is seen in replicates of single plants, the frequency of SNP allele reads is more consistent.

Another possible cause could be the method of allele frequencies estimation from pooled data, known as the “naive” method, where allele reads’ ratio at a locus serves as the estimate [as used by Inbar et al. (2020)]. This method may inflate minor allele frequency estimates, particularly for rare alleles (Chen and Sun, 2013). While tools exist for calling markers with pooled DNA data [see the work of Schlötterer et al. (2014) for a list of methods and for an in-depth comparison between callers], these pipelines require aligning reads to a reference genome (Guirao-Rico and González, 2021). To our knowledge, this is the first instance where read count data from DArTseq has been used for estimating allele frequencies. Regular allele counts from pools of different sizes have been employed in other crops such as Barley (*Hordeum vulgare*; Dziurdziak et al., 2021), chickpea (*Cicer arietinum*; Farahani et al., 2019), cowpea (*Vigna unguiculata*; Ketema et al., 2020), pastures (*Phalaris aquatica*; Gapare et al., 2021), and safflower (*Carthamus tinctorius*; Hassani et al., 2020).

Overall, both *in silico*- and *seq*-pools exhibited high similarity, evidenced by the low AFD, minimal private alleles between pairs, and genetic distances (Figures 4A, B; Supplementary Table 2). Despite that *in silico*-pools do discover more markers (Supplementary Table 2), predominantly low-frequency SNPs (Figure 4C), the overall difference between pools of the same accession was small. However, sample similarity was also influenced by the within-population diversity, as heterogeneous accession groups revealed higher MRD between samples of the same accession (Figure 3; Supplementary Figure 11), potentially indicating single plants’ insufficient representation of an accession. Regarding the single plant datasets, consistency across multiple random sampling runs was observed (Supplementary Figures 5–10) and with either the *seq*-pool or *in silico*-pool data (Figure 7). Nevertheless, a significant discrepancy was noted in the number of detected SNPs in this dataset (Figure 2; Supplementary Figure 6, Supplementary Table 2), suggesting that single plant data underestimates within-accession variation, which is crucial for comprehending species diversity.

After SNP filtering across datasets, a notable disparity in the count of polymorphic sites within accessions was observed. For

instance, the variance in polymorphic markers between *in silico*-pools of accessions G12709B and G20592 was substantial, with 3,492 vs. 372 SNPs, respectively. This difference was even more evident in the *seq*-pool data, with counts of 1,357 vs. 32 SNPs, respectively. In contrast, the difference between single plants of these accessions was minimal, with 16 vs. 9 SNPs (Supplementary Table 2). When examining gene diversity (expected heterozygosity, H_e) across *in silico*-pool or *seq*-pool data, the H_e estimates suggest that certain populations harbor minor alleles with moderate to high frequencies, indicating potential population sub-structure or outcrossing events. Conversely, the H_e estimates derived from either pooled dataset present a nuanced view of accession diversity across our panel. Although H_e varies considerably across populations/accessions, the values remain quite small (ranging from 0.0004 to 0.0128 for *in silico*-pools data and from 0.000034 to 0.008151 for *seq*-pools data), which fits better with a species that is mostly self-pollinating. Estimating H_e based on single plant dataset would not accurately represent the entire accession. Furthermore, the distribution of genetic distances was notably influenced by the presence of low-frequency alleles. Although the shape of the distribution across all datasets appeared similar (refer to Supplementary Figure 13), the distance matrix derived from the *in silico*-pool data was consistently smaller in magnitude (Table 2). This difference between datasets nearly disappeared when shared markers between pools were used to calculate genetic distances (Supplementary Figure 13). The presence of low-frequency SNPs reduces the MRD by increasing the denominator (2N) in the MRD formula (see Materials and Methods). Similarly, the AFD distribution was comparable between the *seq*-pool and the single plant data (Table 2), whereas the *in silico*-pool data displayed greater similarity between accessions, indicating that this metric is also sensitive to a substantial fraction of very rare alleles.

Variation in the within-population diversity of landraces of common bean was observed (Figures 3, 5), potentially attributed to the diverse origins of the included accessions in this study (Supplementary Table 1) and the fact that landraces are generally more genetically diverse compared to modern counterparts (Byrne et al., 2020; Wilker et al., 2020). Across the accessions included in this study, there are some homozygous accessions for almost all loci with some residual heterozygosity (e.g., G10298 and G1368), whereas other accessions are more heterozygous (e.g., G17187 and G21242). The more heterogeneous accessions suggest that they could be a mixture of seeds, a frequent scenario in common bean, potentially enhancing diversity (Blair et al., 2010; García-Narváez et al., 2020). This contrasts with the expected low within-population diversity of a mostly selfing species like *P. vulgaris*, noting that crossing rates may vary from 2.5% up to 70% (Wells et al., 1988; Ibarra-Perez et al., 1997; Ferreira et al., 2000; Royer et al., 2002; Chacón-Sánchez et al., 2021). While DNA pooling is uncommon in common bean genetic diversity studies, its application has focused on variations between gene pools (Papa et al., 2007) or used in different marker systems like microsatellites (Zhang et al., 2008; Asfaw et al., 2009) and simple sequence repeats (Özkan et al., 2022). Because the most diverse accessions coincided between *seq*-pools and *in silico*-pools (Supplementary Table 2), *seq*-pools offers a promising approach for identifying accessions with high genetic diversity (heterogenous

accessions). This information is valuable not only for gene bank users but also for seed collection curators. This highlights a limitation of single plant data because one individual may not adequately represent the diversity of an entire population/accession. This limitation is particularly relevant in the study of landraces, wild forms of *P. vulgaris*, and cross-pollinating *Phaseolus* species. After all, fewer polymorphic SNPs were detected within accessions compared to both *seq-pool* and *in silico-pool* data, emphasizing the importance of pooled sequencing methods for comprehensive diversity assessment (Figure 4A; Supplementary Table 2).

Apart from the mentioned challenge of estimating allele frequencies, a key limitation associated with the use of *seq-pools* lies in the difficulty in accurately estimating the observed heterozygosity within populations of an accession, as highlighted previously by Chen et al. (2022). In our study, we were unable to compare estimates of H_o across datasets. This metric can only be calculated using the *in silico-pools* dataset, where individual genotypes are available and not with *seq-pools* or single plants. Additionally, pooling does not allow us to distinguish whether a heterogeneous accession results from a recent cross or a seed mixture.

As mentioned above, PCoA, hierarchical clustering, and “snmf,” revealed consistent patterns of population structure within *P. vulgaris*, identifying two major ancestral groups across all datasets: *seq-pool*, *in silico-pool*, and single plant datasets. These findings align with the current consensus of domesticated *P. vulgaris* having two major gene pools: the Mesoamerican and the Andean groups (Blair et al., 2012). We also identified G21242 as a potential hybrid, consistent with previous research (Blair et al., 2006). Our results parallel the findings of Arca et al. (2023) in maize pools, demonstrating the consistency of PCoA, hierarchical clustering, and admixture coefficients, albeit utilizing microarray and measurement of fluorescence ratios data for allele frequency estimation. Whereas the PCoA and the hierarchical clustering exhibited similar patterns across datasets, the PCoA based on allele frequencies from *seq-pool* data revealed more distinct groups along the second and third axes compared to *in silico-pool* or single plant data (Supplementary Figure 14). Notably, the division within major groups appeared to segregate American and non-American accessions, which could be attributed to the selection process after introduction into new environments. The “snmf” analysis with *in silico-pool*'s utilized all 1,086 individual samples, leading to a significant difference in estimating the optimal number of ancestral populations compared to *seq-pool* and the random individual data (Supplementary Figure 15). This discrepancy could be attributed to data redundancy or a bias from abundant rare alleles with low informativeness (Linck and Battey, 2019). Conversely, the analysis with the *seq-pool* samples showed less sensitivity, possibly due to the smaller number of accessions studied ($n = 44$), which may not have sufficient for rare alleles to exert significant influence. Nevertheless, the *seq-pool* data remained highly consistent with the estimated ancestry coefficients derived from *in silico-pools* and single plants at $K = 2$ (Supplementary Figure 16).

Our findings demonstrate that using pooled DNA for studying the genetic diversity of domesticated *Phaseolus vulgaris* yields

comparable insights to sequencing individuals, despite certain limitations such as challenges in estimating intermediate allele frequencies and lack of individual genotypes. Despite these limitations, pooled samples remain the most practical sampling strategy for large-scale genotyping efforts of germplasm collections. Genotyping individuals significantly multiplies the workload and resources required by a factor of “ n ” (where “ n ” represents the number of samples to be pooled). This increased demand extends not only to field and lab work but also to sequencing efforts, genotyping, and all subsequent data analyses, requiring substantially larger computational resources and processing time. Although other alternatives, such as WGS or arrays, exist to genotype plant genetic resources, the former remains costly for large-scale projects, although it has the advantage of generating significantly more data. Microarrays, on the other hand, have well-known issues with ascertainment bias (Arca et al., 2023), and the amount of data generated would be insufficient for association studies or analyses beyond genetic diversity.

This study provides valuable guidance for gene bank researchers undertaking genotyping initiatives, aiding in effective collection management, and facilitating marker-trait association studies for identifying candidate markers associated with key traits.

Data availability statement

The data presented in the study are deposited in the Dataverse repository, accession number <https://doi.org/10.7910/DVN/MQCSC4>.

Author contributions

MCA: Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Methodology. JO: Formal analysis, Methodology, Writing – review & editing. PW: Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition. MC-Y: Methodology, Supervision, Writing – review & editing, Conceptualization, Project administration, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by The Crop Trust, the Federal Ministry for Economic Cooperation and Development (BMZ), the GeneBank Platform, and the CGIAR GeneBank Initiative. This support enabled the conduct of the research and the provision of researcher positions and covered publication fees. MCA was co-funded by the Center for International Migration and Development (CIM) of the German Government.

Acknowledgments

The authors would like to thank Luis Guillermo Santos and the seed conservation group at the Alliance Bioversity-CIAT for providing the seeds used for DNA extractions as well as Vincent Johnson for editing 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.

References

- Adjebeng-Danquah, J., Manu-Aduening, J., Asante, I. K., Agyare, R. Y., Gracen, V., and Offei, S. K. (2020). Genetic diversity and population structure analysis of Ghanaian and exotic cassava accessions using simple sequence repeat (SSR) markers. *Heliyon* 6, e03154. doi: 10.1016/j.heliyon.2019.e03154
- Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., et al. (2016). Next generation sequencing of pooled samples: guideline for variants' Filtering. *Sci. Rep.* 6, 33735. doi: 10.1038/srep33735
- Arca, M., Gouesnard, B., Mary-Huard, T., Le Paslier, M.-C., Bauland, C., Combes, V., et al. (2023). Genotyping of DNA pools identifies untapped landraces and genomic regions to develop next-generation varieties. *Plant Biotechnol. J.* 21, 1123–1139. doi: 10.1111/pbi.14022
- Ariani, A., Berny Mier y Teran, J. C., and Gepts, P. (2018). Spatial and temporal scales of range expansion in wild *Phaseolus vulgaris*. *Mol. Biol. Evol.* 35, 119–131. doi: 10.1093/molbev/msx273
- Asfaw, A., Blair, M. W., and Almekinders, C. (2009). Genetic diversity and population structure of common bean (*Phaseolus vulgaris* L.) landraces from the East African highlands. *Theor. Appl. Genet.* 120, 1–12. doi: 10.1007/s00122-009-1154-7
- Berner, D. (2019). Allele frequency difference AFD—an intuitive alternative to FST for quantifying genetic population differentiation. *Genes* 10, 308. doi: 10.3390/genes10040308
- Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., et al. (2017). Beans (*Phaseolus* spp.) as a model for understanding crop evolution. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00722
- Blair, M. W., Giraldo, M. C., Buendia, H. F., Tovar, E., Duque, M. C., and Beebe, S. E. (2006). Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 113, 100–109. doi: 10.1007/s00122-006-0276-4
- Blair, M. W., González, L. F., Kimani, P. M., and Butare, L. (2010). Genetic diversity, inter-gene pool introgression and nutritional quality of common beans (*Phaseolus vulgaris* L.) from Central Africa. *Theor. Appl. Genet.* 121, 237–248. doi: 10.1007/s00122-010-1305-x
- Blair, M. W., Soler, A., and Cortés, A. J. (2012). Diversification and population structure in common beans (*Phaseolus vulgaris* L.). *PLoS One* 7, e49488. doi: 10.1371/journal.pone.0049488
- Byrne, P., Richards, C., and Volk, G. (2020). *From Wild Species to Landraces and Cultivars*, in *Crop Wild Relatives and their Use in Plant Breeding*. Available online at: <https://colostate.pressbooks.pub/cropwildrelatives/chapter/from-wild-species-to-landraces-and-cultivars/> (Accessed August 22, 2023).
- 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
- Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J.-B., et al. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinf.* 15, 125. doi: 10.1186/1471-2105-15-125
- Chacón-Sánchez, M. I., Martínez-Castillo, J., Duitama, J., and Deboucq, D. G. (2021). Gene flow in phaseolus beans and its role as a plausible driver of ecological fitness and expansion of cultigens. *Front. Ecol. Evol.* 9. doi: 10.3389/fevo.2021.618709
- Chen, X., Listman, J. B., Slack, F. J., Gelernter, J., and Zhao, H. (2012). Biases and errors on allele frequency estimation and disease association tests of next generation sequencing of pooled samples. *Genet. Epidemiol.* 36, 549–560. doi: 10.1002/gepi.21648
- Chen, C., Parejo, M., Momeni, J., Langa, J., Nielsen, R. O., Shi, W., et al. (2022). Population structure and diversity in european honeybees (*Apis mellifera* L.)—An

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.2024.1338332/full#supplementary-material>

empirical comparison of pool and individual whole-genome sequencing. *Genes* 13, 182. doi: 10.3390/genes13020182

Chen, Q., and Sun, F. (2013). A unified approach for allele frequency estimation, SNP detection and association studies based on pooled sequencing data using EM algorithms. *BMC Genomics* 14, S1. doi: 10.1186/1471-2164-14-S1-S1

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

Dellaporta, S. L., Wood, J., and Hicks, J. B. (1983). A plant DNA miniprep: Version II. *Plant Mol. Biol. Rep.* 1, 19–21. doi: 10.1007/BF02712670

Dorant, Y., Benestan, L., Rougemont, Q., Normandeau, E., Boyle, B., Rochette, R., et al. (2019). Comparing Pool-seq, Rapture, and GBS genotyping for inferring weak population structure: The American lobster (*Homarus americanus*) as a case study. *Ecol. Evol.* 9, 6606–6623. doi: 10.1002/ece3.5240

Doyle, J. J., and Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus* 12 (1), 13–15.

Dziurdziak, J., Gryziak, G., Groszyk, J., Podyma, W., and Boczkowska, M. (2021). DArTseq genotypic and phenotypic diversity of barley landraces originating from different countries. *Agronomy* 11, 2330. doi: 10.3390/agronomy11112330

Eaton, D. A. R., Spriggs, E. L., Park, B., and Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biol.* 66, 399–412. doi: 10.1093/sysbio/syw092

Farahani, S., Maleki, M., Mehrabi, R., Kanouni, H., Scheben, A., Batley, J., et al. (2019). Whole genome diversity, population structure, and linkage disequilibrium analysis of chickpea (*Cicer arietinum* L.) genotypes using genome-wide DArTseq-based SNP markers. *Genes* 10, 676. doi: 10.3390/genes10090676

Ferreira, J. J., Alvarez, E., Fueyo, M. A., Roca, A., and Giraldez, R. (2000). Determination of the outcrossing rate of *Phaseolus vulgaris* L. using seed protein markers. *Euphytica* 113, 257–261. doi: 10.1023/A:1003907130234

Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression. Third. Thousand Oaks CA: Sage*. Available online at: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

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

Fung, T., and Keenan, K. (2014). Confidence intervals for population allele frequencies: the general case of sampling from a finite diploid population of any size. *PLoS One* 9, e85925. doi: 10.1371/journal.pone.0085925

Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186, 207–218. doi: 10.1534/genetics.110.114397

Gapare, W. J., Kilian, A., Stewart, A. V., Smith, K. F., and Culvenor, R. A. (2021). Genetic diversity among wild and cultivated germplasm of the perennial pasture grass *Phalaris aquatica*, using DArTseq SNP marker analysis. *Crop Pasture Sci.* 72, 823–840. doi: 10.1071/CP21112

García-Narváez, A. L., Hernández-Delgado, S., Chávez-Servia, J. L., and Mayek-Pérez, N. (2020). Variabilidad morfológica y agronómica de germoplasma de frijol cultivado en Oaxaca, México. *Rev. Bio Cienc.* 7, 12 pág–12 pág. doi: 10.15741/revbio.07.e876

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdellhué, C., Pudlo, P., et al. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22, 3165–3178. doi: 10.1111/mec.12089

- Gelaw, Y. M., Eleblu, J. S. Y., Ofori, K., Fenta, B. A., Mukankusi, C., Emam, E. A., et al. (2023). High-density DArTseq SNP markers revealed wide genetic diversity and structured population in common bean (*Phaseolus vulgaris* L.) germplasm in Ethiopia. *Mol. Biol. Rep.* 50, 6739–6751. doi: 10.1007/s11033-023-08498-y
- Gouda, A. C., Ndjiondjop, M. N., Djedatin, G. L., Warburton, M. L., Goungoulou, A., Kpeki, S. B., et al. (2020). Comparisons of sampling methods for assessing intra- and inter-accession genetic diversity in three rice species using genotyping by sequencing. *Sci. Rep.* 10, 13995. doi: 10.1038/s41598-020-70842-0
- Greene, S. L., Hart, T. C., and Afonin, A. (1999). Using geographic information to acquire wild crop germplasm for ex situ collections: II. Post-collection analysis. *Crop Sci.* 39, 843–849. doi: 10.2135/cropsci1999.0011183X003900030038x
- Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). DartR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* 18, 691–699. doi: 10.1111/1755-0998.12745
- Guirao-Rico, S., and González, J. (2021). Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. *Mol. Ecol. Resour.* 21, 1216–1229. doi: 10.1111/1755-0998.13343
- Hale, M. L., Burg, T. M., and Steeves, T. E. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS One* 7, e45170. doi: 10.1371/journal.pone.0045170
- Hassani, S. M. R., Talebi, R., Pourdad, S. S., Naji, A. M., and Fayaz, F. (2020). In-depth genome diversity, population structure and linkage disequilibrium analysis of worldwide diverse safflower (*Carthamus tinctorius* L.) accessions using NGS data generated by DArTseq technology. *Mol. Biol. Rep.* 47, 2123–2135. doi: 10.1007/s11033-020-05312-x
- Haupt, M., and Schmid, K. (2020). Combining focused identification of germplasm and core collection strategies to identify genebank accessions for central European soybean breeding. *Plant Cell Environ.* 43, 1421–1436. doi: 10.1111/pce.13761
- Huang, H., and Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic Biol.* 65, 357–365. doi: 10.1093/sysbio/syu046
- Ibarra-Perez, F. J., Ehdaie, B., and Waines, J. G. (1997). Estimation of outcrossing rate in common bean. *Crop Sci.* 37, 60–65. doi: 10.2135/cropsci1997.0011183X003700010009x
- Inbar, S., Cohen, P., Yahav, T., and Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Comput. Biol.* 16, e1007653. doi: 10.1371/journal.pcbi.1007653
- Ketema, S., Tesfaye, B., Keneni, G., Fenta, B. A., Assefa, E., Greliche, N., et al. (2020). DArTseq SNP-based markers revealed high genetic diversity and structured population in Ethiopian cowpea [*Vigna unguiculata* (L.) Walp.] germplasms. *PLoS One* 15, e0239122. doi: 10.1371/journal.pone.0239122
- Khoury, C. K., Castañeda-Alvarez, N. P., Achicanoy, H. A., Sosa, C. C., Bernau, V., Kassa, M. T., et al. (2015). Crop wild relatives of pigeon pea [*Cajanus cajan* (L.) Millsp.]: Distributions, ex situ conservation status, and potential genetic resources for abiotic stress tolerance. *Biol. Conserv.* 184, 259–270. doi: 10.1016/j.biocon.2015.01.032
- Linck, E., and Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* 19, 639–647. doi: 10.1111/1755-0998.12995
- Martins, S. R., Vences, F. J., Sáenz de Miera, L. E., Barroso, M. R., and Carnide, V. (2006). RAPD analysis of genetic diversity among and within Portuguese landraces of common white bean (*Phaseolus vulgaris* L.). *Scientia Hort.* 108, 133–142. doi: 10.1016/j.scienta.2006.01.031
- Nadeem, M. A., Gündoğdu, M., Ercişli, S., Karaköy, T., Saracoğlu, O., Habyarimana, E., et al. (2020). Uncovering phenotypic diversity and DArTseq marker loci associated with antioxidant activity in common bean. *Genes* 11, 36. doi: 10.3390/genes11010036
- Nadeem, M. A., Habyarimana, E., Çiftçi, V., Nawaz, M. A., Karaköy, T., Comertpay, G., et al. (2018). Characterization of genetic diversity in Turkish common bean gene pool using phenotypic and whole-genome DArTseq-generated silicoDArT marker information. *PLoS One* 13, e0205363. doi: 10.1371/journal.pone.0205363
- O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., and Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* 27, 3193–3206. doi: 10.1111/mec.14792
- Özkan, G., Haliloğlu, K., Türkoğlu, A., Öztürk, H. I., Elkoca, E., and Pocza, P. (2022). Determining genetic diversity and population structure of common bean (*Phaseolus vulgaris* L.) landraces from Türkiye using SSR markers. *Genes* 13, 1410. doi: 10.3390/genes13081410
- Papa, R., Bellucci, E., Rossi, M., Leonardi, S., Rau, D., Gepts, P., et al. (2007). Tagging the signatures of domestication in common bean (*Phaseolus vulgaris*) by means of pooled DNA samples. *Ann. Bot.* 100, 1039–1051. doi: 10.1093/aob/mcm151
- Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633
- Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., and D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00447
- R Core Team (2022). *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: <https://www.R-project.org/>.
- Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., and Fischer, M. C. (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One* 8, e80422. doi: 10.1371/journal.pone.0080422
- Royer, M. R., Gonçalves-Vidigal, M. C., Scapim, C. A., Soares, P., Filho, V., and Terada, Y. (2002). Outcrossing in common bean. *Cropps Breed. Appl. Biotechnol.* 2, 49–54. doi: 10.12702/1984-7033.v02n01a07
- Rubin, B. E. R., Ree, R. H., and Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PLoS One* 7, 1–12. doi: 10.1371/journal.pone.0033394
- Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petrol, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11, 4572. doi: 10.1038/s41467-020-18404-w
- Sansaloni, C., Petrol, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., et al. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proc.* 5, P54. doi: 10.1186/1753-6561-5-S7-P54
- Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803
- Schmidt, T. L., Jasper, M., Weeks, A. R., and Hoffmann, A. A. (2021). Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods Ecol. Evol.* 12, 1888–1898. doi: 10.1111/2041-210X.13659
- Smith, S. E., Johnson, D. W., Conta, D. M., and Hotchkiss, J. R. (1994). Using climatological, geographical, and taxonomic information to identify sources of mature-plant salt tolerance in alfalfa. *Crop Sci.* 34, 690–694. doi: 10.2135/cropsci1994.0011183X003400030017x
- Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., and Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* 61, 839–852. doi: 10.1002/csc2.20377
- Valdisser, P. A. M. R., Pereira, W. J., Almeida Filho, J. E., Müller, B. S. F., Coelho, G. R. C., de Menezes, I. P. P., et al. (2017). In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. *BMC Genomics* 18, 423. doi: 10.1186/s12864-017-3805-4
- Wamalwa, E. N., Muoma, J., and Wekesa, C. (2016). Genetic diversity of cowpea (*Vigna unguiculata* (L.) Walp.) accession in Kenya gene bank based on simple sequence repeat markers. *Int. J. Genomics* 2016, e8956412. doi: 10.1155/2016/8956412
- Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9
- Wells, W. C., Isom, W. H., and Waines, J. G. (1988). Outcrossing rates of six common bean lines. *Crop Sci.* 28, 177–178. doi: 10.2135/cropsci1988.0011183X002800010038x
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). Available at: <https://ggplot2.tidyverse.org>.
- Wiens, J. J. (2006). Missing data and the design of phylogenetic analyses. *J. Biomed. Inf.* 39, 34–42. doi: 10.1016/j.jbi.2005.04.001
- Wilker, J., Humphries, S., Rosas-Sotomayor, J. C., Gómez Cerna, M., Torkamaneh, D., Edwards, M., et al. (2020). Genetic diversity, nitrogen fixation, and water use efficiency in a panel of honduran common bean (*Phaseolus vulgaris* L.) landraces and modern genotypes. *Plants* 9, 1238. doi: 10.3390/plants9091238
- Wright, S. (1978). *Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations* (Chicago: University of Chicago Press).
- Wu, W.-D., Liu, W.-H., Sun, M., Zhou, J.-Q., Liu, W., Zhang, C.-L., et al. (2019). Genetic diversity and structure of Elymus tangutorum accessions from western China as unraveled by AFLP markers. *Heredity* 156, 8. doi: 10.1186/s41065-019-0082-z
- Yi, X., and Latch, E. K. (2022). Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Mol. Ecol. Resour.* 22, 602–611. doi: 10.1111/1755-0998.13498
- Zhang, X., Blair, M. W., and Wang, S. (2008). Genetic diversity of Chinese common bean (*Phaseolus vulgaris* L.) landraces assessed with simple sequence repeat markers. *Theor. Appl. Genet.* 117, 629–640. doi: 10.1007/s00122-008-0807-2

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

